

- New Waveforms for 5G Networks
- 5G Radio Access Architecture and Technologies
- Green Communications and Computing Networks
- Communications Education and Training:
Educational Services Board



Now...

2 Ways to Access the IEEE Member Digital Library

With two great options designed to meet the needs—and budget—of every member, the IEEE Member Digital Library provides full-text access to any IEEE journal article or conference paper in the IEEE *Xplore*® digital library.

Simply choose the subscription that's right for you:

IEEE Member Digital Library

Designed for the power researcher who needs a more robust plan. Access all the IEEE content you need to explore ideas and develop better technology.

- 25 article downloads every month

IEEE Member Digital Library Basic

Created for members who want to stay up-to-date with current research. Access IEEE content and rollover unused downloads for 12 months.

- 3 new article downloads every month

Get the latest technology research.

Try the IEEE Member Digital Library—FREE!

www.ieee.org/go/trymdl



IEEE Member Digital Library is an exclusive subscription available only to active IEEE members.

Director of Magazines

Raouf Boutaba, University of Waterloo (Canada)

Editor-in-Chief

Osman S. Gebizlioglu, Huawei Tech. Co., Ltd. (USA)

Associate Editor-in-Chief

Tarek El-Bawab, Jackson State University (USA)

Senior Technical Editors

Nim Cheung, ASTRI (China)

Nelson Fonseca, State Univ. of Campinas (Brazil)

Steve Gorshe, PMC-Sierra, Inc (USA)

Sean Moore, Centripetal Networks (USA)

Peter T. S. Yum, The Chinese U. Hong Kong (China)

Technical Editors

Mohammed Atiqzazzaman, Univ. of Oklahoma (USA)

Guillermo Atkin, Illinois Institute of Technology (USA)

Mischa Dohler, King's College London (UK)

Frank Effenberger, Huawei Technologies Co., Ltd. (USA)

Tarek El-Bawab, Jackson State University (USA)

Xiaoming Fu, Univ. of Goettingen (Germany)

Stefano Galli, ASSIA, Inc. (USA)

Admela Jukan, Tech. Univ. Carolo-Wilhelmina zu Braunschweig (Germany)

Vimal Kumar Khanna, mCalibre Technologies (India)

Yoichi Maeda, Telecommun. Tech. Committee (Japan)

Nader F. Mir, San Jose State Univ. (USA)

Seshrathi Mohan, University of Arkansas (USA)

Mohamed Moustafa, Egyptian Russian Univ. (Egypt)

Tom Oh, Rochester Institute of Tech. (USA)

Glenn Parsons, Ericsson Canada (Canada)

Joel Rodrigues, Univ. of Beira Interior (Portugal)

Jungwoo Ryoo, The Penn. State Univ.-Altoona (USA)

Antonio Sánchez Esguevillas, Telefonica (Spain)

Mostafa Hashem Sherif, AT&T (USA)

Tom Starr, AT&T (USA)

Ravi Subrahmanyam, InVisage (USA)

Danny Tsang, Hong Kong U. of Sci. & Tech. (China)

Hsiao-Chun Wu, Louisiana State University (USA)

Alexander M. Wyglinski, Worcester Poly. Institute (USA)

Jun Zheng, Nat'l. Mobile Commun. Research Lab (China)

Series Editors

Ad Hoc and Sensor Networks

Edoardo Biagioni, U. of Hawaii, Manoa (USA)

Ciprian Dobre, Univ. Politehnica of Bucharest (Romania)

Silvia Giordano, Univ. of App. Sci. (Switzerland)

Automotive Networking and Applications

Wai Chen, Telcordia Technologies, Inc (USA)

Luca Delgrossi, Mercedes-Benz R&D N.A. (USA)

Timo Kosch, BMW Group (Germany)

Tadao Saito, University of Tokyo (Japan)

Consumer Communications and Networking

Ali Begen, Cisco (Canada)

Mario Kolberg, University of Sterling (UK)

Madjid Merabti, Liverpool John Moores U. (UK)

Design & Implementation

Vijay K. Gurbani, Bell Labs/Alcatel Lucent (USA)

Salvatore Loreto, Ericsson Research (Finland)

Ravi Subrahmanyam, Invisage (USA)

Green Communications and Computing Networks

Song Guo, University of Aizu (Japan)

John Thompson, Univ. of Edinburgh (UK)

RangaRao V. Prasad, Delft Univ. of Tech. (The Netherlands)

Jinsong Wu, Alcatel-Lucent (China)

Honggang Zhang, Zhejiang Univ. (China)

Integrated Circuits for Communications

Charles Chien, CreoNex Systems (USA)

Zhiwei Xu, SST Communication Inc. (USA)

Network and Service Management

George Pavlou, U. College London (UK)

Juergen Schoenwaelder, Jacobs University (Germany)

Networking Testing and Analytics

Ying-Dar Lin, National Chiao Tung University (Taiwan)

Erica Johnson, University of New Hampshire (USA)

Irena Atov, InClusive Technologies (USA)

Optical Communications

Admela Jukan, Tech. Univ. Braunschweig, Germany (USA)

Xiang Lu, Futurewei Technologies, Inc. (USA)

Radio Communications

Thomas Alexander, Ixia Inc. (USA)

Amitabh Mishra, Johns Hopkins Univ. (USA)

Columns

Book Reviews

Piotr Cholda, AGH U. of Sci. & Tech. (Poland)

History of Communications

Steve Weinsten (USA)

Regulatory and Policy Issues

J. Scott Marcus, WIK (Germany)

Jon M. Peha, Carnegie Mellon U. (USA)

Technology Leaders' Forum

Steve Weinsten (USA)

Very Large Projects

Ken Young, Telcordia Technologies (USA)

Publications Staff

Joseph Milizzo, Assistant Publisher

Susan Lange, Online Production Manager

Jennifer Porcello, Production Specialist

Catherine Kemelmacher, Associate Editor



IEEE

IEEE ComSoc
IEEE Communications Society

- 4 THE PRESIDENT'S PAGE
- 6 BOOK REVIEWS
- 8 CONFERENCE CALENDAR
- 9 GLOBAL COMMUNICATIONS NEWSLETTER
- 224 ADVERTISERS' INDEX

5G RADIO ACCESS ARCHITECTURE AND TECHNOLOGIES

GUEST EDITORS: DAVID SOLDANI, PERIKLIS CHATZIMISIOS, ABBAS JAMALIPOUR, BERNARD BARANI, SIMONE REDANA, AND SUNDEEP RANGAN

- 14 GUEST EDITORIAL
- 16 A SCALABLE AND FLEXIBLE RADIO ACCESS NETWORK ARCHITECTURE FOR FIFTH GENERATION MOBILE NETWORKS
Andreas Maeder, Amaanat Ali, Anand Bedekar, Andrea F. Cattoni, Devaki Chandramouli, Subramanya Chandrashekar, Lei Du, Matthias Hesse, Cinzia Sartori, and Samuli Turtinen
- 24 5G RADIO ACCESS NETWORK ARCHITECTURE: DESIGN GUIDELINES AND KEY CONSIDERATIONS
Patrick Marsch, Icaro Da Silva, Ömer Bulakci, Milos Tesanovic, Salah Eddine El Ayoubi, Thomas Rosowski, Alexandros Kaloxylas, and Mauro Boldi
- 33 SPECTRUM POOLING IN MMWAVE NETWORKS: OPPORTUNITIES, CHALLENGES, AND ENABLERS
Federico Boccardi, Hossein Shokri-Ghadikolaei, Gabor Fodor, Elza Erkip, Carlo Fischione, Marios Kountouris, Petar Popovski, and Michele Zorzi
- 40 INITIAL ACCESS IN 5G MMWAVE CELLULAR NETWORKS
Marco Giordani, Marco Mezzavilla, and Michele Zorzi
- 48 THE NEW FRONTIER IN RAN HETEROGENEITY: MULTI-TIER DRONE-CELLS
Irem Bor-Yaliniz and Halim Yanikomeroglu
- 56 ENERGY CONSUMPTION MINIMIZATION FOR F5G ENHANCED LTE-A HETNETS WITH UE CONNECTION CONSTRAINT
Jiajia Liu, Hongzhi Guo, Zubair Md. Fadlullah, and Nei Kato

NEW WAVEFORMS FOR 5G NETWORKS

GUEST EDITORS: CHARLIE JIANZHONG ZHANG, JIANGLEI MA, GEOFFREY YE LI, WEI YU, NIHAR JINDAL, YOSHIHISA KISHIYAMA, AND STEFAN PARKVALL

- 64 GUEST EDITORIAL
- 66 INTRODUCTION TO QAM-FBMC: FROM WAVEFORM OPTIMIZATION TO SYSTEM DESIGN
Chanhong Kim, Yeo Hun Yun, Kyeongyeon Kim, and Ji-Yun Seol
- 74 ON THE WAVEFORM FOR 5G
Xi Zhang, Lei Chen, Jing Qiu, and Javad Abdoli
- 82 INTERFERENCE MANAGEMENT VIA SLIDING-WINDOW CODED MODULATION FOR 5G CELLULAR NETWORKS
Kwang Taik Kim, Seok-Ki Ahn, Yong-Seok Kim, Jeongho Park, Chiao-Yi Chen, and Young-Han Kim
- 90 WAVEFORM AND NUMEROLOGY TO SUPPORT 5G SERVICES AND REQUIREMENTS
Ali A. Zaidi, Robert Baldemair, Hugo Tullberg, Hakan Björkegren, Lars Sundström, Jonas Medbo, Caner Kilinc, and Icaro Da Silva
- 99 GENERALIZED DFT-SPREAD-OFDM AS 5G WAVEFORM
Gilberto Berardinelli, Klaus I. Pedersen, Troels B. Sørensen, and Preben Mogensen

2016 IEEE Communications Society Elected Officers

Harvey A. Freeman, *President*
Luigi Fratta, *VP-Technical Activities*
Guoliang Xue, *VP-Conferences*
Stefano Bregni, *VP-Member Relations*
Nelson Fonseca, *VP-Publications*
Robert S. Fish, *VP-Industry and Standards Activities*
Sergio Benedetto, *Past President*

Members-at-Large

Class of 2016

Sonia Aissa, Hsiao Hwa Chen
Nei Kato, Xuemin Shen

Class of 2017

Gerhard Fettweis, Araceli García Gómez
Steve Gorshe, James Hong

Class of 2018

Leonard J. Cimini, Tom Hou
Robert Schober, Qian Zhang

2016 IEEE Officers

Barry L. Shoop, *President*
Karen Bartleson, *President-Elect*
Parviz Famouri, *Secretary*
Jerry L. Hudgins, *Treasurer*
Howard E. Michel, *Past-President*
E. James Prendergast, *Executive Director*
Celia Desmond, *Director, Division III*

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

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

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

COPYRIGHT AND REPRINT PERMISSIONS: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limits of U.S. Copyright law for private use of patrons: those post-1977 articles that carry a code on the bottom of the first page provided the per copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For other copying, reprint, or republication permission, write to Director, Publishing Services, at IEEE Headquarters. All rights reserved. Copyright © 2016 by The Institute of Electrical and Electronics Engineers, Inc.

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 Erie, ON L2A 6C7.

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

ADVERTISING: Advertising is accepted at the discretion of the publisher. Address correspondence to: Advertising Manager, *IEEE Communications Magazine*, 3 Park Avenue, 17th Floor, New York, NY 10016.

SUBMISSIONS: The magazine welcomes tutorial or survey articles that span the breadth of communications. Submissions will normally be approximately 4500 words, with few mathematical formulas, accompanied by up to six figures and/or tables, with up to 10 carefully selected references. Electronic submissions are preferred, and should be 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>. For further information contact Zoran Zvonar, Associate Editor-in-Chief (zoran.zvonar@mediatek.com). All submissions will be peer reviewed.



106 FLEXIBLE DFT-S-OFDM: SOLUTIONS AND CHALLENGES

Alphan Sahin, Rui Yang, Erdem Bala, Mihaela C. Beluri, and Robert L. Olesen

COMMUNICATIONS EDUCATION AND TRAINING: EDUCATIONAL SERVICES BOARD

GUEST EDITORS: RULEI TING, DAVID G. MICHELSON, AND MICHELE ZORZI

114 GUEST EDITORIAL

116 EDUCATION AND TRAINING IN COMSOC: RECENT ACHIEVEMENTS

Michele Zorzi

121 THE COMSOC HANDS-ON LAB EXCHANGE

Rhys Bowley, Erik Luther, and David G. Michelson

124 TRAINING AND NETWORKING FOR YOUNG SOCIETY MEMBERS: THE COMSOC SUMMER SCHOOL PROGRAM

Fabrizio Granelli

130 INTEGRATING THE SCHOLARSHIP OF TEACHING, LEARNING, AND RESEARCH SUPERVISION INTO COMMUNICATIONS EDUCATION

David G. Michelson

GREEN COMMUNICATIONS AND COMPUTING NETWORKS

SERIES EDITORS: JINSONG WU, JOHN THOMPSON, HONGGANG ZHANG,
RANGARAO VENKATESHA PRASAD, AND SONG GUO

134 SERIES EDITORIAL

136 GREEN TOUCHABLE NANOROBOTIC SENSOR NETWORKS

Yifan Chen, Tadashi Nakano, Panagiotis Kosmas, Chau Yuen, Athanasios V. Vasilakos, and Muhamad Asvial

143 SIMULTANEOUS INFORMATION AND ENERGY FLOW FOR IOT RELAY SYSTEMS WITH CROWD HARVESTING

Weisi Guo, Sheng Zhou, Yunfei Chen, Siyi Wang, Xiaoli Chu, and Zhisheng Niu

150 LIFETEL: MANAGING THE ENERGY-LIFETIME TRADE-OFF IN TELECOMMUNICATION NETWORKS

Luca Chiaraviglio, Lavinia Amorosi, Andrea Baiocchi, Antonio Cianfrani, Francesca Cuomo, Paolo Dell'Olmo, and Marco Listanti

158 DYNAMIC ENERGY TRADING FOR WIRELESS POWERED COMMUNICATION NETWORKS

Yong Xiao, Dusit Niyato, Ping Wang, and Zhu Han

166 POWER-SAVING METHODS FOR INTERNET OF THINGS OVER CONVERGED FIBER-WIRELESS ACCESS NETWORKS

Dung Pham Van, Bhaskar Prasad Rimal, Jiajia Chen, Paolo Monti, Lena Wosinska, and Martin Maier

176 SUSTAINABILITY INFORMATION MODEL FOR ENERGY EFFICIENCY POLICIES

Ana Carolina Riekstin, Bruno Bastos Rodrigues, Viviane Tavares Nascimento, Claudia Bianchi Progetti, Tereza Cristina Melo de Brito Carvalho, and Catalin Meirosu

185 SOFTWARE DEFINED NETWORKING, CACHING, AND COMPUTING FOR GREEN WIRELESS NETWORKS

Ru Huo, Fei Richard Yu, Tao Huang, Renchao Xie, Jiang Liu, Victor C.M. Leung, and Yunjie Liu

194 GREEN DATAPATH FOR TCAM-BASED SOFTWARE-DEFINED NETWORKS

Huawei Huang, Song Guo, Jinsong Wu, and Jie Li

202 TOWARD THE DEVELOPMENT OF A TECHNO-SOCIAL SMART GRID

S. N. Akshay Uttama Nambi and R. Venkatesha Prasad

ACCEPTED FROM OPEN CALL

210 IEEE 5G SPECTRUM SHARING CHALLENGE: A PRACTICAL EVALUATION OF LEARNING AND FEEDBACK

Sreeraj Rajendran, Bertold Van den Bergh, Tom Vermeulen, and Sofie Pollin

217 NEW TECHNOLOGIES AND TRENDS FOR NEXT GENERATION MOBILE BROADCASTING SERVICES

Alejandro de la Fuente, Raquel Pérez Leal, and Ana García Armada

INNOVATE FASTER

WITH FIELD-DEPLOYED 5G PROOF-OF-CONCEPT SYSTEMS

In the race to design next-generation wireless technologies, research teams must rely on platforms and tools that accelerate their productivity. Using the NI software defined radio platform and LabVIEW Communications, leading researchers are innovating faster and building 5G proof-of-concept systems to demonstrate new technologies first.

Accelerate your innovation at ni.com/5g



LabVIEW Communications System Design Software, USRP-2943R SDR Hardware



THREE-PHASE APPROACH TO INVESTING IN COMSOC'S FUTURE

Volunteers and staff create prosperity for ComSoc and its members through dedicated efforts including conferences, workshops, publications, standards, certification, training, education, and intellectual property. This prosperity is carefully managed and used to keep dues as low as possible and to invest in the future. In order to maintain positive momentum, ComSoc uses a three-phase approach to securing its future: careful financial management, investing in the future, and matching membership benefits with the dues structure. This approach is described in this article by Robert C. (Bob) Shapiro who is the Society's current Treasurer.

Bob Shapiro is an Independent Land Mobile Radio and Wireless Engineer and Consultant serving the Public Safety, Electric Utility, Energy, Transportation and Telecommunications industries based in Plano, Texas, USA. Besides being the ComSoc Treasurer, he is the TIA TR8.18 Sub-committee Vice Chair and elected Board of Governor and Private Land Mobile Radio Committee Chair for the IEEE Vehicular Technology Society and Director-Elect of IEEE Region 5. His past IEEE leadership volunteer activities were ComSoc BOG Member at Large, Director, North American and Marketing/Industry Relations, and IEEE Dallas CVT Chapter Chair. He has also served as Vice President-Conferences for the IEEE Technology Management Council and Vice President-Mobile Radio for the IEEE Vehicular Technology Society and IEEE Region 5 South Area and Membership Chair and Secretary. Bob has held his Professional Engineering Certification in Texas since 1990. He is a Fellow of the Radio Club of America and a Member of the IEEE since 1981 and Senior Member since 2003.

CAREFUL FINANCIAL MANAGEMENT TO ENSURE SUCCESS

ComSoc employs consistent financial planning to ensure success today and for the future. The ComSoc Finance Committee, or FINCOM, is charged to protect, preserve, and invest in the future of the Society. To this end, the FINCOM meets at the semi-annual Board of Governors meetings to plan the budgets and forecasts and, with guidance from the volunteer Treasurer, Bob Shapiro, and Staff Director of Finance and Business Operations, Bruce Worthman, sets the course forward for the financial planning.

Budgets are submitted from the FINCOM to the Board of Governors for approval and then sent to IEEE Technical Activities. IEEE Technical Activities and ComSoc follow a calendar fiscal year. The yearly business cycle starts in Jan-



Harvey Freeman



Robert (Bob) C. Shapiro

uary at the Management Retreat and uses a baseline pro-forma budget from the Director of Finance and Business Operations. This baseline budget reflects many years of historical data and adds a collaboration with the President and Vice Presidents to determine income and expense estimates for the year.

The budget is prepared after inputs are received and reviewed at the Operations Committee meeting normally held in March or April. The budget is edited and modified with further input from the stakeholders and submitted to the FINCOM around one month prior to the Board of Governors meeting in May or June. The FINCOM meets, adjusts, then approves the budget to be submitted to the Board of Governors and, if approved at this level, is submitted to the Technical Activities FINCOM for approval and rollout to all of Technical Activities.

This budget includes income from membership, conferences, publications, and IEEE Intellectual Property sales and expenses such as membership programs, conferences, publications, and IEEE and ComSoc overhead. Mapping income to expenses is used to load balance money flows and helps determine dues levels and what amounts to charge for conference attendance and publications and other services. This process also determines the amount to spend on ComSoc staff, other overhead, volunteer meetings and travel, and programs benefiting our members.

The most important transaction, leading eventually to future revenues, is the investment into ComSoc's new areas of interest and initiatives. Conference and publications surpluses and revenue from IEEE intellectual property sales are used to first offset costs to run the Society, and then to invest to ensure current levels of revenue are at least maintained, if not growing.

INVESTING TO INCUBATE AND DEVELOP NEW INITIATIVES

Many programs are currently funded to prepare ComSoc's future such as the 5th Generation Mobile Networks, Fog Computing, Internet of Things, RFID, Sensors, and industry events that extend the reach of the Society to its local chapters with practitioner marketed content.

The 5th Generation Mobile Networks initiatives are now integrated into the Society's conferences and publications. This effort reaped the benefits of many years of incubation, development, and investment, and is now a primary focus in the wireless industry. ComSoc is in the forefront of conferences, publications, and most importantly, leadership in

local workshops. The 5G Summits are a big success. These summits are typically full day events around the globe and are prepared with local volunteer and chapter support and marketed toward the practitioner.

The OpenFog Consortium is a collaboration of industry and academia with ComSoc square in the middle as the only non-university professional society. From the OpenFog Consortium website, "Fog computing is a system-level horizontal architecture that distributes resources and services of computing, storage, control, and networking along the continuum from Cloud to Things such as Horizontal Architecture, Cloud-to-Thing Continuum of Services, and System Level Architecture." In addition to membership in the Consortium, ComSoc has its own initiative in Fog Computing for which it is currently seeking IEEE initiative funding. ComSoc is in a leadership position for content, similar to the 5th Generation Mobile Networks initiatives.

The Internet of Things, combining the 5th generation of mobile networks and other wireless protocols and Fog Computing, is the connectivity of physical devices such as vehicles, appliances, buildings, and people using low-power sensors. The Internet of Things Initiative within IEEE is now led by ComSoc, with the goal to bring clarity and disseminate information globally. The Internet of Things Initiative serves as the home for the global community of engineering and technical professionals within industry, academia, and government to bring together content including conferences, publications, videos, articles, standards, webinars, workshops, and a web portal. Similar to Fog Computing, the investments in the leadership of the Internet of Things Initiative will drive content and engineers to ComSoc as the leader in the field, and as the initiative grows, the Society will grow also.

The Council on Radio Frequency Identification is an IEEE focus area on radio frequency identification and the Internet of Things. ComSoc has been a member Society since the Council was a Committee and continues to support the initiative. The Council on Radio Frequency Identification has two international conferences, RFID and RFID Technical Applications in addition to workshops, conference program tracks, and technical forums to foster technical exchanges on Radio Frequency Identification. ComSoc supports the Council due to the inter-workings of wireless technologies, Fog Computing, Internet of Things, Radio Frequency Identification, and Sensors. This relationship helps support technical communities and chapters where there are mutual benefits to collaborate.

The Sensors Council, similar to the Council on Radio Frequency Identification, is an IEEE focus area on the theory, design, fabrication, manufacturing, and applications of devices for sensing, and is a cornerstone of the Internet of Things. ComSoc is a member of the Council and continues to support the tracks that coincide with those in its conferences, publications, and events. The Sensors Council has a flagship conference and two journals: the *IEEE Sensors Journal* and the *Internet of Things Journal*. Similar to the IEEE Council on Radio Frequency Identification, ComSoc supports the Council due to the inter-workings of wireless technologies, Fog Computing, Internet of Things, and Radio Frequency Identification. This relationship also helps support technical communities and chapters where there are mutual benefits to collaborate.

Industry events allow ComSoc to reach industry partners and their practitioners. These events, in the form of webinars, online collaborations, and communities, include workshops, awards and recognition, research and development networking and outreach, applications development focus groups, and product showcases. Including industry and practitioners, especially in local communities, increases the viability of chapters around the globe and allows more members to connect with the Society.

MATCHING MEMBERSHIP DUES WITH AN EVER INCREASING BENEFITS OFFERING

Benefits and features offered to members of ComSoc are numerous, and we are continuing to offer new benefits regularly, from distinguished lecturers to chapter support, from discounts for meetings and conferences to online and in-person educational opportunities, from student travel grants to awards and recognition, from workshops to student competitions, and from mentoring to young professionals.

ComSoc works closely with its chapters to coordinate activities by providing funding and helping to coordinate distinguished lecturers. The chapters are the heartbeat of the Society and where most of the members interact. Investing in the success of the chapters is taken very seriously, and funds are budgeted each year for as many activities as possible across more than 150 chapters worldwide.

Most IEEE and all ComSoc conferences, meetings, workshops, tutorials, and training and certification sessions offer discounts for members to attend. These discounts, if taken, pay for membership in one or two sessions. Also offered to ComSoc members are student travel grants. These grants help student authors meet the costs to attend and present their papers or poster boards at ComSoc events.

The Awards Committees is tasked with identifying and recognizing members' accomplishments. There are numerous awards that are funded from the Society's surpluses to provide travel grants and in some cases, monetary rewards. Some of the award categories are geared toward industry and practitioners, such as the following career awards: Education, Edwin Howard Armstrong Achievement, Distinguished Industry Leader, and Industry Innovation and Public Service in the Field of Telecommunications and Service. Service awards include the Donald W. McLellan Meritorious Service, Harold Sobol Award for Exemplary Service to Meetings and Conferences, Joseph LoCicero Award for Exemplary Service to Publications, and the COMSOC/KICS Exemplary Global Service.

All in all, considering the benefits offered, membership dues are affordable while investing in future events and prudent fiscal management help to keep dues low while increasing these benefits. The costs to maintain a membership base, manage these benefits, and run programs far exceeds the cost of dues. The remainder of the funds come from conferences and publications and IEEE intellectual property along with the returns on the investments made in the other activities mentioned in this article. ComSoc is committed to investing in the future and keeping dues in check, and will use as many tools as possible to maintain an equitable system.

INDUSTRIAL NETWORK SECURITY: SECURING CRITICAL INFRASTRUCTURE NETWORKS FOR SMART GRID, SCADA, AND OTHER INDUSTRIAL CONTROL SYSTEMS

By Eric D. Knapp and Joel T. Langill, Elsevier, 2015, Second Edition, ISBN 978-0-12-420114-9, 439 pages

Reviewer: Marcin Jekot

This book presents a holistic view of the overall security of Industrial Control Systems (ICS). The author considers all the key aspects of ICS security: network, protocol and application characteristics, regulatory compliance issues, an approach to risk assessment and management of the aforementioned systems, etc.

This book has a well-organized and readable form and consists of 13 chapters followed by three appendices providing background material to the matters described beforehand. The structure of the book and logical flow allows for easy understanding of the presented ideas. What is more, particular chapters can be read independently, allowing for familiarization only with the specific aspect of interest, such as security assessment of ICS or the related network protocols.

The first five chapters following Chapter 1 (Introduction) provide the basics of ICS. Definitions from both the cyber-security (e.g., defense in depth, access control) and ICS (such as smart grids or industrial protocols) worlds are introduced. Technical concepts of network and architecture along with the respective industrial network protocols are discussed in Chapters 4-6. Chapter 3 describes the history and evolution of cyber threats to industrial systems with examples dating as far back as the 1902 attack on the Marconi Wireless Telegraph system. The second part of the chapter describes modern trends presently observable in the industry (APTs, cyber-warfare, hacktivism, etc.).

Chapters 7 and 8 focus on the possible attack vectors for the particular components present in ICSs and provide a description of methods employed to detect and triage related vulnerabilities and risks. The author delivers not only a detailed breakdown of the attack targets along with the possible consequences, but also elaborate on lessons learned from past incidents (Stuxnet), giving an overview of how a change in reasoning is required with respect to the current ICS security posture. The analysis presented considers usage of

weaponized malware by state-sponsored agencies. Chapter 8 focuses on the formal side of security testing and assessment activities, providing basic information about risk elements. Additionally, it gives valuable input for developing customized testing and assessment methodologies for ICS.

The next four chapters are devoted to building and maintaining a secure ICS that can withstand attacks described in the prior sections. Chapters 9 and 10 give a detailed insight of the concept of security zones, along with a set of security and access controls that should be implemented. The application of secure zones in ICS is overviewed, considering specific constraints, such as the use of particular protocols including DNP3 for SCADA, plan level control processes, etc. Chapters 11 and 12 describe various aspects of monitoring system operations and reporting on suspicious behavior, including reporting on exceptions from established policies and anomaly detection, as well as best practices in operational and procedural security monitoring. The important problem related to information overflow associated with excessive monitoring is discussed, and suggestions on developing appropriate log retention and collection processes are described.

The last chapter is an extensive description of standards and best practices that can be used to develop and audit ICS infrastructures. High-level recommendations of required security controls for compliance with a given standard are also listed.

The most important advantage of the book is its duality, coupling the worlds of IT and operational technology security, providing insights into baselines of both worlds, and allowing professionals coming from both these areas to speak in a common language. Real examples of how standard IT security correlates with operational industrial systems makes it easier to understand both areas. The author has done solid work by summarizing the overall security approach to ICS, not focusing solely on a particular technology or use case. Therefore, the ideas presented can be useful for securing a variety of ICSs, e.g., supporting smart grids, intelligent buildings, or car manufacturers.

This book is worth recommendation for people who are interested in modern industry control systems security. Additionally, it will be advantageous for university researchers and gradu-

ate students in the network security field, as well as to the industry specialists in the area of ICS. The book will be especially beneficial for individuals who are already familiar with IT security, but need guidance in terms of applying appropriate security paradigms in the field of ICS. Additionally, as highlighted by the authors, compliance officers who are responsible for ensuring that their systems meet particular regulatory or internal standards, can use the book as a high level baseline for planning and implementing necessary cyber security controls in a way that enables them to fulfil security audit requirements.

CLOUD SERVICES, NETWORKING, AND MANAGEMENT

Edited by Nelson L. S. da Fonseca and Raouf Boutaba, Wiley, 2015, ISBN 978-1-118-84594-3, hardcover, 407 pages

Reviewer: Piotr Borylo

Cloud Computing is a topic that is attracting the attention of the vast majority of scientists around the world. The impact of Cloud Computing on everyday life motivates researchers to address the issue from different perspectives. What is more, Cloud Computing is inseparably coupled with other popular topics, such as: data centers, Software-Defined Networking (SDN), mobility, energy consumption, risk management, security, and Big Data. Most of the aforementioned issues are addressed in the book edited by Fonseca and Boutaba, which makes it relevant and noteworthy, especially as the content is not only up-to-date but also comprehensive and well structured.

The book is divided into 15 chapters grouped into four logical parts. Chapter 1 describes the reasons for Cloud Computing's popularity, basic definitions, and cloud-enabling technologies. Chapters 2 and 3 cover virtualization and migration of virtual instances. These chapters address virtualization requirements on advanced management functions, security improvements, and consistent interfaces applicable in heterogeneous infrastructures. Chapter 4 describes the data center network infrastructure. Innovative technologies supporting incremental expansion, Layer 2 and Layer 3 addressing, as well as data center-specific traffic profiles are also addressed. The book additionally considers inter-data-center networks in Chapter 5. Optical data transmission is indicated as the most prominent

technology, exemplary optimization problems are formulated, and energy-aware aspects are thoroughly investigated. In Chapter 6, SDN and OpenFlow are thoroughly introduced as a perfect environment for modern cloud applications. The complete framework for cloud and OpenFlow cooperation is additionally proposed. The risk management approach is utilized to study the issue of mobile Cloud Computing in Chapter 7. The next chapter focuses on optimization of energy consumption in data centers. Models for energy consumption of various components are described and solutions to enhance energy efficiency of cloud data centers are provided. In Chapter 9, improvements in management are indicated as necessary to support the transition from private clouds to multi-tenant clouds. The X-Cloud Application Management Platform is proposed and exhaustively assessed on the basis of a created testbed. A formal problem statement for application sched-

uling and virtual machine allocation followed by exemplary solutions are provided in Chapter 10. Cloud-based Intrusion Detection/Prevention Systems are introduced in Chapter 11 along with secure cloud design issues and requirements for secure clouds. Additionally, FlowIPS is proposed as the authors' solution based on the Open vSwitch. Chapter 12 forms a survey on cloud survivability and considers availability-aware resource allocation schemas. Scientific workflow scheduling in the cloud is considered in the subsequent chapter, where a novel scheduler is also proposed by the authors. Challenging multimedia services offered in the cloud are analyzed for requirements and delivery models in Chapter 14. The exemplary multimedia services embrace: cloud-gaming, user-generated live-streaming, and time-shifting on-demand TV. Finally, Chapter 15 considers Big Data in the context of Cloud Computing, which serves simultaneously as supply and demand of Big Data.

Each chapter of the book is a separate self-contained part. The length of each chapter is adequately selected, while the balance between an introduction for beginners and details for advanced readers is excellently preserved. For a majority of the covered topics, a comprehensive survey and thoughtful taxonomy are provided. The presented issues are followed by the exemplary solutions, which are not only conceptual, but very often describe practical deployments. Furthermore, the reference lists are complete, while indicated directions for further studies are relevant and valuable. The only minor drawback is that some chapters repeat fundamental information already mentioned in other preceding chapters.

Summarizing, I recommend this book as a source of very relevant and valuable information about popular and emerging cloud-related topics. An additional advantage is that security and energy efficiency are also addressed.

**IEEE
ComSoc™**

Join our Community!
www.comsoc.org



THE GLOBAL COMMUNITY OF COMMUNICATIONS PROFESSIONALS



Member Benefits

IEEE Communications Magazine (electronic & digital delivery)

IEEE Communications Surveys and Tutorials (electronic)

Online access to IEEE Journal of Lightwave Technology, IEEE OSA Journal of Optical Communications and Networking and IEEE RFID Virtual Journal

Member Discounts

Valuable discounts on conferences, publications, IEEE WCET Certification program, IEEE Training courses and other exclusive member-only products.

These membership and exclusive benefits will expand your technical community and valuable networking opportunities. Join Today!

UPDATED ON THE COMMUNICATIONS SOCIETY'S WEB SITE
www.comsoc.org/conferences

2016

NOVEMBER

MILCOM 2016 — Military Communications Conference, 1–3 Oct.

Baltimore, MD
<http://events.afcea.org/milcom16/Public/enter.aspx>

IEEE SmartGridComm — IEEE Int'l. Conference on Smart Grid Communications, 6–9 Nov.

Sydney, Australia
<http://sgc2016.ieee-smartgridcomm.org/>

IEEE ANTS 2016 — IEEE Int'l. Conference on Advanced Networks and Telecommunications Systems, 6–11 Nov.

Bangalore, India
<http://ants2016.ieee-comsoc-ants.org/2016/02/17/about-ants-2016/>

IEEE RIVF 2016 — IEEE RIVF International Conference on Computing & Communication Technologies, 7–9 Nov.

Hanoi, Vietnam
<http://rivf2016.tlu.edu.vn/>

IEEE NFV-SDN 2016 — IEEE Conference on Network Function Virtualization and Software Defined Networks, 7–10 Nov.

Palo Alto, CA
<http://nfvsdn2016.ieee-nfvsdn.org/>

FRUCT19 2016 — 19th Conference of Open Innovations Association FRUCT, 7–11 Nov.

Jyvaskyla, Finland
<http://fruct.org/cfp>

ANIL-FRUCT 2016 — Artificial Intelligence and Natural Language FRUCT 2016 Conference, 13–16 Nov.

St. Petersburg, Russia
<http://ainlconf.ru/>

WPMC 2016 — Int'l. Symposium on Wireless Personal Multimedia Communications, 13–16 Nov.

Shenzhen, China
<http://www.wpmc2016.org/>

ITU-K 2016 — ITU Kaleidoscope: ICTs for a Sustainable World, 14–16 Nov.

Bangkok, Thailand
<http://www.itu.int/en/ITU-T/academia/kaleidoscope/2016/Pages/default.aspx>

IEEE OnlineGreenComm — IEEE Online Conference on Green Communications, 14–16 Nov.

Online
<http://onlinegreencomm2016.ieee-online-greencomm.org/>

NOF 2016 — Int'l. Conference on the Network of the Future, 16–18 Nov.

Buzios, Brazil
<http://www.network-of-the-future.org/>

NTMS 2016 — IFIP Int'l. Conference on new Technologies, Mobility and Security, 21–23 Nov.

Larnaca, Cyprus
<http://www.ntms-conf.org/ntms2016/>

IFIP PEMWN 2016 — IFIP Int'l. Conference on Performance Evaluation and Modeling in Wired and Wireless Networks, 22–25 Nov.

Paris, France
<https://sites.google.com/site/pemwn2016/>

CloT 2016 — Cloudification of the Internet of Things, 23–25 Nov.

Paris, France
<http://www.dnac.org/DNAC/iot/>

D E C E M B E R

IEEE GLOBECOM 2016 — 2016 IEEE Global Communications Conference, 4–8 Dec.

Washington, DC
<http://globecom2016.ieee-globecom.org/>

IEEE VNC 2016 — IEEE Vehicular Networking Conference, 8–10 Dec.

Columbus, OH
<http://www.ieee-vnc.org/>

IEEE WF-IOT — IEEE World Forum on Internet of Things, 12–14 Dec.

Reston, VA
<http://wfiot2016.ieee-wf-iot.org/>

ICT-DM 2016 — Int'l. Conference on Information and Communication Technologies, 13–15 Dec.

Vienna, Austria
<http://ict-dm2016.ait.ac.at/>

IEEE ICCS 2016 — IEEE Int'l. Conference on Communication Systems, 14–16 Dec.

Shenzhen, China
<http://www.ieee-iccs.org/>

ICSPCS 2016 — Int'l. Conference on Signal Processing and Communication Systems, 19–21 Dec.

Surfers Paradise, Australia
http://www.dspscs-witps.com/icspcs_2016/index.html

SCNS 2016 — Smart Cloud Networks & Systems Workshop, 19–21 Dec.

Dubai, UAE
<http://www.scns-workshop.org/>

2017

J A N U A R Y

COMSNETS 2017 — Int'l. Conference on Communication Systems & Networks, 4–8 Jan.

Bangalore, India
<http://www.comsnets.org/>

IEEE CCNC 2017 — IEEE Consumer Communications and Networking Conference, 8–11 Jan.

Las Vegas, NV
<http://ccnc2017.ieee-ccnc.org/>

ICNC 2017 — Int'l. Conference on Computing, Networking and Communications, 26–29 Jan.

Santa Clara, CA
<http://www.conf-icnc.org/2017/>

F E B R U A R Y

ICACT 2017 — Int'l. Conference on Advanced Communication Technology, 19–22 Feb.

Pyeongchang, Korea
<http://www.icaact.org/>

–Communications Society portfolio events appear in bold colored print.

–Communications Society technically co-sponsored conferences appear in black italic print.

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



November 2016

ISSN 2374-1082

SISTER AND RELATED SOCIETIES

Sister and Related Societies: Reaching Out to ComSoc's Global Community

Interview with Curtis Siller, Director of Sister and Related Societies

By Stefano Bregni, Vice-President for Member and Global Activities, and Curtis Siller, Director of Sister and Related Societies

Following the series of articles published about two years ago during my previous term as Vice-President for Member Relations, with this issue we begin a new series of eight interviews with the Directors of the IEEE ComSoc Member and Global Activities Council, which will be published every month in the Global Communications Newsletter.

In this series of articles, I will introduce the six Directors on the Member and Global Activities Council (namely: Sister and Related Societies; Membership Services; AP, NA, LA, EMEA Regions), and the Chairs of the Women in Communications Engineering (WICE) and Young Professionals (YP) Standing Committees. They will present their sector activities and plans.

Opening the series, this month we begin with Curtis Siller, Director of Sister and Related Societies. Curtis has served on the IEEE Communications Society Board of Governors for 18 years. Among his service to our society, he was President (2004–2005), a Director for three terms and a Vice-President for two terms. He was an IEEE Division Director and Editor-in-Chief of IEEE Communications Magazine in 1993–1995. Curtis is an IEEE Life Fellow. He worked for more than 30 years at Bell Labs (where he was named a Bell Labs Fellow), then later in several other high-tech and consulting companies.

Bregni: Curtis, let us begin by explaining what are the Sister and Related Societies of ComSoc.

Siller: The IEEE Communications Society has a long tradition of global outreach. Dating back more than 20 years, ComSoc instituted a novel program of entering into relationships with Sister and Related Societies. Sister Societies are those that have a charter that overlaps ComSoc's technical scope. These are usually national professional societies. Related Societies are those whose focus complements but does not overlap ComSoc's technical orientation.

Bregni: Are Sister and Related Societies important elements in ComSoc's strategy toward globalization?

Siller: They certainly are! Both are important ingredients in ComSoc's international initiative. Allow me a moment to share a few benchmarks. As a robust Society, ComSoc is among IEEE's most prominent Organizational Units, with a membership that exceeds thousands and is experiencing membership growth.

Notably, 2004 was a very significant year. It was then that our Society transitioned to having more international members than those in the United States. Today, less than half of our member-

ship is in the United States. More are international, with especially significant growth in Asia and Southeast Asia. By way of contrast, IEEE, our parent organization, attained a majority of international membership in late 2011/early 2012, nearly eight years later.

Bregni: To how many Sister Societies worldwide is ComSoc linked?

Siller: A full list of Sister and Related Societies is found at www.comsoc.org/about/sistersocieties

There you will find that ComSoc has reached out to more than 30 Sister Societies worldwide and a variety of Related Societies. Each of those societies adds to our global community! In early 1995, six societies entered into these agreements. Since then, the number has grown several times, as noted above.

Bregni: How does ComSoc reach out to Sister and Related Societies? Are our relationships regulated in some way? And what types of activities are mainly addressed by cooperation?

Siller: There are at least two essential elements that bind us to these societies: conferences and publications. We'll share more about that in a moment. The bases for these relationships are two Memoranda of Understanding (MOUs). Generally, these span two to four years, with the ComSoc President and a corresponding Sister or Related Society president as signatories. Essential ingredients include: submission of papers to ComSoc conferences, journals, transactions, and magazines; discounted participation at our conferences; discounted subscriptions to ComSoc publications; affiliate memberships in ComSoc; streamlined technical co-sponsorship of conferences; participation in ComSoc technical committees; and advertising promotions.

Bregni: Tell us something more about Related Societies.

Siller: In addition to Sister Society agreements, ComSoc enjoys Related Society agreements. These naturally include operational units within IEEE, such as the Signal Processing, Computer, Circuits and Systems, and Power and Energy Societies. Among those outside of the IEEE, let me note the East-West Institute, a "think tank" in New York City that is dedicated to facilitating world harmony, and the Pacific Telecommunications Council.

Bregni: In conclusion, would you share with us your plans going forward as Director of Sister and Related Societies?

Siller: The opportunities are numerous. MOUs that are expiring need to be renewed, and additional societies, not yet among those noted above, need to be identified. Additionally, elements will be added to these agreements to make them more meaningful to the Sister and Related Societies and for ComSoc. Further, in the past I think there has been little direct, personal contact between ComSoc members and individuals in our reciprocal organizations. One way to enhance these contacts is to offer social events at ComSoc's most notable international conference and meeting venues so we can greet each other on a personal level and discuss ideas that would bond us more fully.

ComSoc celebrates all of these relationships. They are essential to our presence in the world community. Please contact me if you know of other societies that might invite our engagement.



Stefano Bregni



Curtis Siller

Highlights from IEEE HPSR 2016: 17th International Conference on High Performance Switching and Routing

By Naoaki Yamanaka, General Co-Chair, Keio University, Japan;
Eiji Oki, TPC Co-Chair, The University of Electro-Communications,
Tokyo, Japan

The IEEE 17th International Conference on High Performance Switching and Routing (HPSR 2016) was held at Keio University, Yokohama, Japan, on 14-17 June, 2016. Yokohama is a major port city to the south of Tokyo. The conference was sponsored by the IEEE Communications Society and the IEICE Communications Society, and was co-located with 12th International Conference on IP+Optical Network (iPOP). The conference was supported by the IEICE Photonic Network Technical Committee, the National Institute of Information and Communications Technology, Japan, the Support Center for the Advanced Telecommunications Technology Research Foundation, Japan, and the Yokohama Convention & Visitors Bureau, Japan.

HPSR addresses numerous challenges of today's data networks, which are being subjected to significant changes driven by cloud computing, the Internet of Things, and other new concepts. As a result, new technologies are needed to efficiently and effectively cope with the resulting traffic demands. This conference brought together researchers from around the world to present the latest advances in the fields of high-performance switching and routing. The participants discussed switching and routing capabilities that ought to be more intelligent, efficient, and reliable than ever before.

The conference program included a rich technical program comprising 31 excellent technical full-paper presentations and eight poster-paper presentations, three keynote speeches, one invited speech, four tutorials, technical tours, and workshop presentations. The number of participants was 101 (including 69 from academia, 26 from industry, and four from government, among others).



Technical visit.



Conference lunch.



HPSR Best Paper Award ceremony at banquet. From left to right: Naoaki Yamanaka, General Co-Chair, Andrew Mundy, Award Recipient, and Eiji Oki, TPC Co-Chair.



Plenary session at Fujiwara Memorial Hall.



Walking tour in Yokohama bay area.

A total of 80 eligible papers were submitted from 26 countries, including Asia, North America, and Europe. The submitted papers were carefully peer-reviewed by our Technical Program Committee (TPC). Each paper received at least three reviews, thus hopefully providing valuable feedback to the authors and ensuring high confidence in the outcome of the review process. In total, the TPC completed 302 reviews, for an average of 3.8 reviews per paper. For each track, the accepted papers were selected based on all the review results, including reviewers' comments.

On Tuesday, three tutorials were offered. Abbas Jamalipour (University of Sydney, Australia) spoke about scaling dense-traffic cellular networks through software defined networking. Noriaki Kamiyama (Osaka University & NTT Network Technology Laboratories, Japan) presented a tutorial on advances in reducing Web response time. Dimitri Papadimitriou (Nokia - Bell Labs, Belgium) reviewed a number of open challenges in network optimization. Each tutorial attracted approximately 70-80 attendees.

On Wednesday afternoon, two keynotes were delivered at the HPSR and iPOP joint plenary session. Before the two keynotes, a piano concert was held. Rutsuko Yamagishi, renowned pianist, played F. Liszt and S. Rachmaninov on the piano at Fujiwara Memorial Hall at Keio University. Ken-ichi Sato (Nagoya University, Japan) gave a keynote on how optical technologies are expected to help mitigate the adverse effects of the imminent demise of Moore's Law. Tarik Taleb (Aalto University, Finland) presented

(Continued on Newsletter page 4)

Talk of Desmond McLernon from Leeds University at IEEE Jordan ComSoc Chapter

By Ala' Khalifeh, IEEE ComSoc Jordan Chapter

The IEEE Jordan Communications Society chapter had a vibrant start in 2016. On Thursday 19 May, 2016, the chapter organized a technical talk by Dr. Desmond McLernon from Leeds University titled "Applications of Mobile Robots and Drones in Future Wireless Communication Systems" at the German Jordan University. The talk was originally geared toward academics and professionals.

Surprisingly, many students attended the talk. As organizers, we felt some unease since the talk's content was technically beyond the B.Sc. level taught to our students. So we expected the students to lose interest and miss the main objective of the talk, which is to engage engineers in serving the community's needs by researching new technologies. However, the talk "engineered" such that all the audience (including us) left the auditorium pleased.

First, the speaker credited the talk to his students. The reaction of our students was memorable when the speaker crossed out his name from the introductory slide and put the name of his student. Furthermore, the mathematical ideas in the talk were presented as logical and intuitive concepts in lieu of rigid equations. Also, to aid in explaining some problems encountered in the research presented, famous mathematical examples were used such as the marriage problem from optimal stopping theory. The interest of the audience, academics and



A group picture with Dr. Desmond McLernon surrounded by the students who attended the lecture.

students alike, was evident in the question and answer session after the talk.

In conclusion, we were thrilled by how this talk was received, and we encourage our speakers to engage B.Sc. level students in research oriented events.

ABOUT THE IEEE JORDAN SECTION

The IEEE Jordan Section was established in 1999. The membership in this section is generally rising and reached more than 1,000 members in 2014. About two thirds of these members are student members. The section has four active chapters, two active affinity groups, and eight active student branches. The four chapters are: Joint Computational Intelligence Society and Computer Society Chapter; Communications Society Chapter; Joint Power and Energy/Dielectrics Society and Electrical Insulation Society Chapter; and the Robotics and Automation Society Chapter. The two affinity groups are the Women in Engineering Affinity Group and the Young Professionals Affinity Group.

(Continued on Newsletter page 4)

International Symposium on Networks, Computers and Communications (ISNCC): The Flagship Event of the IEEE ComSoc Tunisia Chapter

By Tarek Bejaoui – IEEE ComSoc Tunisia Chapter Chair

The IEEE Communications Society Tunisia Chapter has been in operation since June 2009. From its inception, major steps have been taken to expand its activities, and its outstanding success could not have been achieved without the hard work and persistence of its volunteers.

Currently, the IEEE ComSoc Tunisia Chapter is actively engaged

in various actions, including distinguished lecturers tours, technical lecturers, and conferences. This year, the Chapter was pleased to support ISW-5G, a winter school on 5G Networks and Technologies, and ISNCC 2016, the International Symposium on Networks, Computers and Communications, held 11–13 May, 2016 at Hammamet, Tunisia.

This flagship conference, technically co-sponsored by the IEEE and the IEEE Tunisia Section, in addition to the IEEE ComSoc Tunisia Chapter, covered theoretical and practical aspects related to information systems, communication networks, and computing technologies. Its multi-thematic program focused on the major future scientific challenges related to these fields.

The conference featured a strong technical program in the area of networking, communications, and information technology, and the Technical Program Committee members took on the challenging job of evaluating the submitted papers. Their dedicat-

(Continued on Newsletter page 4)



ISNCC 2016 conference session attendees.

HPSR 2016/Continued from page 2

a keynote on network softwarization toward 5G. On Thursday morning, a keynote was presented by Akihiro Nakao (University of Tokyo, Japan) on the software defined data plane and applications. The joint plenary attracted more than 300 attendees.

The technical program, from Wednesday through Friday, comprised eight regular technical sessions: Data Center Networks, Routing, Optical Switching and Networking, Software Defined Networks, Secure and Green Technologies, Switches/Packet Processors/Traffic Monitoring, Network Virtualization, and Resource Allocation. On Thursday afternoon there was a poster session with eight poster presentations, and a workshop on high performance IP and photonic networks including 18 poster presentations. On Friday morning, Luigi Rizzo (Università di Pisa, Italy) gave an invited talk focusing on how to build efficient network data planes in software, and Dimitri Papadimitriou (Nokia-Bell Labs, Belgium) gave an invited talk on research challenges and perspectives toward Information-driven networks.

As a part of the conference technical program, on Tuesday two technical tours were conducted, one on the Keio K2 Campus to visit advanced science and technical labs, and one of Keio Digital Media Contents (DMC) to emphasize research on digital media and content. The Keio K2 Campus tour, which was guided by Naoaki Yamanaka (Keio University), included three lab visits and technical demonstrations, covering the robotics and artificial technology project dedicated to advance medical sciences; the photonics polymer project for innovative ultra high-definition liquid-crystal displays; and the elastic lambda aggregation project for high-speed future networks. The elastic lambda aggregation project, which is one of the largest Japanese national projects in the area of elastic optical networks, presented the first live demonstration of elastic multi-port wavelength selective switches. The entire tour was enjoyable and the percipients learned new technologies via lectures from top-leading professors and researchers. The Keio DMC tour was guided by Kunitake Kaneko (Keio University). Participants experienced the same content of a testing movie, but taken in different environments. Moreover, the participants experienced a demonstrated 3D movie, which is one of the innovative future high-speed network applications. The participants were allowed to see the server room of the system. This studio is not only doing the testing but also managing the archive of the media. Apart from the studio, researchers of this lab introduced their works. After the two tours were finished, more than 70 participants attended the Get Together Party, which was conducted at the Yagami campus, Keio University.

Before the conference banquet on Tuesday, the Yokohama walking tour was held. The walking tour covered about 2 km in the Yokohama bay area. The walking tour started at Minatomirai station and reached Peking Hanten Restaurant, Chinatown, which was the banquet venue.

During the banquet, the conference best paper award was presented to Andrew Mundy, Jonathan Heathcote, and Jim D. Gar-side (University of Manchester, United Kingdom) for their paper entitled "On-chip Order-Exploiting Routing Table Minimization for a Multicast Supercomputer Network." Andrew Mundy received the award certificate plaque and Japanese traditional happi coat from Eiji Oki. The best paper was selected by the HPSR 2016 Award Committee based on all the review results, including reviewers' comments and reviews by the Committee.

Malathi Veeraraghavan and Weiqiang Sun, HPSR 2017 TPC Co-Chairs, announced that the next HPSR would be held in Campos do Jordão, Brazil, on 27–30 June, 2017. More information can be found at <http://www.ieee-hpsr.org/>

ISNCC/Continued from page 3

ed and professional work made ISNCC 2016 a very successful event.

The IEEE Communications Society Tunisia Chapter generously sponsored the keynote talks that were at the technical leading edge. The conference attendees fully enjoyed the talk given by Prof. Ashfaq Khokhar from Illinois Institute of Technology, USA, discussing Big Data Challenges related to Electronic Health Record Systems. They benefited tremendously from the speech given by Prof. Giuseppe Bianchi from The University of Roma Tor Vergata, Italy, about Software Defined Networking (SDN), and took full advantage of the experiences and lessons learned about Automatic Cyberdefense that were shared by Prof. Zonghua Zhang from Telecom Lille, France.

In addition to the technical program, ISNCC attendees enjoyed the beautiful and attractive Hammamet. It is the major tourist destination in Tunisia, best known for its wide sandy beaches and water sports.

The next edition of this conference (ISNCC 2017) will take place in Marrakesh, the magical "Red City" of Morocco. The Technical Committee is composed of senior researchers with a strong background, and the program will feature highly reputable keynote speakers from the scientific and research community. Stay tuned!

JORDAN CHAPTER/Continued from page 3

The eight student branches are in the following Jordanian universities: Hashemite University, The University of Jordan, Yarmouk University, Jordan University of Science and Technology, Princess Sumaya University for Technology, Al-Balqa Applied University, Mutha University, and Al-Hussein Bin Talal University.

The section organizes several activities, including organizing the biannual IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT) conference series, technically co-sponsoring local conferences, conducting technical and professional lectures and workshops, and holding the annual general meetings, in addition to a large variety of student activities.

**GLOBAL COMMUNICATIONS NEWSLETTER**

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

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

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



Freedom through Communications

4-8 December 2016 // Washington, DC USA

The Premier Communications Event

- Access to more than 1,500 presentations exploring next generation advancements in broadband, wireless, multimedia, Internet, image and voice communications.
- Experience the Exhibition Hall featuring interactive sessions, exhibits and demonstrations showcasing the latest in 5G, IoT, IPv6, NFV, SDN and more.
- Network and exchange ideas with Industry's most distinguished leaders

Register today!

<http://globecom2016.ieee-globecom.org>

PATRONS



HUAWEI

QUALCOMM

NOKIA

SAMSUNG

KEYSIGHT
TECHNOLOGIES



5G RADIO ACCESS ARCHITECTURE AND TECHNOLOGIES



David Soldani

Periklis Chatzimisios

Abbas Jamalipour

Bernard Barani

Simone Redana

Sundeep Rangan

The global research on 5G networks has identified the need to rework the radio access network (RAN) architecture, beyond incremental evolution of current and upcoming Third Generation Partnership Project (3GPP) Long Term Evolution (LTE) releases. Novel services and usage scenarios, new classes of traffic, and increased pressure on infrastructure valorization place a number of new requirements on RANs, specifically in terms of reconfigurability and flexibility. Future RANs are thus expected to optimally exploit a broad portfolio of enabling technologies and frequency bands, departing from the “one system fits all services” approach. They will leverage virtualization, software defined networking, and cloud computing technologies to adapt execution logic to specific service requirements with composition and instantiation of access and network functions and corresponding states in the most appropriate physical locations of the infrastructure for joint performance improvement.

This Feature Topic (FT) addresses promising approaches to RAN architecture, functions, interfaces, and protocols redesign toward IMT-2020 (fifth generation, 5G). Design constraints of user and control planes and novel site configurations make it possible to provide novel services and achieve new-fangled objectives far beyond the stretched capabilities of today’s mobile network and corresponding evolution. This is especially true for massive and mission-critical machine type communications requiring a high degree of reliability and extremely low latency. The possibility of supporting different software-defined air interface variants, frequency bands, multi-connectivity, heterogeneous radio access technologies, network slicing (multi-tenancy), and context-aware adaptation of network functions and applications are a part of the challenges that need to be efficiently solved.

In *IEEE Communications Magazine*, this timely FT brings together key contributions of researchers from industry and academia, which address the above challenges, and sheds light on some fundamental technical aspects of the 5G RAN architecture and key enabling technologies.

In response to our Call for Papers, 45 manuscripts were received. The submissions underwent a rigorous review

process, following which only six outstanding papers were selected for publication. The six articles provide guidelines to 5G RAN architecture design, spectrum pooling (sharing) and initial access for mmWave networks, multi-tier drone-cell deployment, and energy consumption minimization for heterogeneous networks. These articles are expected to stimulate new ideas and contributions within the global research and innovation community, in addition to providing readers with relevant background information and feasible solutions to the main technical design issues of future 5G RANs.

The first article, “A Scalable and Flexible 5G RAN Architecture,” is by A. Maeder, A. Ali, A. Bedekar, A. F. Cattoni, D. Chandramouli, S. Chandrashekar, L. Du, M. Hesse, C. Sartori, and S. Turtinen. The article presents new services and business models along with a flexible 5G RAN functional architecture and pertinent 5G deployment scenarios using heterogeneous RANs, different carrier frequencies, and site configurations.

The second article, “5G Radio Access Network Architecture: Design Guidelines and Key Considerations,” by P. Marsch, I. Da Silva, Ö. Bulakci, M. Tesanovic, S. E. El Ayoubi, T. Rosowski, A. Kaloxylas, and M. Boldi, proposes a 5G RAN architecture, functions and interfaces supporting novel air interface variants (AIVs), as well as key enabling technologies for lower-layer service prioritization, and a novel radio resource control (RRC) state model.

In the third article, “Spectrum Pooling in mmWave Networks: Opportunities, Challenges, and Enablers,” F. Boccardi, H. Shokri-Ghadikolaei, G. Fodor, E. Erkip, C. Fischione, M. Kountoris, P. Popovski, and M. Zorzi analyze the performance of partial and full spectrum pooling with respect to an exclusive spectrum allocation, with and without coordination between operators, at different carrier frequencies and base station density, using beamforming and omnidirectional antennas.

The fourth article, “Initial Access in 5G mmWave Cellular Networks,” by M. Giordani, M. Mezzavilla, and M. Zorzi, compares the performance of three initial access techniques for 5G mmWave cellular networks (i.e., exhaustive and iterative search, and pure and enhanced context

information [CI]), in terms of misdetection probability and discovery time.

The fifth article, “The New Frontier in RAN Heterogeneity: Multi-Tier Drone-Cells,” by I. Bor-Yaliniz and H. Yanikomeroglu, introduces a drone-cell management framework (DMF) for opportunistic utilization of low-altitude unmanned aerial platforms (drones) equipped with base stations (drone-BSs), and analyzes the gains yielded by the proposed framework in terms of cost savings and number of users served by a drone-cell.

The sixth article, “Energy Consumption Minimization for FiWi Enhanced LTE-A HetNets with UE Connection Constraint,” by J. Liu, H. Guo, Z. Md. Fadlullah, and N. Kato, closes this FT with a proposal on how to reduce energy consumption with the user equipment (UE) connection constraint in fiber-wireless (FiWi) enhanced LTE-Advanced heterogeneous networks using a heuristic greedy solution to find the optimal list of active BSs and their associated UEs.

In closing, we would like to thank all the stakeholders who have made this FT possible, and hope it meets readers’ expectations, for whom this FT has been prepared.

BIOGRAPHIES

DAVID SOLDANI (david.soldani@nokia.com) received his M.Sc. degree in engineering with Magna Cum Laude Approbatur from the University of Florence, Italy, in 1994, and his D.Sc. degree in technology with distinction from Aalto University, Finland, in 2006. He was appointed visiting professor at the University of Surrey, United Kingdom, in 2014, and industry professor at UTS, AU, in 2016. He

is currently head of 5G Technology, E2E, Global, at Nokia. Prior to that, he served Huawei as head of the Central Research Institute in Europe.

PERIKLIS CHATZIMISIOS [SM] serves as an associate professor and a division head for the Department of Informatics at Alexander TEI of Thessaloniki (ATEITHE), Greece. He is involved in several standardization activities, and is the author/editor of 8 books and more than 100 peer-reviewed papers on performance evaluation and standardization of mobile/wireless communications, the Internet of Things, big data, and vehicular networking. He received his Ph.D. from Bournemouth University, United Kingdom (2005), and his B.Sc. from ATEITHE (2000).

ABBAS JAMALIPOUR [S’86, M’91, SM’00, F’07] is the Professor of Ubiquitous Mobile Networking at the University of Sydney, and holds a Ph.D. from Nagoya University. He is a Fellow of the IEICE and Institution of Engineers Australia. He has authored 17 books and book chapters, over 450 technical papers, and 5 patents. He is an elected member of the IEEE Vehicular Technology Society (VTS) Board of Governors and Editor-in-Chief of *VTS Mobile World*. He has received a number of prestigious awards from IEEE ComSoc.

BERNARD BARANI joined the European Commission (EC) with responsibility for implementation of research and policy issues in wireless communication. He is currently acting head of unit in charge of research and innovation on network technologies in the CONNECT Directorate General of the EC. He is responsible for the definition and implementation of the research strategy and related policy issues in the field of future networks, and the implementation of the 5G Public Private Partnership launched in 2013, as the flagship EC initiative in support of 5G.

SIMONE REDANA is head of the Mobile Network Architecture & Systems Research Group at Nokia Bell Labs and Chairman of the 5GPPP Architecture Working Group. He received his M.Sc. and Ph.D. degrees from the Politecnico di Milano, Italy, in 2002 and 2005, respectively. He has coordinated the EU funded project 5G NORMA. His current research interests are in novel architecture solutions for the 5G era.

SUNDEEP RANGAN [F] received his B.A.Sc. from the University of Waterloo, Canada, and his M.Sc. and Ph.D. from the University of California, Berkeley. He held postdoctoral appointments at the University of Michigan, Ann Arbor and Bell Labs. In 2000, he co-founded Flarion Technologies, which was subsequently acquired by Qualcomm, where he served as director of engineering. In 2010, he joined the ECE Department at New York University (NYU) Tandon. He is the director of NYU WIRELESS.

“The best way to predict the future is to invent it.”

-Alan Kay



IEEE COMSOC
TRAINING
www.comsoc.org/training

A Scalable and Flexible Radio Access Network Architecture for Fifth Generation Mobile Networks

Andreas Maeder, Amaanat Ali, Anand Bedekar, Andrea F. Cattoni, Devaki Chandramouli, Subramanya Chandrashekar, Lei Du, Matthias Hesse, Cinzia Sartori, and Samuli Turtinen

Compared to 3GPP LTE, fifth generation radio access networks need to support much more diverse requirements with a wide range of deployment options for infrastructure and spectrum availability. Based on this insight, the authors discuss a flexible and scalable radio access network architecture design.

ABSTRACT

The fifth generation of mobile networks is expected to become the key enabling technology for new services and businesses in the Internet of Things realm, including automotive and industry communications. At the same time, the demand for the “bread and butter” services of mobile broadband will continue to increase, driving the need for ubiquitous data capacity everywhere. Compared to 3GPP LTE, fifth generation radio access networks need to support much more diverse requirements with a wide range of deployment options for infrastructure and spectrum availability. Based on this insight, this article discusses a flexible and scalable radio access network architecture design.

INTRODUCTION: MAIN DRIVERS OF THE FIFTH GENERATION RADIO NETWORKS

The development of the fifth generation (5G) of mobile networks is driven by three main factors: new services and markets, a new way of deployment and operation, and the ever increasing need for more data communication capacity and higher bandwidth.

Mobile Internet is one of the most important enablers of the connected society. In the near future, the increase of smartphones and tablets will continue to drive the huge traffic growth of global mobile IP traffic. By 2030 there is likely to be as much as 10,000 times more wireless data traffic than there was in 2010 [1].

In addition, society will enormously benefit from the power of connectivity taking place in almost every industry. In particular, high-priority service areas are education, health, government, public safety, and disaster relief, as well as public and private transportation. In this context, a diverse set of services is emerging and shows significantly different characteristics from today’s dominant human-to-human and human-to-machine traffic.

The design of the 5G radio access network (RAN) architecture needs to have a healthy balance between evolution and revolution in order

to fulfill the novel requirements and operation paradigms on one hand, and cost efficiency and best utilization of already existing deployments on the other hand.

The main drivers of the 5G RAN architecture design are the following.

New services and business models: Efficient support of new and enhanced services, including mobile broadband with very high data rates, massive machine-type communication, and ultra-reliable and ultra-low-latency communication. Here, network slicing is widely recognized as a key technology enabler to tailor network operations to new services and business in an efficient way [2].

Cloud RAN will drive a paradigm shift in operation and deployment of mobile networks. Based on on-demand provisioning of resources, centralization of network functions and network functions virtualization (NFV) will impact functional components and interfaces of the logical network architecture.

Integration of multiple radio access technologies (RATs): In order to achieve improved user experience and cost efficiency, 5G should provide a framework for tight integration of new 5G and existing radio interfaces in the RAN, including WiFi and LTE.

Several standards organizations and industry fora have published requirements and design objectives for 5G, including the Next Generation Mobile Networks (NGMN) Alliance in [2] with a comprehensive list of requirements, the International Telecommunication Union — Radio-communication Standardization Sector (ITU-R) in [3] with a general set of design objectives for IMT-2020 systems, and the Third Generation Partnership Project (3GPP) in [4] with an analysis of use cases and requirements. Furthermore, operators and the academic community have proposed high-level design concepts, such as in [5] from the end-to-end system perspective, and in [6, 7], where key challenges are described. This article analyzes the technical requirements and features to address the flexibility and scalability objectives of 5G systems, and proposes an architecture that addresses said requirements.

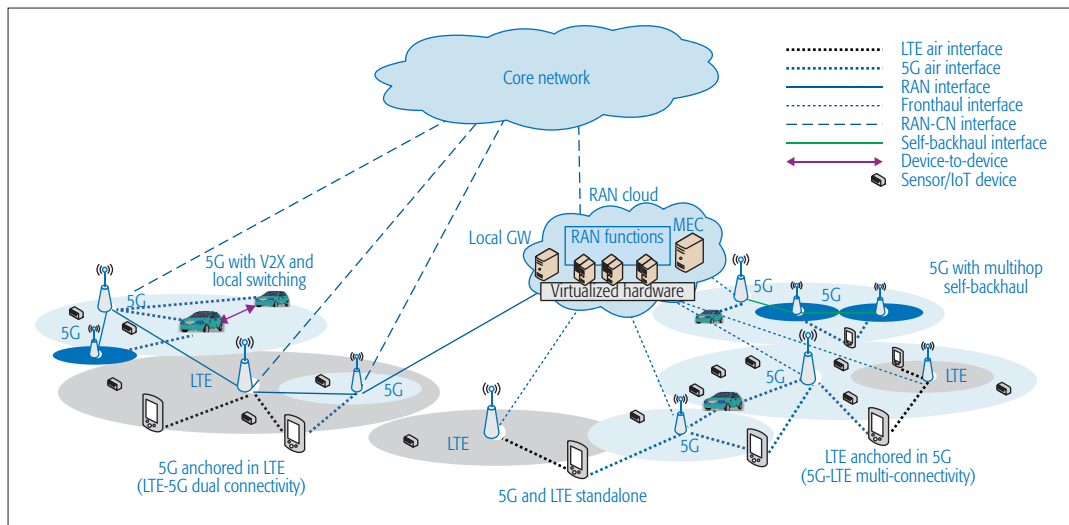


Figure 1. 5G RAN scenarios.

5G SCENARIOS AND REQUIREMENTS

5G will allow industry stakeholders to significantly broaden the range of use cases from enhanced mobile broadband (eMBB) and interactive services to massive machine-type communication (mMTC) and ultra-reliable, critical machine-type communications (cMTC) and many more. Both traditional and new use cases impose a set of challenging requirements to the 5G system [3]:

- For eMBB a peak cell rate of 20 Gb/s and an experienced user data rate of 1 Gb/s with at least 100 Mb/s at cell edges is expected. The end-to-end latency should not be above 10 ms. Vehicular speeds up to 500 km/h shall be supported for high-speed train use cases.
- For mMTC, while experiencing usually low data rates not above 100 kb/s, a density of up to 106 devices per km² is expected.
- For cMTC, less than 1 ms over-the-air latency is expected for certain use cases.

The requirements of these use cases are vastly different in terms of network performance indicators, but also in the categories of cost efficiency and functional requirements. As an example, for massive MTC the cost factor for chipset and module is the most important due to the expected number of devices, and the low peak rates and restricted periods of service and mobility.

The challenge for the 5G system architecture is to support all requirements with one flexible system. Figure 1 illustrates how the 5G main use cases integrate into a common system perspective. One important aspect is that 5G will not take over Long Term Evolution (LTE) spectrum disruptively. In contrast, LTE deployments will be utilized and tightly integrated into 5G, with support of dual connectivity and multi-connectivity between LTE and 5G air interfaces. Multihop self-backhauling will be supported for coverage extension of high frequency spectrum above 6 GHz. Device-to-device communication is required for critical MTC communication for very low latency between devices such as in vehicle-to-vehicle communications. Latency-tolerant RAN functions will be shifted into cloud deployments, depending on the infrastructure

and fronthaul/backhaul capabilities. Mobile edge computing (MEC) and local services, dedicated for specific use cases like industry automation, are co-located and operate on the same virtualized or physical infrastructure. For more use cases with some specific service requirements, see [4].

FLEXIBLE AND SCALABLE DESIGN

The logical 5G RAN architecture needs to support logical functions and interfaces such that flexible placement and scaling of logical entities is enabled. Both *centralized* and *distributed* network deployments will be supported by the architecture design. It depends on the operator's infrastructure and offered services which functions will be centralized and where they will be placed. For example, dedicated infrastructure for small cell may be necessary due to insufficient transport network capacities [8]. In such a case, small cell traffic can be aggregated in dedicated network nodes or edge cloud facilities. Nevertheless, coordination between macro and small cells is required. One possibility would be a control/user (C/U)-plane split with low-footprint user plane nodes at aggregation nodes or in the cloud, while radio resource coordination and the control plane are still anchored in the macrocell. This also enables over-the-air C/U-plane split for increased mobility robustness and overhead reduction [5].

Furthermore, depending on the fronthaul capabilities, both C-plane and U-plane RAN functionality can be placed in centralized nodes. In this configuration, the signaling effort for coordination between functions in the cloud is small, and enables coordination gains [7].

To this end, in comparison to 3GPP LTE, on one hand a functional decomposition of RAN is necessary, and on the other hand requirements for flexibility and scalability need to already be considered in the design phase with the following features.

Flexible RAN Functional Split: Classic C-RAN approaches are based on I/Q sample forwarding as specified by the Common Public Radio Interface (CPRI) industry cooperation. The very high bandwidth requirements of 5G

The logical 5G RAN architecture needs to support logical functions and interfaces such that a flexible placement and scaling of logical entities is enabled. Both centralized and distributed network deployments will be supported by the architecture design. It depends on the operator's infrastructure and offered services which functions will be centralized and where to be placed.

Within the RAN, some functions have tight real-time requirements at millisecond or below time-scales and must complete their execution within hard deadlines, e.g., the MAC scheduler and PHY layer functions. Other protocols do not have such real-time requirements, such as the packet data convergence protocol in LTE, or radio resource control functions.

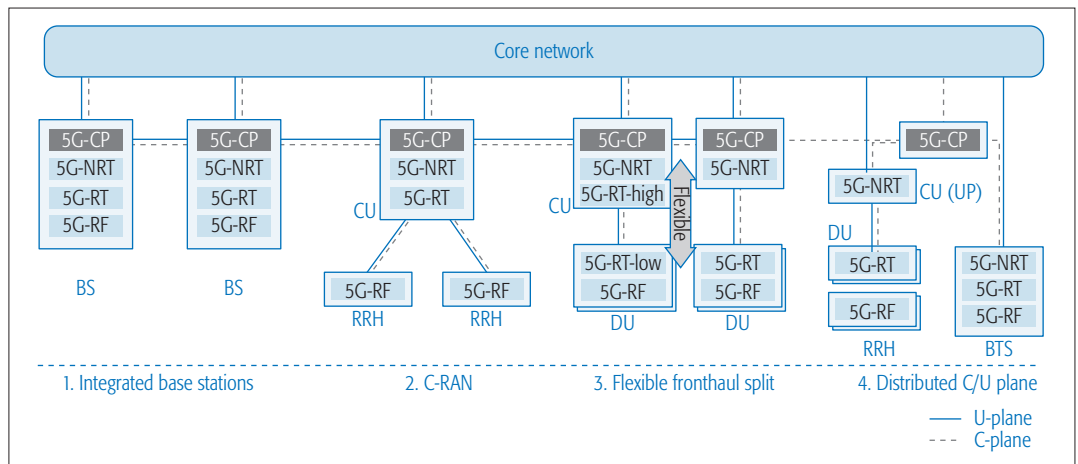


Figure 2. Flexible RAN architecture with fronthaul split and distributed C/U-plane.

would make such an approach very costly for many mobile network operators (MNOs). However, there is wide support in the industry fora for introducing centralization in 5G, driven by the demand for cloud computing principles, and expected performance gains and cost reduction in the longer term. In order to be able to adapt to the individual needs of operators, 5G RAN architecture will support different functional splits of the RAN protocol stack.

C/U-Plane Split: Compared to monolithic deployments, cloud computing principles allow more flexible resource assignment and scaling of network functions. In a RAN, the C-plane and U-plane often scale differently, depending on the type of traffic that needs to be supported. For instance, mMTC applications often generate only very small data packets of several bytes, but multiplied by tens of thousands of individual connections per cell. In this case, C-Plane functions will be much more loaded than U-Plane ones. On the other hand, very high data rates require relatively low C-Plane effort, but immense computing resource for the U-plane. Thus, a clear split between C- and U-plane functions enables independent scaling and placement of these functions [9].

Flexible Network Functions Placement: The placement of RAN functions is driven on one hand by the trade-off between low-latency requirements of services (e.g., mMTC) and the function itself (e.g., functions relying on control-loops between user equipment [UE] and network such as hybrid automatic repeat request [HARQ]), and on the other hand by centralization gains due to coordination of radio resources, and pooling and statistical multiplexing gains on processing in the central site [10], and on the transport network requirements [11]. This implies that both the U-plane and C-plane need to support flexible placement of network functions.

Separation of Per-Cell and Per-User Functions: 5G should enable separation of per-user functions from per-cell functions in order to maximize the ability to pool and the ability to elastically scale the processing resources allocated to each type of function.

Differentiation between Real-Time and Non-Real-Time Functions: Within the RAN, some

functions have tight real-time (RT) requirements at millisecond or below timescales and must complete their execution within hard deadlines, such as the medium access control (MAC) scheduler and physical (PHY) layer functions. Other protocols do not have such real-time requirements, such as the packet data convergence protocol (PDCP) in LTE, or radio resource control (RRC) functions. A key design tenet for the 5G RAN functions and interfaces is to enable separation of RT functions from non-real-time (NRT) functions so that the right execution environment and scaling paradigm can be applied to each class of functions.

An example of the resulting flexible 5G RAN architecture concept is illustrated in Fig. 2. On the left side, an integrated base station (BS) deployment is shown as the state of the art in 3GPP LTE. All C- and U-plane functions are integrated in one logical and physical network entity. The terms RT and NRT denote the real-time and non-real time parts of the RAN protocol stack functions. In LTE, the NRT part corresponds to PDCP and upper parts of RLC, including link reliability schemes. Different functional splits are possible; here, two examples are shown which may correspond to intra-PHY split (left of Fig. 2) and split between RT and NRT functions. Note that most radio resource control plane (CP) functions are not time-critical on below millisecond scales.

The left and right of the center show different options for centralizing RAN functions. C-RAN deployments with fully centralized functions in a central unit (CU) and distributed remote radio head (RRH) are used when very capable fronthaul is available. For other cases, only parts of the protocol stack are centralized in a CU, while the lower part is distributed in a distributed unit (DU).

The right side shows a configuration with separated control- and user-plane functions. The NRT user plane is aggregated in a central unit, exploiting centralization gains from cloud implementation and resource management independent from C-plane functions, which allows independent scaling, load balancing, and placement.

The C-plane is centralized as well, but not necessarily co-located with the U-plane functions.

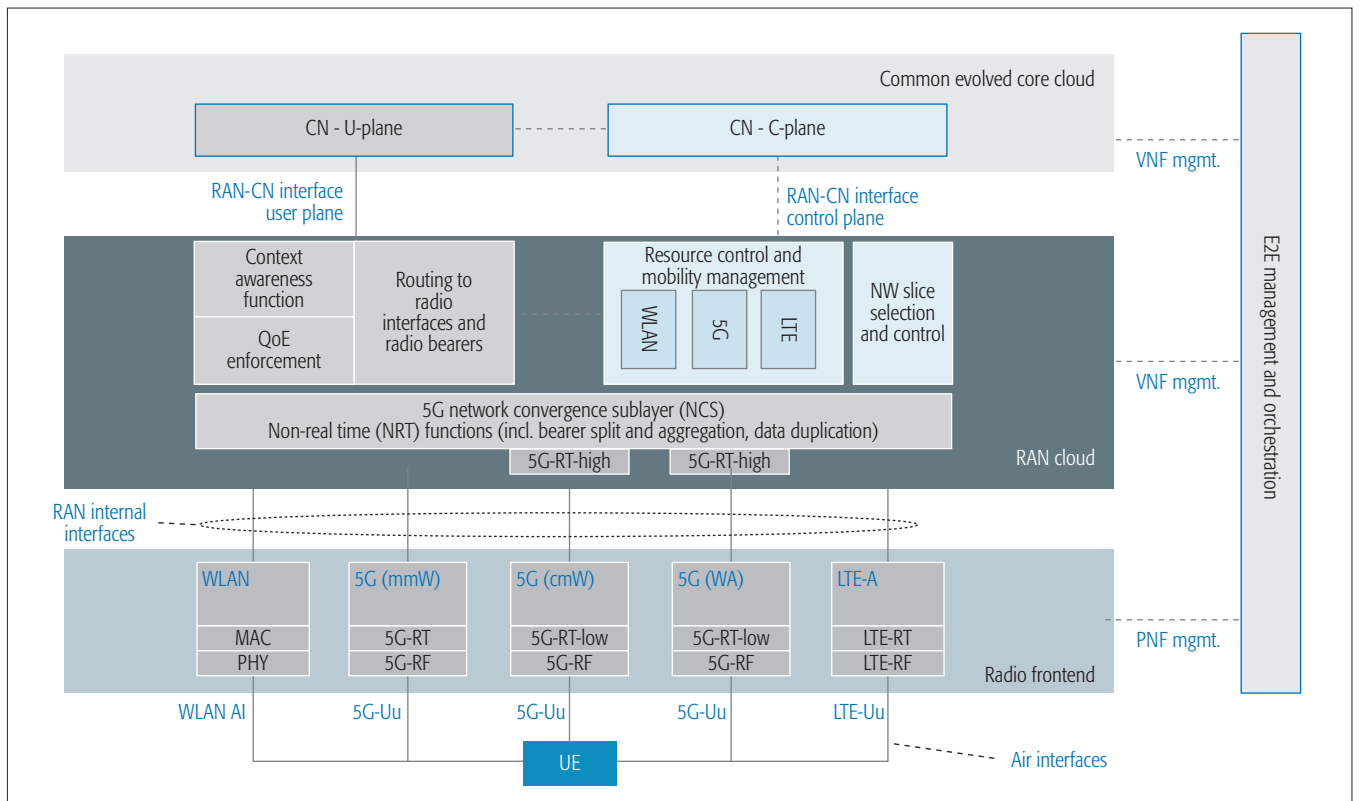


Figure 3. 5G functional RAN architecture.

This enables coordinated control of integrated macro sites and small cells in heterogeneous network deployments as described above.

FUNCTIONAL ARCHITECTURE

The 5G RAN functional architecture aims to achieve a unified framework which minimizes the need for specialized functions for different RATs (e.g., 5G, LTE, WiFi) or radio interfaces (e.g., wide area [WA], centimeter-wave, and mmWave air interfaces).

GENERAL PRINCIPLES

Figure 3 illustrates the general principles of the functional 5G RAN architecture. It is assumed that the core network is separated into C-plane and U-plane functions, which are located in operators' core cloud data centers. It is further assumed that core network functions are as far as possible *access-agnostic* in the sense that independent of the RAT, the same functionality can be used. This also facilitates independent evolution of core and radio (e.g., if a new RAT is to be deployed in operators' networks) and the use of a single RAN-CN interface for different RAT types.

Correspondingly, the same principle is applied to the RAN, where common functions for different RATs should be used as far as possible. This includes a common U-plane protocol, the network convergence sublayer (NCS), as the anchor for multiple RAT and multi-connectivity features, routing and traffic handling functions, and quality of service/experience (QoS/QoE) enforcement. The C-plane includes radio resource and mobility management, as well as radio RRC protocol for different RATs, which can have common and access-specific aspects.

Furthermore, integrated radio resource management and control of multiple RATs for fast coordination of radio resources in different sites is a basic principle of the proposed architecture. Close coordination and joint processing in cloud RAN enables gains for system and cell edge capacity, as well as improved user experience with seamless and transparent inter-RAT access [7].

The upper layers may be deployed in RAN cloud sites; however, as explained earlier, stand-alone and "cloudified" deployments of various degrees should be supported. This is illustrated by the RT and NRT parts of the protocol, which can be located in the RAN cloud or in the radio front-end part.

USER PLANE

5G RAN U-plane functions need to support very high data rates as well as low-latency services over different radio interfaces (multi-connectivity) and RATs.

The NCS is an enhancement of the LTE PDCP sublayer for 5G. Like PDCP, the NCS shall provide ciphering and header compression, and in-sequence delivery to upper layers. The NCS additionally provides in-service-flow differentiation and multi-connectivity anchoring. As indicated in Fig. 3, NCS is a universal transport protocol also used for control-plane messages.

The NCS layer acts as a multi-connectivity anchor which is responsible for traffic routing, splitting, and packet duplication according to the configuration by the RRC. NCS can receive feedback from the lower layers of each multi-connectivity leg that allows NCS to determine how much traffic to send to each multi-connectivity leg (i.e., flow control).

The 5G control plane needs to support diverse use cases. Compared to LTE, it needs to support configuration of multi-connectivity with several nodes in centralized and distributed deployment configurations over potentially discontinuous spectrum bands for supporting high data throughput sessions at several gigabits per second.

NCS is envisioned to become the common U-plane protocol for all RATs which are tightly integrated in a common 5G architecture. As such, it should also be optimized for cloud deployment:

- Protocol design optimized for parallel processing in cloud processing environments
- Support for dynamic relocation and instantiation of U-plane entities for load balancing and service-aware networking

The latter implements the concept of flexible functional placement, which is based on the UE multi-connectivity configuration, the end-to-end path that the traffic is expected to follow, the end-to-end latency and bandwidth expectations of the service flow, the transport network topology (including knowledge of bandwidth and latency situations in the network), and the processing load at candidate sites that could host the NCS functionality.

At lower layers, the 5G equivalent of radio link control (RLC) is responsible for segmentation of data into MAC transport blocks, and for link reliability by means of ARQ. The 5G MAC layer is responsible for multiplexing and packet scheduling of control and user data. For 5G, HARQ functions in MAC shall be designed such that fronthaul constraints from the HARQ feedback loop are reduced and configurable, enabling a flexible fronthaul split as described in Section III.

CONTROL PLANE

The 5G C-plane needs to support diverse use cases. Compared to LTE, it needs to support configuration of multi-connectivity with several nodes in centralized and distributed deployment configurations over potentially discontinuous spectrum bands for supporting high data throughput sessions at several gigabits per second. For mMTC, devices transmitting infrequent small packet data and targeting prolonged battery lifetimes, it is expected that the signaling and protocol overheads will be very low. However, for cMTC, it needs to support low-latency connection setup, and enhanced mobility robustness and signaling reliability.

From an architectural viewpoint, a key design feature of 5G C-Plane is to implement a strict separation of the logical C-plane from the transport plane for control messages to support massive centralization for the C-/U-plane split. From a logical point of view, the UE should communicate as a default with one RRC entity in the network. The transport of signaling between UE and network is the task of the NCS sublayer, which can utilize different radio interfaces, and perform duplication of C-plane messages on different radios to increase the overall signaling reliability for cMTC use cases. This solution is flexible to adapt C-plane key performance indicators to different use case, and scalable by letting the NCS decide per message, based on feedback, to which radio interfaces to send. Note that C-plane and U-plane are not necessarily co-located in the case of dedicated U-plane anchors on traffic aggregation nodes.

MTC traffic introduces a larger share of small and infrequent packet data into the net-

work (e.g., from sensors). Similarly, very popular smartphone applications for RT social interaction regularly exchange keep-alive and status messages, and generate traffic only sporadically and with low intensity. Because frequent transitions between connected and idle mode increase the signaling load in the network and the energy consumption of the UE, novel mechanisms for signaling reduction and mobility on demand are necessary.

MOBILITY ON DEMAND

Experience from the field has shown that more than 70 percent of devices camping in mobile networks are stationary or nomadic, meaning that devices are either fixed to a certain location or moving slowly within a small area. Based on this insight, compared to LTE systems, scalable solutions for mobility management based on the actual *demand* and capabilities of the application and device are required. Furthermore, mobility events should be hidden as much as possible from the core network in order to reduce overall signaling load in the network. This can be achieved by the following means.

Configuration of the level of mobility support according to service and application requirements, such that for certain devices which are nomadic or immobile, session continuity (from CN perspective) between network nodes is not supported. Depending on the application needs and device capabilities, the network determines the right level of active and idle mode mobility to be assigned to a certain device.

Centralization of the RAN mobility anchor in aggregation nodes of the network. Path switch events that require signaling to the core network can be avoided, specifically for small cells, which are aggregated in high numbers under the umbrella of the NCS layer located in the RAN cloud.

Reduction of state transitions to idle mode with corresponding signaling (including authentication and location updates) to the core network. A new RRC state keeps the UE context available in the RAN, while the UE is inactive and saving battery power by switching off transceiver components. In this state, the access network configures the UE to perform an autonomous, lightweight mobility procedure in which, upon cell reselection, the UE context is retrieved and forwarded from the last visited base station. Since the UE is connected, paging procedures (for cellular terminals) are managed in the access network to avoid additional signaling to the core network, thus reducing the overall signaling load while maintaining full connectivity from the core network point of view.

Zero latency mobility is needed for support of cMTC applications by means of “make before break” (i.e., first establishing the new link to the target BS before removing the link to the source) and utilizing multi-connectivity mechanisms.

Note that similar as the proposal in [12], configuration of transport infrastructure and data path switch based on software defined networking (SDN) principles can be implemented. In this case, relocation of the NCS protocol entity is initiated upon path switching by a corresponding SDN controller.

For efficient management of QoE, the radio access should be able to react quickly and appropriately to application requirements in a dynamic manner [13]. Therefore, QoE and QoS should be managed close to the interface representing a potential bottleneck, without the need for immediate end-to-end signaling for bearer setup and modification as in 3GPP LTE systems.

The following components are part of this framework in RAN:

- The context awareness function (CAF) ensures, based on policy information, that QoE is enforced in a coordinated manner with any potential core network mechanisms. It derives dynamic QoS targets and local enforcement actions in the RAN based on policies provided by a corresponding CN function. Information on the application type can be derived from local detection or packet marking from the core network. Two levels of policies are foreseen: *explicit QoS targets* (e.g., for the support of IMS-based voice services and any other services like critical communication for verticals) and *intent level policies*, which comprise high-level policies and guidelines for QoE management.

- An *application scheduler* differentiates between application traffic streams on a connection, such that bandwidth scheduling and prioritization can be applied and policies are enforced. The application scheduler is independent from the radio access protocol stack, but should be capable of taking feedback on radio interface capacity into account.

- RAN-based mapping* of traffic flows to different radio interfaces and radio bearers independent of core network signaling. The granularity of the mapping depends on the characteristics of the application traffic. For example, best effort application traffic with small data burst could be aggregated in a single queue, while interactive video is mapped to a dedicated queue.

- QoS enforcement* is done by the MAC scheduler by means of parameterization of the scheduling weights according to the QoS queue characteristics.

The same principles apply to operator managed traffic like voice or business services, which often need to fulfill stricter requirements than for Internet traffic.

SUPPORT FOR NETWORK SLICING

A network slice refers to a logical end-to-end network that shares resources on a common physical infrastructure with other network slices. Network slicing enables operators to offer tailored networks for specific business operations, such as verticals from the automotive and transportation, energy, or health sector. Network slicing requires a certain degree of isolation between resources in the RAN, as well as operational and management means for life cycle management and configuration of slices [2].

A virtualized implementation of C/U-plane functions enables, together with a flexible C/U-plane configuration, network slices that are tailored to specific use cases. For example, a network slice for a low-latency cMTC use case (e.g., industry automation) would be instantiated

with C- and U-plane functions close to the radio interfaces. A network slice for mobile broadband on part of the same physical infrastructure could be instantiated with U-plane functions in network nodes with higher aggregation such that multiplexing and pooling gains can be exploited. This flexible instantiation of slices is achieved by means of network slice blueprints, which map service-level requirements by means of business and policy decisions to a set of network functions.

One of the key issues in a RAN is how to realize device access to an end-to-end network slice [14]. In the proposed functional architecture, a *network slice selection and control function* is responsible for associating the entities belonging to a logical end-to-end slice to a device based on service requirements or other information. This information can be available in the network (e.g., in the subscription data) or provided by the user. Depending on the instantiation of the slice, RAN functions are configured by network management for customized utilization by different RAN slices according to the service requirements. Furthermore, enforcement of QoS should be supported according to slice requirements by slice-specific policies from which then QoS parameters are derived by the CAF.

In the radio access, some enhancements are necessary in order to support the requirements for cryptographic and resource isolation between slices. For example, slice-specific key assignment and placement of security-related protocol functions (i.e., NCS) may be necessary. In order to achieve end-to-end latency key performance indicators (KPIs), slice-specific configuration of the radio resource scheduler, as well as provisioning of dedicated, slice-specific resources may be required as well.

NETWORK MANAGEMENT AND ORCHESTRATION

5G will be the first generation of mobile network systems that is designed for NFV and cloud RAN. This requires an automated NMOS for the configuration, performance, fault, and (life cycle) management of network functions. One of the main challenges is the integration of virtual network functions (VNF) and physical network functions (PNF) management. While VNFs can be scaled according to the load situation, PNFs are static and need to be planned in such a way that peak capacity requirements are fulfilled. If VNFs and PNF are operated jointly, NMOS has to ensure that the resource allocation for VNFs match the requirements of PNFs in the radio front-end of the network. Furthermore, the (virtual) transport network needs to match the requirements and logical topology of the mobile network, including network slice instantiations, which is achieved by jointly managing network functions and data layer by means of SDN-based control in the end-to-end framework of NMOS.

Network slicing introduces an additional dimension: apart from slice instantiation, NMOS performs inter-slice orchestration, which coordinates the allocation of according resources and prioritizes access to the VNF instance among slices in the RAN cloud, and configures PNFs in the radio front-end.

For efficient management of QoE, the radio access should be able to react quickly and appropriately to application requirements in a dynamic manner. Therefore, QoE and QoS should be managed close to the interface representing a potential bottleneck, without the need for immediate end-to-end signaling for bearer setup and modification as in 3GPP LTE systems.

The road to the 5G radio and network architecture transformation is not trivial and market needs are not homogeneous around the globe. To this end, the 3rd Generation Partnership Project (3GPP) considers the introduction of the 5G specification in phases. This ensures that the 5G system for the foundational network is realized as soon as possible.

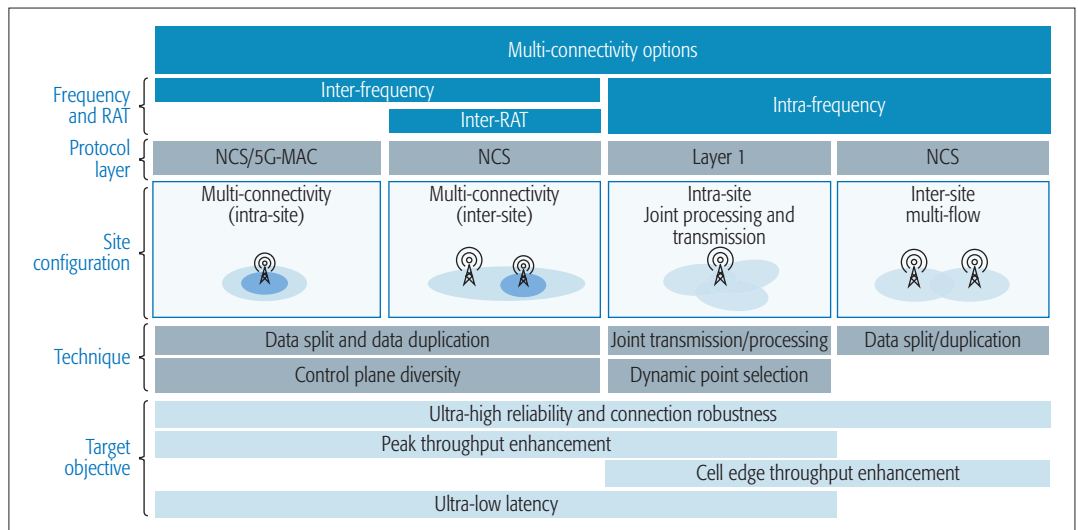


Figure 4. Multi-connectivity techniques and objectives.

FLEXIBLE AIR INTERFACE: MULTI-CONNECTIVITY AND HETRAN

Heterogeneous RAN (HetRAN) refers to a network where different carrier frequencies and site locations are deployed in a single network. Multi-connectivity denotes a mode of operation where the UE is configured with radio resources from more than one radio access (or cell).

Multi-connectivity is a key technology to meet 5G target objectives for throughput, reliability, latency, and connection robustness. As shown in Fig. 4, multi-connectivity is an umbrella term for several techniques that are applicable on a case-by-case basis, depending on the radio and site configuration and on the intended objective.

Intra-site denotes the case where the radio transmission endpoints are co-located in the same physical site, while in case of inter-site, they are located in different physical sites. Inter-frequency multi-connectivity refers to the case where a UE is connected to two radios on different carrier frequencies. These frequencies could also be utilized by different RATs. In contrast, intra-frequency multi-connectivity refers to transmission at the same carrier frequency. Note that inter-site can also refer to different RUs in case of a centralized deployment with fronthaul split.

In 5G, inter-site multi-connectivity is a generalization of the dual connectivity model of 3GPP LTE [15] with the following enhancements:

- Support of data split and duplication to more than two nodes, for example, in the case of HetNet scenarios with LTE, 5G macro, and 5G mmWave cells in different sites.

- Transmission of RRC messages over several radio links (in parallel or selective) for the purpose of increased connection robustness. Note that the handling of U-plane and C-plane is not necessarily congruent: for example, handover signaling messages could be transmitted with packet duplication over different radio interfaces, while for the U-plane, data split is applied in order to increase throughput. While the UE should maintain only a single RRC state in order to reduce management complexity, messages can originate from multiple RRC entities in the network in a coordinated manner.

- Support for split bearer on secondary nodes where the C-plane master is not co-located with the U-plane anchor, for example, to offload 5G traffic from legacy (LTE) or to enable efficient high aggregation nodes in cloud RAN for small cells, with the C-plane still at the macrocell.

- Direct transmission of RRC messages from secondary node (e.g., for MAC/PHY configuration) in order to eliminate the backhaul latency between master and secondary nodes.

The NCS protocol layer is responsible for implementing multi-connectivity U-plane functions, configured by RRC. Due to the design objective of general applicability of NCS, inter-site multi-connectivity can also be applied to multiple RATs, including LTE and WiFi.

OUTLOOK: THE ROAD TOWARD 5G

The road to the 5G radio and network architecture transformation is not trivial, and market needs are not homogeneous around the globe. To this end, 3GPP is considering the introduction of the 5G specification in phases. This ensures that the 5G system for the foundational network is realized as soon as possible, while in later phases, optimizations for novel features will follow the demand of the market.

1. The first phase, with expected first commercial deployments around 2018, solves imminent commercial market needs to boost mobile broadband capacity. At the same time, it leverages existing LTE deployments by tightly connecting 5G radio access to the 4G network via dual connectivity. In parallel, a 5G standalone system with a focused set of functionalities is introduced to build the foundational 5G network. Support for both dual connectivity and a standalone system is necessary for wider acceptance of the 5G standard in the first phase for MBB.

2. The second phase should evolve the 5G system for all use cases such as mMTC, cMTC, and further enhancements of MBB.

The phased approach will allow operators to leverage from existing deployments to provide higher data rate, better capacity in the near term, and at the same time, introduce a future-proof network architecture to support new use cases and services in the longer term. Thus, it will

enable operators have the right tools at hand to harvest the opportunity and unlock new revenue potentials by offering new services via their converged 5G network in a timely manner.

In this context it is important to recognize that the specification and development of “5G” will not be restricted to 3GPP only, which nevertheless will take on the main work for the radio access. However, operation and deployment requirements will be addressed in other fora as well, including industry fora such as the NGMN Alliance, open source initiatives such as Open BTS, and infrastructure-related standards organizations such as CPRI and the Next Generation Fronthaul Interface (NGFI) initiative [16].

While it is difficult to forecast the outcome of the many standardization processes related to 5G, this article provides an overview of the key design objectives for the RAN architecture. It should be emphasized that the main challenge is not to build a system that can cover just the diverse use cases of 5G, but to do so in a scalable and flexible way, such that cost and effort are justified by industry stakeholders and the standards community.

ACKNOWLEDGMENTS

The authors are thankful for the fruitful discussions and invaluable contributions of our colleagues in Nokia and Nokia Bell Labs. Special thanks to Agnieszka Szufarska of Nokia Bell Labs Wroclaw, Poland, and to Christian Mahr, Nokia MN, Ulm, Germany.

REFERENCES

- [1] ITU-R, “Rep. ITU-R M.2370-0; IMT Traffic Estimates for the Years 2020 to 2030,” July 2015.
- [2] NGMN Alliance, “NGMN 5G White Paper,” Feb. 2015.
- [3] ITU-R Rec. M.2083-0, “IMT Vision – Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond,” Sept. 2015.
- [4] 3GPP, “TR 22.891 V1.2.0; Study on New Services and Markets Technology Enablers,” Nov. 2015.
- [5] P. K. Agyapong *et al.*, “Design Considerations for a 5G Network Architecture,” *IEEE Commun. Mag.*, vol. 52, no. 11, Nov. 2014, pp. 65–75.
- [6] P. Demestichas *et al.*, “5G on the Horizon: Key Challenges for the Radio-Access Network,” *IEEE Vehic. Tech. Mag.*, vol. 8, 2013, pp. 347–53.
- [7] P. Rost *et al.*, “Cloud Technologies for Flexible 5G Radio Access Networks,” *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 68–76.
- [8] NGMN Alliance, “Small Cell Backhaul Requirements,” June 2012.
- [9] R. Agrawal *et al.*, “Cloud RAN Challenges and Solutions,” *Proc. 19th Conf. Innovations in Clouds, Internet, and Networks*, Paris, France, Mar. 2016.
- [10] P. Rost, S. Talarico, and M.C. Valenti, “The Complexity-Rate Tradeoff of Centralized Radio Access Networks,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, Nov. 2015, pp. 6164–76.
- [11] J. Bartelt *et al.*, “Fronthaul and Backhaul Requirements of Flexibly Centralized Radio Access Networks,” *IEEE Wireless Commun.*, vol. 22, no. 5, Oct. 2015, pp. 105–11.
- [12] R. Trivisonno *et al.*, “SDN-Based 5G Mobile Networks: Architecture, Functions, Procedures and Backward Compatibility,” *Trans. Emerging Telecommun. Tech.*, vol. 26, no. 1, Jan. 2015, pp. 82–92.
- [13] N. Radics, P. Szilágyi, and Csaba Vulkán, “Insight Based Dynamic QoS Management in LTE,” *Proc. IEEE PIMRC 2015*, Hong Kong, China, Aug. 2015.
- [14] X. An *et al.*, “On End to End Network Slicing for 5G Communication Systems,” *Trans. Emerging Telecommun. Tech.*, June 2016, doi: 10.1002/ett.3058.
- [15] 3GPP, “TS 36.300 V12.8.0; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description,” Jan. 2016.
- [16] C.L. I *et al.*, “Rethink Fronthaul for Soft RAN,” *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015.

BIOGRAPHIES

ANDREAS MAEDER (andreas.maeder@nokia-bell-labs.com) is a senior radio researcher at Nokia Bell Labs in Munich, Germany, where he is leading the research on 5G RAN architecture work. He received his Ph.D. in 2008 from the University of Wuerzburg, Germany. He contributed to the standardization

of IEEE 802.16m on WirelessMAN-Advanced and IEEE 802.16p on M2M communications, to 3GPP RAN2, and to 3GPP SA2. He has authored numerous standard contributions, conference papers, journal articles, book chapters, and patents.

AMAANAT ALI (amaanat.ali@nokia.com) is a 3GPP RAN2 standardization delegate at Nokia based in Bangalore, India, for HSPA and 5G topics, working on architecture and control plane protocol areas since 2012. Since 2009, he has been involved in end-to-end system architecture and specification work within Nokia, interacting closely with product implementation and design teams to introduce new roadmap features. He is also a student at Aalto University’s Electrical Engineering Department pursuing his M.Sc. in wireless communication.

ANAND BEDEKAR (anand.bedekar@nokia.com) leads a team on RAN and E2E architecture in the Technology Vision and Architecture group in Nokia’s Mobile Networks CTO Organization. His research interests are in the architecture of next-generation networks and algorithms for scheduling and radio resource management. He received his B.Tech. from the Indian Institute of Technology, Bombay, and his M.S. and Ph.D. from the University of Washington.

ANDREA F. CATTONI (andrea.cattoni@keysight.com) received his M.Sc. and Ph.D., both from the University of Genoa, Italy, in 2004 and 2008, respectively. From 2008 to 2015 he was first a postdoctoral researcher and then an associate professor at Aalborg University, Denmark, where he investigated wireless technologies and networks ranging from cognitive radio networks, LTE-A, Wi-Fi advances, and lately to 5G. In January 2016 he joined Keysight Technologies as a research engineer. His research interests are 5G systems and networks, software defined networks, and network virtualization.

DEVAKI CHANDRAMOULI (devaki.chandramouli@nokia.com) is currently with Nokia Bell Labs. Her focus areas include architecture development for 5G research and standards development for 3GPP access and LTE/SAE related topics. She has been involved in 5G research since the beginning of 5G architecture research work and is a Co-Rapporteur for the Next Generation architecture study in 3GPP. She has co-authored numerous academic papers and internal concept papers in the area of wireless technology. She has also (co-) authored over 60 patents in wireless communications.

SUBRAMANYA CHANDRASHEKAR (subramanya.chandrashekar@nokia.com) is a senior specialist at Nokia Networks, Bangalore, India, where he is in the C-plane solution domain responsible for 5G system architecture and design. He has exposure to requirement engineering in both control and user plane domains of the RAN. He was also a key contributor to the back-office support for 3GPP HSPA Rel 10 standardization. He has authored many 3GPP contributions, some journal articles, and more than 20 patents and patent applications.

LEI DU (lei.du@nokia.com) is a senior researcher for Nokia Bell Labs in Beijing, China. She received her Master’s degree from Beijing University of Posts and Telecommunications in 2004 and joined DoCoMo Beijing Labs working on 4G research and standardization. She joined Nokia in 2008 and has been leading working groups and projects on LTE and 5G work, and contributes to 3GPP on several topics. She has authored numerous patents and scientific papers on wireless communication.

MATTHIAS HESSE (matthias.hesse@nokia-bell-labs.com) received his Diplom Ingenieur degree (M.Sc. equivalent) in electrical engineering from Dresden University of Technology in 2006, and his Ph.D. degree from the University of Nice Sophia-Antipolis, France, in 2010. He is a senior radio research engineer at Nokia Bell Labs in Wroclaw, Poland, where he has been leading several research projects and actively contributing to 3GPP standardization. For the past two years his research focus has been on the design of new MAC, RLC, PDCP, and RRC functionality in novel 5G communication systems.

CINZIA SARTORI (cinzia.sartori@nokia-bell-labs.com) is a 5G researcher at Nokia Bell Labs in Munich, Germany. She is engaged in designing the end-to-end 5G network architecture in Nokia as well as in the European H2020 5G NORMA project. She also led the Self-Organizing Network (SON) Research and Standardization project in Nokia Siemens Networks, and is co-editor of a book, *LTE Self-Organizing Network*. Earlier she worked in the Network Telecom, RRM, and SS7 departments of Nokia Siemens Networks.

SAMULI TURPINEN (samuli.turtinen@nokia-bell-labs.com) is a senior radio research specialist at NOKIA Bell Labs in Oulu, Finland, where he is contributing to 5G RAN architecture and protocol work. He received his M.Sc. from the University of Oulu. He contributed to the standardization of 3GPP RAN1 on device-to-device proximity services and RAN2 on WiFi interworking, enhancement for data applications, in-device coexistence, and others. He is an author of many standard contributions as well as more than 100 patents and patent applications.

It should be emphasized that the main challenge is not to build a system which can cover just the diverse use cases of 5G, but to do so in a scalable and flexible way, such that cost and effort are justified by industry stakeholders and standards community.

5G Radio Access Network Architecture: Design Guidelines and Key Considerations

Patrick Marsch, Icaro Da Silva, Ömer Bulakci, Milos Tesanovic, Salah Eddine El Ayoubi, Thomas Rosowski, Alexandros Kaloxylas, and Mauro Boldi

The authors provide a comprehensive overview of the 5G RAN design guidelines, key design considerations, and functional innovations as identified and developed by key players in the field. They depict the air interface landscape that is envisioned for 5G, and elaborate on how this will likely be harmonized and integrated into an overall 5G RAN.

ABSTRACT

While there is clarity on the wide range of applications that are to be supported by 5G cellular communications, and standardization of 5G has now started in 3GPP, there is no conclusion yet on the detailed design of the overall 5G RAN. This article provides a comprehensive overview of the 5G RAN design guidelines, key design considerations, and functional innovations as identified and developed by key players in the field.¹ It depicts the air interface landscape that is envisioned for 5G, and elaborates on how this will likely be harmonized and integrated into an overall 5G RAN, in the form of concrete control and user plane design considerations and architectural enablers for network slicing, supporting independent business-driven logical networks on a common infrastructure. The article also explains key functional design considerations for the 5G RAN, highlighting the difference to legacy systems such as LTE-A and the implications of the overall RAN design.

INTRODUCTION

After several years of research on fifth generation (5G) wireless and mobile communications, there is broad consensus that 5G will not just be a “business-as-usual” evolution of 4G networks with new spectrum bands, higher spectral efficiencies, and higher peak throughput, but will also target new services and business models. The main 5G service types typically considered are extreme mobile broadband (xMBB, a.k.a. eMBB) with data rates up to several gigabits per second in some areas and reliable broadband access over large coverage areas, massive machine-type communications (mMTC) requiring wireless connectivity for, for example, millions of power-constrained sensors and actuators, and ultra-reliable MTC (uMTC, a.k.a. ultra-reliable and low-latency communications, URLLC) requiring end-to-end latencies of less than 5 ms and 99.999 percent reliability for, say, vehicle-to-anything (V2X) communication [1–3].

Investigations toward an overall 5G radio access network (RAN) architecture that can efficiently support these requirements are still ongoing. This article provides an overview of the

5G RAN design guidelines, key design considerations, and functional innovations identified and developed by key players in the field.¹ Key 5G RAN design requirements are listed, after which the envisioned 5G air interface (AI) landscape is described, including the latest considerations on how different air interface variants (AIVs) may be integrated into one overall AI. The article further captures overall system architecture considerations, such as the logical split between core network (CN) and RAN and related network interfaces, before venturing into key 5G functional design paradigms. Finally, conclusions are drawn.

KEY 5G RAN DESIGN REQUIREMENTS

Due to the diverse and extreme requirements of the main 5G service types, it is clear that the 5G RAN must be designed to operate over a wide range of spectrum bands with diverse characteristics, such as channel bandwidths and propagation conditions [1]. It must further be able to scale to extremes w.r.t. throughput, number of devices, connections, and so on, which is facilitated if the user plane (UP), related to application payload transmission, and control plane (CP), related to control functionality and signaling, are handled individually. For scalability also toward various possible deployments and an evolving application landscape, both 5G RAN and CN must be software configurable, meaning, for example, that it is configurable which logical and physical entities are traversed by CP and UP packets.

A common understanding is that the 5G RAN should allow integrating Long Term Evolution-Advanced (LTE-A) evolution and novel 5G radio on the RAN level, although integration need not always take place on this level.

The 5G RAN should further support more sophisticated mechanisms for traffic differentiation than LTE-A in order to fulfill diverse and more stringent quality of service (QoS) requirements, and facilitate the network slicing vision from Next Generation Mobile Networks (NGMN) [2], enabling independent operation of logical networks for different business cases on a shared physical infrastructure.

Another required feature distinctive from LTE-A is the native and efficient support of

¹ Note that throughout the article, any notion of “it is envisioned” or “it is proposed” refers to the views and proposals of partners of the METIS-II project: www.metis-ii.5g-ppp.eu/

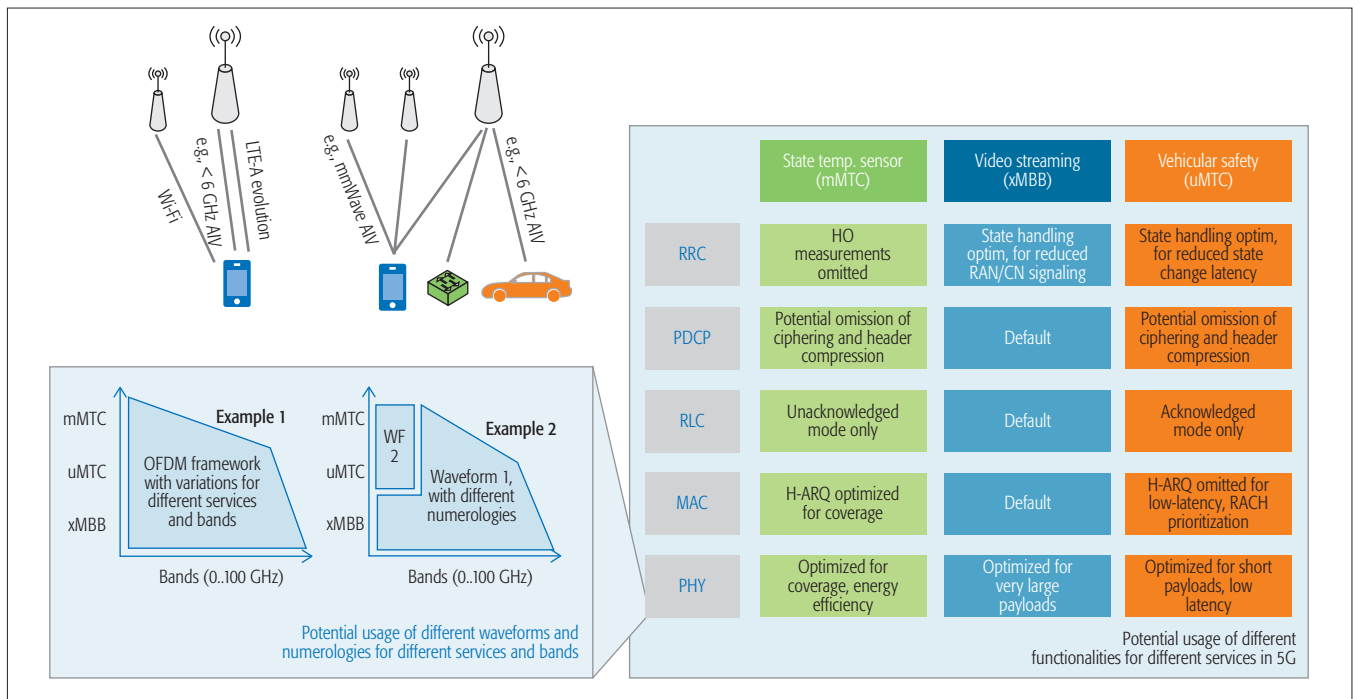


Figure 1. Overall 5G AI envisioned, composed of multiple AIVs with potentially different functionalities for different services and bands.

communication forms like multi-connectivity (e.g., concurrent communications of a device with multiple network nodes) and network-controlled device-to-device (D2D) communication, in the form of point-to-point, multicast, or broadcast communication. The 5G RAN should further support a wide range of physical deployments, from distributed base stations to centralized cloud-RAN deployments or distributed edge clouds. Different types of backhaul and fronthaul shall also be supported with steady performance degradation in case of reduced backhaul or fronthaul quality in terms of delay and capacity. Also, self-backhauling is seen as an important feature, where devices may also act as base stations and self-establish wireless backhaul links to suitable donor base stations.

Finally, the 5G RAN must be highly energy-efficient and future-proof, that is, enabling efficient introduction of new features and services and backward compatibility of devices in future releases.

AIR INTERFACE LANDSCAPE AND INTEGRATION INTO ONE 5G AIR INTERFACE

In order to simultaneously handle the extreme requirements of 5G use cases, an overall 5G AI is envisioned, shown in the top left of Fig. 1, which is composed of LTE-A evolution and novel AIVs tailored to different services, bands, and cell sizes. Also, non-cellular standards like Wi-Fi are considered as part of 5G, although they are not covered in this article. The overall 5G AI is expected to operate over a wide range of spectrum bands, where frequencies below 6 GHz are likely most suitable to support, for example, mMTC services where coverage is important, while spectrum above 6 GHz (up to 100 GHz) is essential to meet xMBB capacity demand. Three

authorization schemes or mixtures thereof are expected to coexist for 5G spectrum usages: primary user mode, licensed shared access (LSA), and unlicensed mode. Exclusive use of spectrum should remain the main and preferred solution, while shared use of spectrum may be a complement to increase spectrum availability, as summarized in [1].

Evolved LTE-A will likely play a pivotal role in the overall 5G AI, with tight interworking with novel AIVs [4], that is, support of inter-AIV mobility with very low interruption times and inter-AIV multi-connectivity. Note that this does not imply LTE assistance, but also stand-alone operation of novel AIVs or their usage as an anchor layer is considered.

Novel 5G AIVs are needed in particular to respond to stringent latency and reliability targets that LTE-A evolution cannot meet, or to enable communication at significantly higher frequencies than in existing systems. For this, some new AIVs are expected to use significantly shorter transmission time intervals (TTIs) and a wider bandwidth compared to LTE. It is also clear that there cannot be a “one-size-fits-all” solution for all novel AIVs. For example, an AIV tailored to lower carrier frequencies, large cell sizes, and high velocity will likely have a physical layer (PHY) designed to be most robust toward delay spread and Doppler spread, whereas an AIV for mmWave frequencies and short-distance communication with limited mobility might instead require robustness toward other impairments like phase noise. Key properties of novel 5G AIVs will be a flexible and scalable numerology (e.g., frequency-dependent TTI length) enabling user- and service-specific adaptations, flexible sub-band configurations, improved spectral and energy efficiency, support for flexible time-division duplex (TDD), reduced out-of-band (OOB)

For the CP, a single set of RRC specifications is envisioned among all novel AIVs. These might differ, however, for evolved LTE-A and novel AIVs in order to simplify the standardization process. Major changes envisioned in the RRC for novel AIVs are the need to support procedures relying on a beam-centric and lean design.

emissions, and support for non-orthogonal multiple access (NOMA) schemes [5].

A key challenge is the integration of different AIVs, including LTE-A evolution, into one overall 5G AI with minimal standardization and implementation complexity, while not sacrificing the performance of individual AIVs. In this respect, it is expected that there will be network functions (NFs) on different protocol stack layers that are tailored to different services (the right side of Fig. 1). For safety-critical vehicular communications, for example, one may need specific medium access control (MAC) functionalities like specialized hybrid automated repeat request (HARQ) for fast multicast communications and prioritized random access channels (RACHs). Together with service-specific higher-layer optimizations like the omission of header compression, the overall latency of emergency message transmission may be substantially reduced [6]. However, it appears desirable to design all NFs such that they can be parametrized from common specifications, and/or be selected from a common pool of atomic functions, for easier standardization and implementation.

Regarding PHY waveforms and numerology, it is envisioned that variations for different bands and services will be required (the bottom left of Fig. 1). It has not yet been concluded whether one would use a single waveform family such as orthogonal frequency-division multiplex (OFDM) parameterized to support all services and bands, or coexistence of different waveforms such as variants of OFDM and filter-bank multi-carrier (FBMC) [7]. However, one should strive for a large extent of harmonization, maximum reuse of signal processing blocks, and so on. While a large degree of NF harmonization among novel 5G AIVs can be enabled by design, the harmonization among novel AIVs and LTE-A evolution is more challenging, as this may pose too stringent limitations on 5G possibilities.

In conclusion, 5G will consist of a diverse set of AIVs, harmonized to a large but not complete extent, requiring novel architectural and functional design paradigms, as covered in the following sections.

OVERALL 5G SYSTEM ARCHITECTURE CONSIDERATIONS

CONSIDERATIONS ON CN, CN/RAN SPLIT, AND NETWORK INTERFACES

In alignment with NGMN [2], it is expected that the majority of CN and service-layer functions are deployed in 5G as virtual network functions (VNFs), running on virtual machines on standard servers, potentially on cloud computing infrastructures (i.e., data centers). Although not necessary, the function design will to some extent explore software-defined networking (SDN) principles such as UP/CP split, and allow for their flexible deployment in operators' networks depending on latency requirements, available transport, processing and storage capacity, and so on. Moreover, different services or network slices may utilize different CN and service-layer VNFs deployed at different network sites.

An important assumption, also considered

by the Third Generation Partnership Project (3GPP) [8], is a logical split between RAN, CN, and service layer functions. This appears beneficial because it allows independent evolution of RAN and CN functionality to accelerate the introduction of new technology, and enables access-agnostic CN functions (e.g., common UP processing). A downside is that cross-CN/RAN optimizations (e.g., collapsing of functionalities) are limited to single-vendor environments.

Regarding the exact logical split between RAN and CN, various changes are being proposed w.r.t. the evolved packet system (EPS) [9]. For example, paging is envisioned to be moved toward the RAN, and a new radio resource control (RRC) state has the potential to simplify the 5G CN, both aspects detailed later.

In previous generations, different radios like 3G and 4G were connected via different CNs and performed CN-based mobility via CN interfaces for handovers and location updates. For 5G, requirements related to latency, reliability, and short interruption times clearly render a tighter interworking between LTE-A evolution and novel AIVs necessary [4] (e.g., enabling RAN-based mobility and multi-connectivity). In this context, it appears beneficial to have a common CN/RAN interface for LTE-A evolution and novel AIVs, herein denoted as S1*, for:

- Fast establishment of multi-connectivity for any user equipment (UE) entering the system via LTE-A or novel AIVs, requiring neither CN/RAN nor non-access stratum (NAS) signaling
- Simplified UE implementation due to a common NAS layer for LTE-A and novel AIVs
- Common context handling enabling dynamic UE activity and roaming across LTE-A and novel AIVs without CN/RAN signaling involved
- Needing only one CN state, reducing the risk of losing state synchronization

One potential drawback of a common CN and CN/RAN interface is the need for a common evolution track for LTE-A and new AIVs, which may be complicated if the 5G CN should diverge strongly from the current design. Note that the envisioned S1* interface would inherently enable tight Wi-Fi integration as in LTE Release 13 via Xw interface.

Also, the existing X2 interface is expected to evolve (denoted herein X2*), in particular to support inter-node multi-connectivity, high-performance mobility, and RAN-based paging. The left side of Fig. 2 illustrates the envisioned logical architecture with mentioned interfaces.

USER PLANE AND CONTROL PLANE ARCHITECTURE

For the UP, a harmonized packet data convergence protocol (PDCP) for LTE-A evolution and novel AIVs can be achieved with a single set of functions captured in a common specification, potentially with different operation modes. This would facilitate the standardization process and the implementation of features such as handover and UP aggregation among AIVs. For aggregation among LTE-A evolution and novel AIVs, a bearer split should be supported where anchoring may occur on either side.

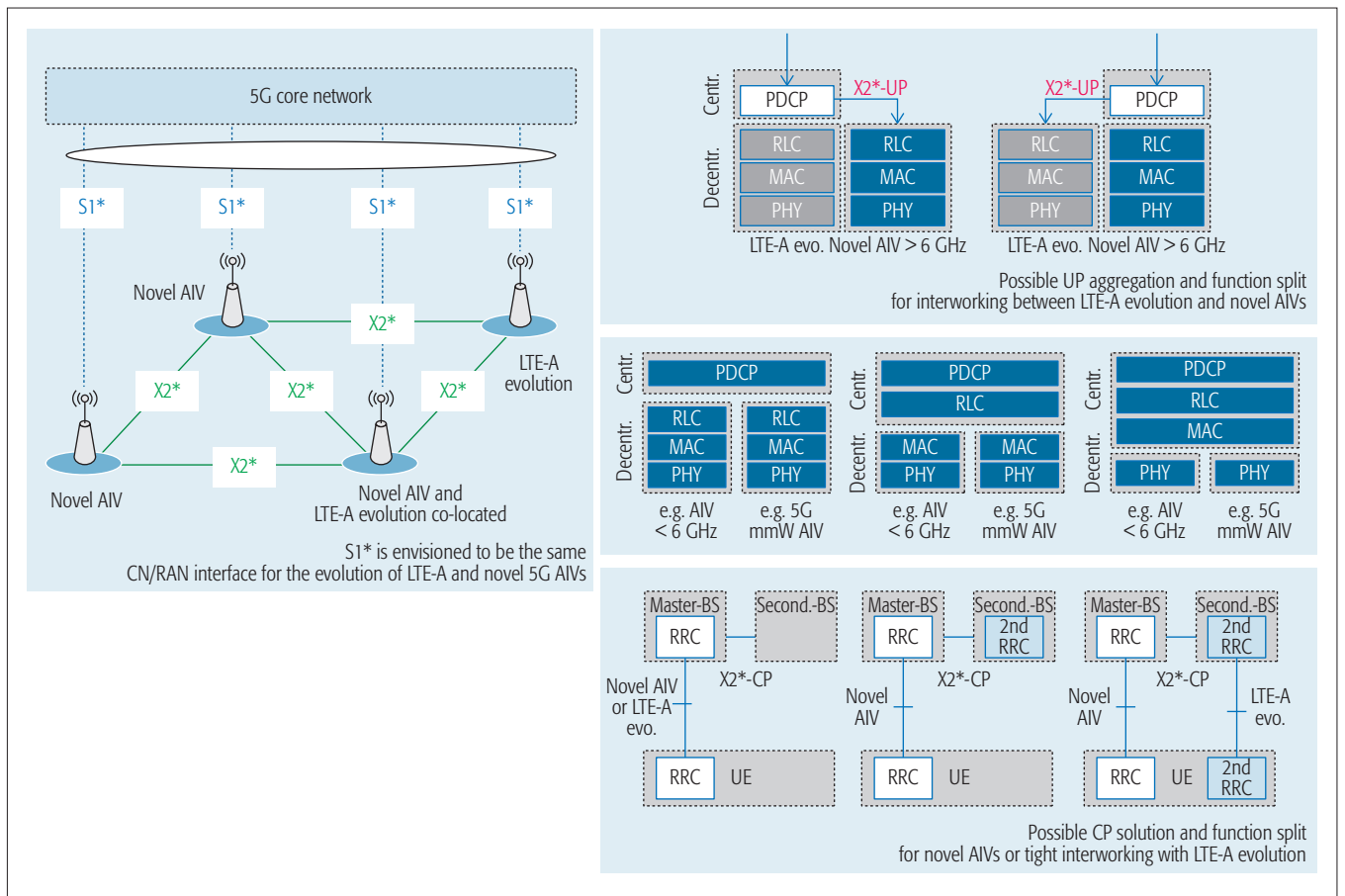


Figure 2. Logical architecture and network interfaces, options for multi-AIV integration, and function splits.

For the CP, a single set of RRC specifications is envisioned among all novel AIVs. These might differ, however, for evolved LTE-A and novel AIVs in order to simplify the standardization process. Major changes envisioned in the RRC for novel AIVs are the need to support procedures relying on a beam-centric and lean design.

For the CP support of multi-connectivity among novel AIVs, a single RRC connection is envisioned (as in LTE dual connectivity), since it accelerates the establishment of a secondary AIV and reduces UE complexity. For multi-connectivity among LTE-A evolution and novel AIVs, different options in Fig. 2 have been investigated, with the current assumption being to have a single RRC connection (i.e., a single state machine at the UE). The stated UP and CP architecture options are shown on the right side of Fig. 2.

DEPLOYMENT ARCHITECTURE, PLACEMENT OF LOGICAL FUNCTIONS, AND RECONFIGURATION

As mentioned before, the 5G RAN should support a wide range of physical deployments and maximally leverage centralization, while also supporting distributed base stations and non-ideal backhaul. A key enabler for this is the implementation of some radio functions as VNFs, as mentioned before in the CN context, allowing flexibly shifting these toward or away from the radio edge depending on the physical architecture and specific application requirements. Preliminary analyses [10] concluded that functions that are time-*asynchronous* to the

radio interface (in LTE these are PDCP and RRC functions related to measurement control and reporting, handover preparation and execution, dual connectivity, random access, RRC state transition, etc.) are mostly suitable to be implemented as VNFs and possibly centralized, as they typically require low data rates on their interfaces, scaling with the number of users and not the overall traffic. Further, these functions can typically cope with larger latency (e.g., tens of milliseconds in LTE).

However, time-*synchronous* functions (in LTE these are PHY, MAC, and radio link control [RLC] functions, e.g., scheduling, link adaptation, power control, interference coordination) typically require high data rates on their interfaces, scaling with the traffic, signal bandwidth, and number of antennas, and benefit from hardware acceleration. The potential for centralization is here most pronounced in deployments with low-latency and high-bandwidth backhaul/fronthaul due to timing and real-time processing requirements.

A key consideration for 5G is to design RAN functions such that strict timing relations between the protocol layers are avoided, and hence multiple possible function splits among physical entities are possible. The right side of Fig. 2 shows some possible function splits, including the aforementioned split between asynchronous and synchronous functionality.

The 5G RAN is envisioned to enable a fast reconfiguration of the mix of supported AIVs and frequency bands, the reconfiguration of the

Beamforming will play an essential role in 5G, in particular in the context of bands beyond 6 GHz, where a larger path loss has to be overcome, but where shorter wavelengths also permit a larger number of antenna elements at reasonable form factor. Different from LTE-A, where only data channels use beamforming, higher bands will also require beamforming of control channels.

related NFs as such, and the mapping of NFs to physical architecture. This reconfiguration functionality is expected to act on the basis of various input parameters, such as available resources (spectrum and hardware), traffic demand, UE capabilities (supported AIVs, frequency bands, etc.), and requested services (e.g., bandwidth and quality of service [QoS]), and could exploit collaboration among network nodes.

RAN ARCHITECTURE SUPPORT FOR NETWORK SLICING

Flexibility and configurability will be key RAN characteristics to support the stated diverse services and related requirements in one common network infrastructure. This may be realized by a protocol architecture supporting a service-specific selection of NFs and service-tailored optimizations, as elaborated earlier. Going beyond the support of diverse services as such, network slicing [2] refers to the operation of multiple independent logical networks for different business cases on a shared physical infrastructure. For instance, a dedicated network slice could be configured for a particular V2X business case, involving service level agreements (SLAs) among stakeholders and a slice-specific selection and configuration of CN, transport network, and RAN functions. It is envisioned that [6]:

- Even if network slices are seen as separate logical networks, efficient reuse of resources like radio spectrum, infrastructure, transport network, and NFs among the slices is essential.
- Network slices (or an abstraction thereof, e.g., groups of service flows) need to be visible to the RAN, such that NFs can take into account overall slice-specific metrics or constraints (e.g., the constraint that all services belonging to a slice may jointly only occupy a certain amount of radio resources).
- Means for slice isolation and protection are needed (e.g., to guarantee that events in one slice cannot negatively impact another slice).
- Performance monitoring solutions need to be aggregated per slice to verify the fulfillment of SLAs and/or properly operate the different businesses associated with different slices.
- Configuration management, self-organizing networks (SON), and so on could be slice-tailored. Additionally, certain functions could be turned on/off and/or possibly configured individually for different slices. This decision should be made before the deployment of a slice; however, the fine-tuning of the configuration of the functions shall be possible during the lifetime of the slice.

One open question related to network slicing is how the system could tell to which slice a newly appearing device or service should belong. This could be solved via network implementation where UE identities are matched with the slice to which they belong. Alternatively, a slice-specific identity could be either transmitted over system information to support the UE to access the proper slice or a slice-aware authentication/attach where the UE has some sort of slice-specific identity.

KEY FUNCTIONAL DESIGN PARADIGMS FOR THE 5G RAN

Beyond the overall system architecture, the envisioned service and band diversity suggests various paradigm changes in the functional design of the 5G RAN, as explained in the following.

BEAM-CENTRIC DESIGN

Beamforming will play an essential role in 5G, in particular in the context of bands beyond 6 GHz, where a larger path loss has to be overcome, but where shorter wavelengths also permit a larger number of antenna elements at reasonable form factor. Different from LTE-A, where only data channels use beamforming, higher bands will also require beamforming of control channels. For analog/hybrid antenna architectures, beam sweeping procedures are proposed for synchronization and reference signals where transmissions do not occur simultaneously over all directions. For initial access, a beam-based system increases inefficiency of broadcasted information, posing a severe issue given the size of the system information in LTE-A. Compression mechanisms are being proposed that minimize the information transmitted in common channels (possibly not beamformed), using dedicated channels for the remaining information when UEs are connected. Mobility reference signals will also require beamforming, and the UE should be able to detect different beams. In addition, due to the volatility of links in higher frequencies, faster and robust beam switching mechanisms should be supported including the UE directly using a pre-configured candidate beam. Clearly, a beam-centric design, which is necessary for higher bands, but may for harmonization reasons be generally introduced in 5G, has a large implication on the overall RAN design [11].

DEDICATED AND SELF-CONTAINED TRANSMISSIONS, AND NEW SCHEMES TO DISTRIBUTE SYSTEM INFORMATION

In LTE-A, mobility and system access procedures rely on “always-on” signals such as synchronization sequences, cell-specific reference signals, and system information. For a lean and future-proof design, active mode mobility should rely, whenever possible, on more dynamically configurable reference signals possibly sparsely transmitted in time and in specific frequency resources. These could be either periodic or configured on demand based on mobility scenarios (highways, semi-static environments, etc.). Besides minimization of broadcast signals, as depicted in Fig. 3a, one should strive to avoid transmitting, for example, reference signals over the entire bandwidth, but instead use self-contained transmissions as shown in Fig. 3b. Here, reference signals used for the estimation of data channels are transmitted jointly with the payload, minimizing the overhead and interference, and are better suited for beam-based transmission and more future-proof (e.g., facilitating the definition of new physical channels when needed) [12].

In particular, in the context of the co-usage of lower and higher frequency bands, new ways to distribute common signals are envisioned, for instance, letting only some nodes, frequency carriers, or AIVs transmit system information or

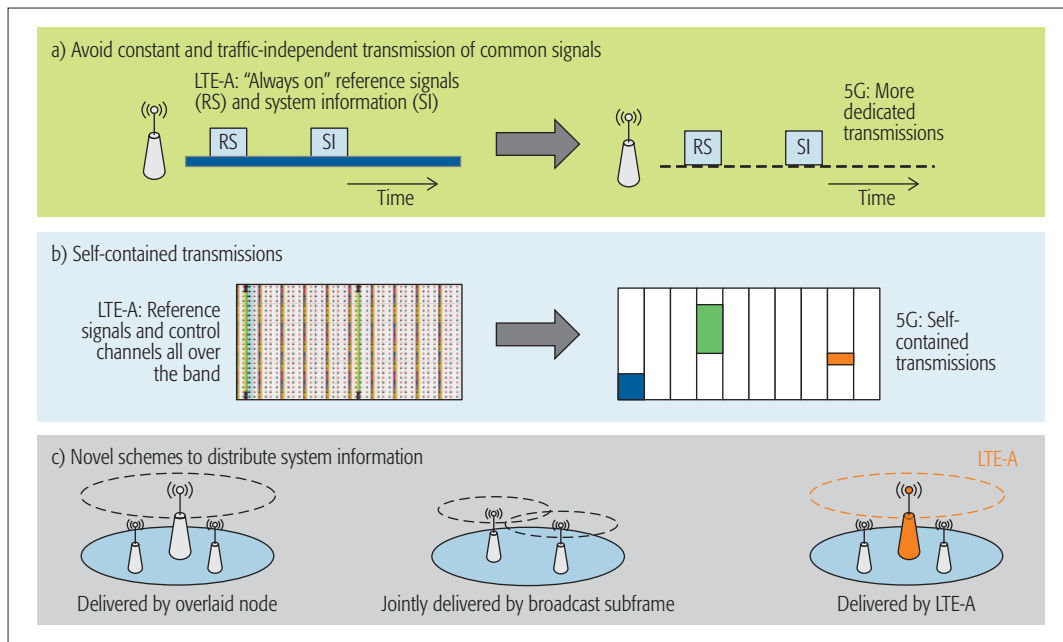


Figure 3. Dedicated and self-contained transmissions, and means to distribute system information in 5G.

parts of it (Fig. 3c). These solutions exploit the fact that multiple carriers and nodes might use exactly the same system information configuration.

SERVICE PRIORITIZATION AT INITIAL ACCESS

In LTE-A, service prioritization is possible through access class barring, where resources are split among services with significantly diverse QoS requirements. A proposed paradigm change in 5G is to provide service differentiation already on the PHY/MAC level. As one specific example, random access requests associated with delay-sensitive services could apply a combination of preamble signatures at a given random access time slot, enabling high-priority service requests to have very small collision rates (up to 100 times less than state-of-the-art access class barring schemes [13]), since they could easily be identified at the receiver side. Without service differentiation at the initial access phase, uncritical services may impair the successful initial access of mission-critical services.

HIERARCHICAL AND DYNAMIC TRAFFIC STEERING AND RESOURCE MANAGEMENT IN 5G

A key challenge in 5G is how to dynamically assign the foreseen wide service range to the different spectrum bands, usage types, related AIVs, and the radio resources therein, also involving novel communication forms like unicast or multicast network controlled D2D. This challenge will further be pronounced through more dynamic network topologies, for instance, resulting from a flexible activation and deactivation of access nodes, antennas and remote radio heads (RRHs), moving cells, and nomadic nodes. Furthermore, interference interdependencies between network nodes will increase, for instance, in dense deployments with TDD and a flexible usage of uplink (UL) and downlink (DL), D2D, or in-band self-backhauling. Consequently, key design principles related to traffic

steering and resource management need reconsideration [13].

A proposed paradigm change in 5G is an extended notion of the term *resource*, considering not only classical radio resources in time and frequency, but also soft network capabilities like processing power, memory/storage capacity and power budget. Also, licensed spectrum resources should be natively complemented by shared (e.g., LSA) or unlicensed resources, where the duration and extent of spectrum sharing may depend on the uncontrolled source of interferers and dynamics in topology changes.

In existing systems, the assignment of services to radio access technologies (RATs, e.g. 3G and 4G) and cells takes place via handover (i.e., RRC-level mechanisms). In 5G, in particular considering more stringent QoS requirements and the relative unpredictability of radio links in higher bands, traffic steering will need to be performed in a more agile way and on a faster timescale.

One specific proposal is a hierarchical traffic steering and resource management approach in the RAN, as illustrated in Fig. 4. Here, the "outer loop" (access network-outer, AN-O) receives the QoS policies from the CN and maintains a cross-AIV 5G RAN view, based on real-time feedback (per-TTI or periodically over a few TTIs) from the AIVs in the "inner loop" (access network-inner, AN-I). This allows the AN-O to do fast traffic re-routing among the AIVs, and also dynamically (de-)activate AIVs and place UEs into suitable multi-AIV multi-connectivity constellations, while the actual TTI-level resource management happens in the AN-I. The exact implementation of the scheme depends on the chosen function split in the RAN. If, for instance, RRC, PDCP, and RLC, are centralized in a cloud RAN, and MAC and PHY are performed in radio units, the former group would naturally constitute the AN-O and the latter the AN-I, as depicted on the right side of Fig. 4. The

A key challenge in 5G is how to dynamically assign the foreseen wide service range to the different spectrum bands, usage types, related AIVs and the radio resources therein, also involving novel communication forms like uni-cast or multi-cast network controlled D2D. This challenge will be further pronounced through more dynamic network topologies.

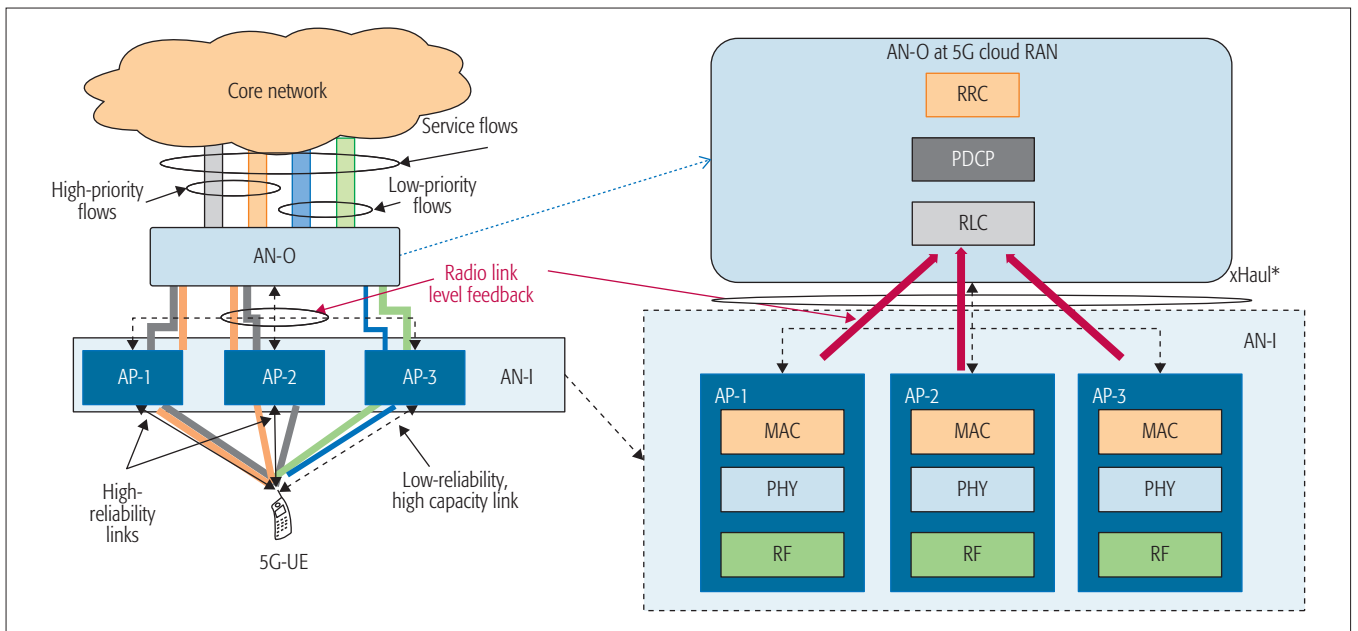


Figure 4. An example scheme for hierarchical traffic steering and resource management in its abstract form (left), and in the context of a RAN function split shown on the right illustrating one specific example, while different split options depending on xHaul* characteristics are possible.

interface between AN-O and AN-I, here denoted as xHaul*, would then have to encompass new RAN measurement information elements for the real-time AIV feedback to the AN-O, and control information elements for the steering of the AN-I resource management schemes by the AN-O. It is still under research to which extent these information elements should be standardized and may possibly also require changes to the Uu interface.

A NOVEL RRC STATE MODEL IN 5G

Moving toward the Internet of Things (IoT), there will be significantly more battery-powered UEs (e.g., sensors, baggage tags, wearable devices) in 5G. Therefore, battery efficiency and duration will be essential, especially for devices with limited accessibility (e.g., in remote locations or restricted areas). However, fast first packet transmission (UL or DL) may be even more important than today. This trade-off between device power efficiency and fast accessibility is often called the “UE sleeping problem.”

For 5G, a novel RRC state model is proposed to address the problem, relying on a novel state, RRC Connected Inactive, in addition to RRC Connected and RRC Idle. This novel state explores the principle of “not discarding previously exchanged information” for inactive UEs, meaning that UEs in the new state still keep parts of the RAN context (e.g., related to security, UE capability information). In addition, signaling is reduced by allowing the UE to move around within a pre-configured area without notifying the network, even among LTE-A evolution and novel AIVs. The new state is envisioned to be highly configurable with a wide range of discontinuous reception (DRX) cycles (from milliseconds to hours) and service-tailored state transitions.

Transitions from RRC Idle to RRC Connected are expected to occur mainly when a UE first

attaches to the network, or as a fallback when devices and/or the network cannot use the previously stored RAN context. However, transitions from RRC Connected Inactive to RRC Connected are expected to occur often and are hence optimized to be fast and lightweight in terms of signaling. This is achieved by keeping the CN/RAN connection alive during inactivity periods and reducing the amount of RRC signaling needed to resume an existing inactive connection via the usage of a RAN context ID — an option inspired by the suspend/resume procedure considered for idle state UEs in LTE [14]. Thanks to the common CN/RAN interface, the UE is able to perform this optimized state transition even when roaming among LTE-A evolution and novel AIVs. The proposed novel state model and expected benefits w.r.t. protocol overhead and CP latency reduction are shown in Fig. 5, with more details and a discussion on similarities and differences of the proposed new RRC state to the URA-PCH state in high-speed packet access (HSPA) given in [15].

RAN-BASED PAGING

Due to the expected massive number of devices and denser deployments in 5G, paging may significantly increase the load on the AI and the CN/RAN interface, thus requiring new solutions for efficient paging and UE location tracking. One proposed option is hierarchical location tracking where the CN tracks UEs in RRC Connected Inactive on the level of groups of RAN locations, whereas the RAN tracks these at cell-level granularity. This can be seen as a combination of the hierarchical location concepts in high-speed packet access (HSPA) with the leaner state model (compared to HSPA) described before [15]. This could involve a lightweight signaling procedure terminated in the RAN, using a security handling mechanism based on retaining and updating the security context from the last attach procedure. The proposed hierarchical

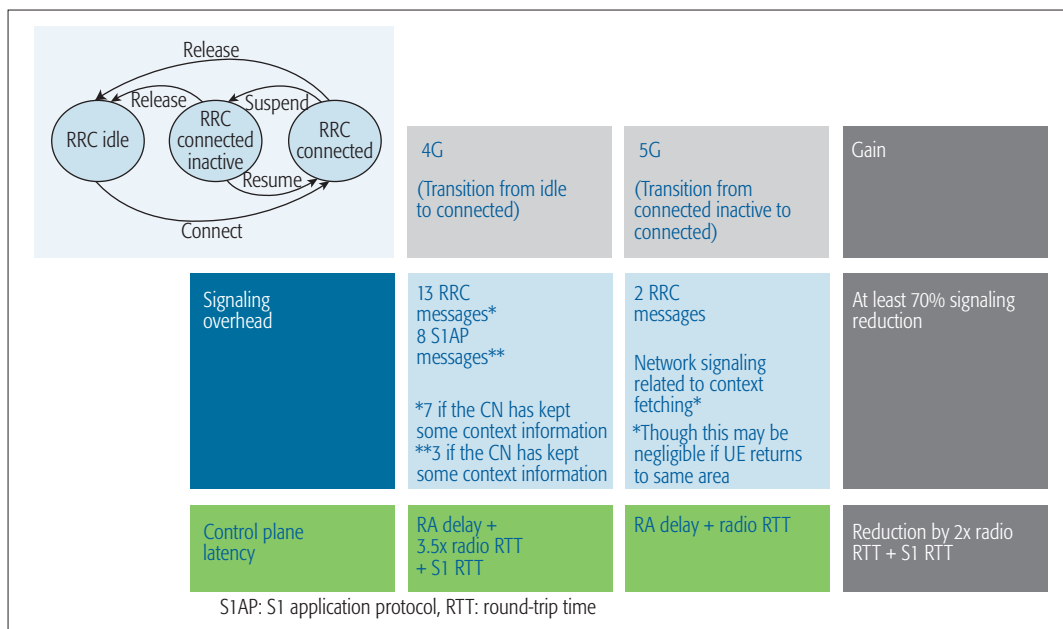


Figure 5. The proposed novel RRC state model and its expected benefits.

5G functional design paradigm	Key benefits	Key difference to LTE-A evolution	Implication on overall RAN design
Beam-centric design	Better coverage, capacity and data rates in higher bands	Narrow beams possibly swept instead of notion of omnidirectional cells	All control signals beamformed. All mobility and initial access procedures need native beam-centric design.
Dedicated and self-contained transmissions, new ways of distributing system information	Energy efficiency and future-proofness, potentially also improved C-plane scalability	Reference signals not always on, not full band, not all subframes	Significantly more configurable reference signals and mobility procedures
Service prioritization at initial access	Service differentiation already at first access, lower latency, and 100x reduction of collision probability for mission-critical services [13]	Different levels of service prioritization for diverse sets of delay requirements	New MAC procedures required for RACH to enable service prioritization. Signaling to higher layers needed.
Hierarchical and dynamic traffic steering and resource management	Faster and more agile resource management, around 20% reduction in mean packet delivery delay [13], higher overall reliability. Essential for uMTC.	Traffic steering performed on comparatively lower protocol stack layer. Dynamic traffic steering instead of handover.	New control information elements and steering options between protocol layers needed. Possible impact on Uu interface.
A novel RRC state model	Reduced UE power dissemination and C-plane latency. Reduction of CN/RAN signaling by at least 70% [15]. Especially suitable for bursty connectivity and massive access.	UEs are always connected from a CN perspective. Significantly larger possibilities for service-specific configuration.	Context fetching needs to be specified and supported. Novel mobility procedures for new state need to be defined.
RAN-based paging	CN/RAN signaling reduced by more than 85%, reduced C-plane latency [13].	In LTE, paging is a CN function, which is now moved into the RAN	Entire re-design of paging functionality, signaling etc. Change of usage of CN/RAN interface.

Table 1. Summary of the key functional design paradigms considered.

location tracking approach would imply moving some paging and mobility related functionalities from CN to RAN, hence changing the CN/RAN split compared to EPS.

Table 1 summarizes the key functional design paradigms explained before, their key benefits, difference from LTE-A, and their implication on the overall 5G RAN design.

CONCLUSIONS

This article has provided a consolidated view of the likely overall 5G RAN architecture, functions, and interfaces, enabling integration of

LTE-A evolution and multiple novel AIVs in 5G. Besides providing details on the likely logical architecture and its mapping to physical architecture as well as means to support network slicing, the article has ventured into key functional paradigm changes that are proposed (e.g., related to a beam-centric design, lower-layer service prioritization, traffic steering in 5G, and a novel RRC state model), emphasizing the key differences from LTE-A and the main RAN design implications. It has to be noted that the topics covered in this article are still under research and subject to finalization in the next months.

Due to the expected massive number of devices and denser deployments in 5G, paging may significantly increase the load on the AI and the CN/RAN interface, thus requiring new solutions for efficient paging and UE location tracking.

ACKNOWLEDGMENT

This work has been performed in the framework of the 5G PPP project METIS-II co-funded by the European Union. The views expressed are those of the authors and do not necessarily represent the project. The authors would like to thank Gerd Zimmermann, Panagiotis Spapis, and Athul Prasad for their contributions and suggestions.

REFERENCES

- [1] ICT-671680 METIS-II, Deliv. 1.1, "Refined Scenarios and Requirements, Consolidated Test Cases, and Qualitative Techno-Economic Assessment," Jan. 2016.
- [2] NGMN Alliance, "NGMN 5G White paper," Feb. 2015.
- [3] ITU-R Working Party WP 5D: Draft New Rec. ITU-R M.2083-0 (09/2015), "IMT Vision – Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," Sept. 2015.
- [4] 3GPP TR 38.913, "Study on Scenarios and Requirements for Next Generation Access Technologies," Mar. 2016.
- [5] L. Dai *et al.*, "Non-Orthogonal Multiple Access for 5G: Solutions, Challenges, Opportunities, and Future Research Trends," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 74–81.
- [6] I. Da Silva, G. Mildh *et al.*, "On the Impact of Network Slicing on 5G Radio Access Networks," *Euro. Conf. Networks and Commun.*, Athens, Greece, June 2016.
- [7] M. Tesanovic *et al.*, "Towards a Flexible Harmonised 5G Air Interface with Multi-Service, Multi-Connectivity Support," ETSI Wksp. Future Radio Technologies: Air Interfaces, Jan. 2016.
- [8] 3GPP S2-153651, "Study on Architecture for Next Generation System," Oct. 2015.
- [9] 3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2," Nov. 2015.
- [10] ICT-317669 METIS, Deliv. 6.4, "Final Report on Architecture," Jan. 2015.
- [11] F. Athley *et al.*, "Providing Extreme Mobile Broadband Using Higher Frequency Bands, Beamforming and Carrier Aggregation," *IEEE Int'l. Symp. Personal, Indoor and Mobile Radio Commun.*, Aug. 2015, pp. 1370–74.
- [12] P. Frenger, M. Olsson, and E. Eriksson, "A Clean Slate Radio Network Designed for Maximum Energy Performance," *IEEE Int'l. Symp. Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sept. 2014, pp. 1300–04.
- [13] ICT-671680 METIS-II, Deliverable 2.2, "Draft Overall 5G RAN Design," June 2016.
- [14] 3GPP TR 23.720, Technical Specification Group Services and Systems Aspects, "Architecture Enhancements for Cellular Internet of Things (Rel. 13)," Nov. 2015.
- [15] I. Da Silva *et al.*, "A Novel State Model for 5G Radio Access Networks," *IEEE ICC Wksp.*, May 2016, pp. 632–37.

BIOGRAPHIES

PATRICK MARSCH (patrick.marsch@nokia-bell-labs.com) received his Dipl.-Ing. and Dr.-Ing. degrees from Technische Universität Dresden, Germany, in 2004 and 2010, respectively. After leading a research group at TU Dresden, Germany, since 2011 he has been heading a research department within Nokia Bell Labs (formerly Nokia Siemens Networks), Wrocław, Poland. He has published 60+ journal or conference papers, co-edited 2 books, received 4 best paper awards, and received the Philipp Reis Prize for pioneering work in the field of coordinated multipoint. He has been General Chair of various IEEE workshops, and has initiated the ULTRA² series of workshops on ultra-low latency and ultra-high reliability in wireless communications. He is the Technical Manager of the 5G PPP project METIS-II.

ICARO DA SILVA is a senior researcher working at Ericsson, Stockholm, Sweden, since 2010. He is currently working as a 3GPP delegate driving technical areas such as 5G mobility and initial access. He was a Work Package Leader in the 5G-PPP project METIS-II responsible for the Overall Control Plane Design for the 5G architecture. He is also a co-author of the chapter about 5G architecture in the book *5G Mobile and Wireless Communications Technology* and

an author of many conference/journal papers in various areas from mobile networks to signal processing and physical layer. He is also an active inventor in the area of mobile networks, architecture, and network management with more than 100 patents. He received his M.Sc. from the Universidade Federal do Ceará, Brazil, in 2007 and since then he has been involved in R&D projects with Ericsson.

ÖMER BULAKCI received his B.Sc. degree in electrical and electronics engineering from the Middle East Technical University in 2006, his M.Sc. degree in communications engineering from the Technical University of Munich in 2008, and his doctoral degree in communications engineering from Aalto University in 2013. From 2009 to 2012, he conducted his research activities with emphasis on relay networks and standardization of relaying at Nokia Siemens Networks, Germany. Since October 2012, he has been conducting research toward 5G and contributing to EU flagship projects METIS and METIS II at Huawei Technologies ERC, Germany. He is an author of 50+ publications and is an inventor of 10+ patent applications. He is currently leading the agile resource management framework in METIS II.

MILOŠ TESANOVIĆ [SM] is principal 5G researcher at Samsung Electronics R&D Institute UK, focused on coordinating collaborative European R&D activities and conducting research into next-generation mobile systems. He is currently leading the 5G-PPP/H2020 METIS-II project's Work Package on 5G air interface design. He obtained his Ph.D. in 2008 from the University of Bristol, United Kingdom, for his work on robust wireless transmission of video over MIMO systems, and his graduate engineering (Dipl. Ing.) degree in electronics, telecommunications, and control in 2004 from the University of Belgrade, Serbia. He has filed more than 35 patents/patent applications, published over 20 refereed journal and international conference papers, and made active contributions to PHY and MAC standardization of LTE and LTE-A.

SALAH EDDINE EL AYOUBI received his Ph.D. and habilitation degrees from University Paris 6, France, in 2004 and 2009, respectively. Since 2004, he has been working at Orange Labs as a senior research engineer. His research interests include design and performance evaluation of mobile networks, and his current focus is on 5G RAN design. He is active in several EU research projects including 5G PPP METIS-II, where he acts as a Work Package Leader, and 5G-PPP FANTASTIC-5G project where he acts as Technical Coordinator.

THOMAS ROSOWSKI received his Diploma degree in electrical engineering from the University for Applied Science in Bingen, Germany, in 1986. He started his career in the area of broadcasting, got experience in international mobile network standardization, also holding posts as Chair and Secretary of Working Groups in ETNO and ETSI. Since 2009 he has been with Deutsche Telekom Laboratories where his areas of work include wireless and mobile technologies and networks, and radio spectrum management. In addition to company internal projects, he participated in the EU funded research projects E3 and METIS, and is currently involved in METIS-II. He is co-author of a number of publications on radio spectrum regulation and mobile network architectures.

ALEXANDROS KALOXYLOS [SM] received his B.Sc. from the University of Crete in 1992, his M.Phil. from Heriot-Watt University in 1994, and his Ph.D. from the University of Athens in 1999. He has participated in numerous EU projects and has published over 110 papers in the area of mobile communications since 1994. He is an Editorial Board member of several journals and a TPC member for numerous conferences. He is an assistant professor at the Department of Informatics and Telecommunications of the University of Peloponnese in Greece. Since June 2014, he is a principal researcher at Huawei's German Research Center in Munich, where he is currently the head of Radio Access Network Department.

MAURO BOLDI graduated at Politecnico di Torino, Italy, in 1998. He joined the Telecom Italia research centre (CELT at that time) in the same year, where he has been involved in radio access network topics, with activities on antenna and propagation, radio over fiber solutions, collaborative networks, and energy efficiency. He is an author of patents and papers on the same topics, and he has been involved in many European projects, including Winner (where he led the activity on CoMP), Artist4G, Earth, and Metis/Metis-II. He is also a co-author of books on LTE-Advanced and 5G after the projects Winner and METIS. He is Telecom Italia delegate in the ETSI EE Technical Committee, where he plays the role of Rapporteur of the European Specification on Energy Efficiency for Radio Access Networks and Vice-Chairman of the EEPS group.

Spectrum Pooling in MmWave Networks: Opportunities, Challenges, and Enablers

Federico Boccardi, Hossein Shokri-Ghadikolaei, Gabor Fodor, Elza Erkip, Carlo Fischione, Marios Kountouris, Petar Popovski, and Michele Zorzi

ABSTRACT

Motivated by the specific characteristics of mmWave technologies, we discuss the possibility of an authorization regime that allows spectrum sharing between multiple operators, also referred to as *spectrum pooling*. In particular, considering user rate as the performance measure, we assess the benefit of coordination among networks of different operators, study the impact of beamforming at both base stations and user terminals, and analyze the pooling performance at different frequency carriers. We also discuss the enabling spectrum mechanisms, architectures, and protocols required to make spectrum pooling work in real networks. Our initial results show that, from a technical perspective, spectrum pooling at mmWave has the potential to use the resources more efficiently than traditional exclusive spectrum allocation to a single operator. However, further studies are needed in order to reach a thorough understanding of this matter, and we hope that this article will help stimulate further research in this area.

INTRODUCTION

The demand for mobile wireless services is predicted to increase significantly in the coming years. The scarcity of available microwave spectrum, which cannot satisfy this increased demand, has led to the emergence of millimeter-wave (mmWave) as the new frontier of wireless communication. Recently, as part of the harmonization process that will lead to a new mobile spectrum, the 2015 World Radio Conference selected different bands, ranging from about 24 GHz to 86 GHz, for further studies for use in future fifth generation (5G) systems. Unfortunately, the availability of spectrum for mobile services presents limitations even at mmWave frequencies, particularly if one considers the requirements of other systems that may also use these bands in the future, including satellite and fixed services. This is further exacerbated if we also consider the need to license mobile bands to multiple operators and thereby foster healthy competition in the market. Therefore, it is essential to seek an optimal use of the spec-

trum, with the ultimate goal of maximizing the benefits for citizens.

The type of access scheme plays a fundamental role in achieving efficient usage of the spectrum. Spectrum sharing allows multiple service providers to access the same band for the same or different uses. This article investigates the case of spectrum sharing for the same use — mobile services — between different mobile operators, also referred to as *spectrum pooling*. The specific features of mmWave frequencies, for example, the propagation characteristics and the operation based on directional beamforming, are expected to be critical enablers for spectrum pooling, but also call for judiciously designed new paradigms.

Spectrum pooling has recently been considered for cellular systems at microwave frequencies. For example, in [1, references therein], it was shown that orthogonal spectrum pooling, whereby frequency channels are dynamically but exclusively allocated to one operator at a time, results in significant throughput gains, on the order of 50–100 percent. In addition, if frequency channels can be allocated simultaneously to multiple operators, called non-orthogonal spectrum pooling, further gains can be obtained. To achieve these gains, coordination mechanisms may be required, both within an operator, hereafter called intra-operator coordination, and among networks of different operators, hereafter called inter-operator coordination.

Two main different architectural approaches have been investigated to enable spectrum sharing, with or without sharing of the radio access network (RAN) infrastructure, respectively. The benefits of spectrum pooling with RAN sharing are discussed in [2]. In the context of microwave heterogeneous networks (HetNets), it was shown in [3] that a RAN sharing strategy might be optimal for small cells. When spectrum pooling is used without RAN sharing, interference becomes the main limiting factor, and traditional interference management techniques may lead to suboptimal spectrum utilization. Therefore, interference-aware techniques have been studied, and the benefits of spectrum pooling with smart scheduling have been discussed in, for example, [4].

The authors discuss the possibility of an authorization regime for mmWave networks that allows spectrum sharing between multiple operators, also referred to as spectrum pooling. In particular, considering user rate as the performance measure, they assess the benefit of coordination among networks of different operators, study the impact of beamforming at both base stations and user terminals, and analyze the pooling performance at different frequency carriers.

G. Fodor was funded by Wireless@KTH and his work is part of the Wireless@KTH project BUSE.

The work of E. Erkip was partially supported by the National Science Foundation under Grant 1547332.

The work of M. Zorzi was partially supported by NYU-Wireless and by the Villum Foundation, Denmark.

The work of P. Popovski was partially supported by the ERC Consolidator Grant, Horizon 2020.

Federico Boccardi is with Ofcom. However, his work for this article was carried out in his personal capacity; the views expressed here are his own and do not reflect those of his employer; Hossein Shokri-Ghadikolaei and Carlo Fischione are with KTH Royal Institute of Technology; Gabor Fodor is with the KTH Royal Institute of Technology and also with Ericsson Research; Elza Erkip is with New York University; Marios Kountouris is with Huawei Technologies Co. Ltd.

However, the views expressed here are his own and do not reflect those of Huawei; Petar Popovski is with Aalborg University;

Michele Zorzi is with the University of Padova and Aalborg University.

Digital Object Identifier: 10.1109/MCOM.2016.1600191CM

With an uncoordinated approach, network operators do not exchange any information and make independent decisions on how to allocate spectrum. However, common rules need to be in place in order to ensure equitable spectrum allocation, in a way that each operator has the same probability to access spectrum.

Recent works have also considered the benefits of spectrum sharing at mmWave frequencies. In [5], a mechanism that allows two different IEEE 802.11ad access points to transmit over the same time/frequency resources was proposed. This is realized by introducing a new signaling report broadcast by each access point, thereby facilitating the establishment of an interference database to support scheduling decisions. A similar approach was proposed in [6] for mmWave cellular systems, with both centralized and distributed inter-operator coordination. In the centralized case, a new architectural entity receives information about the interference measured by each network and determines which links cannot be scheduled simultaneously. In the decentralized case, the victim network sends a message to the interfering network with a proposed coordination pattern, which can be further refined via multiple stages. In [7], the authors studied the feasibility of spectrum pooling in mmWave cellular networks, and showed that under certain conditions (e.g., ideal antenna pattern), spectrum pooling could be beneficial, even without any coordination among different operators.

The aim of this article is twofold. First, we aim to discuss the technical enablers required to make spectrum pooling work under realistic assumptions and constraints. We discuss the trade-offs among the architectural solutions, the type of coordination, the amount and type of information exchange required, and the new enabling functionalities. We argue that further works are needed to assess other spectrum access regimes, for example, those built on the aggregation between licensed and license-exempt spectrum. Second, we aim to present technical evidence, by means of simulation results, that reveals under which assumptions and conditions spectrum pooling at mmWave frequencies is beneficial. To this end, we start from the studies in [6–8] and substantially extend them. In contrast to [6], where the emphasis was on the multiple access control (MAC) layer, we jointly consider physical and MAC layers. The emphasis in [7] was on the physical layer without coordination, whereas we consider the effects of both intra- and inter-operator coordination. We show that while coordination may not be needed under ideal assumptions, it does provide substantial gains when considering realistic channel and interference models and antenna patterns. Moreover, we evaluate the impact of beamforming, antenna array size, different carrier frequencies, and different base station (BS) densities on the pooling performance.

Finally, we note that this work makes an important contribution toward answering the following fundamental question related to future mmWave networks: *For a given amount of spectrum for mobile applications at a given mmWave frequency, what is the access scheme that allows its optimal utilization?* This work provides to standardization bodies a discussion on different aspects related to this important question and an initial set of performance assessments.

PROTOCOL AND ARCHITECTURAL ENABLERS

In the following, we discuss different coordination types, supporting architectures, and functional enablers of spectrum pooling in mmWave networks.

TYPES OF SPECTRUM POOLING COORDINATION

Coordination is distributed when decisions are made in each operator domain, aided by the supporting information that can be exchanged prior to resource allocation decisions. For example, a participating operator may report the set of subbands, resource blocks, or beams within the shared spectrum pool that are not used within a geographical area or in a given cell (e.g., due to low traffic load). In addition, information between participating operators may be exchanged in a reactive fashion. For example, high interference levels measured in subbands within the shared spectrum pool can trigger a request from a participating operator to its peer operator to reschedule some of the served traffic to other resources. Such inter-operator distributed coordination schemes can be seen as extensions of the RAN sharing scenarios studied by the Third Generation Partnership Project (3GPP) in TR 22.852 [9]. With centralized coordination, the actions are decided by a logical central entity, such as a spectrum broker [10] or a module for making network policy, supported by a network-wide database [11]. We note that the feasibility of centralized coordination also depends on the latency of the exchanged information. If this latency is sufficiently low, any architecture that supports distributed coordination can support centralized coordination as well, by electing one of the distributed entities as a leader and thus a logical coordination center. We further discuss architectural aspects in the next section.

With an uncoordinated approach, network operators do not exchange any information and make independent decisions on how to allocate spectrum. However, common rules need to be in place in order to ensure equitable spectrum allocation in such a way that each operator has the same probability to access spectrum.

Inter-operator coordination can be near real time (operating on a timescale of hundreds of milliseconds) or long-term (operating on a timescale of seconds, minutes, or even coarser scales). In the first case, BSs coordinate at the level of resource block scheduling. Such near-real-time coordination is similar to the X2-application-protocol-based intra-operator mechanisms, which can be deployed in Long Term Evolution (LTE) networks, including inter-BS signaling schemes on traffic load and high interference indications [12]. In the second case, spectrum usage is supported by information exchange on a coarse timescale, typically implemented as part of the operation and maintenance (O&M) infrastructure in each participating operator's network. This type of long-term spectrum usage coordination can operate on the basis of an inter-operator agreement on a usage portion of the pooled spectrum resources during different times of the day or associated with specific days of the week. It can also include a maximum level of energy emissions on specific parts of the spectrum pool. Long-term spectrum usage coordination can be conveniently realized by information exchange at the O&M level rather than employing protocol messages between RAN nodes, as in the case of near-real-time coordination schemes.

SUPPORTING ARCHITECTURES

The different spectrum pooling mechanisms discussed above can be implemented through different supporting architectures. Figure 1 gives a high-level summary of the main alternatives as follows.

Interface at the RAN: The alternative in Fig. 1a refers to the introduction of a new interface (or an extension of the X2 interface used for LTE [12]) between BSs belonging to different networks to enable distributed coordination. From a logical architectural perspective, the alternative in Fig. 1a allows fast information exchange between two different networks, and therefore near-real-time spectrum pooling is possible.

Interface at the core network (CN): The alternative in Fig. 1b refers to an architecture where the interface between the different networks is at the CN. Due to the latency involved, Fig. 1b does not enable real-time spectrum pooling. On the other hand, CN-level coordination can handle a large number of cellular BSs by exchanging a few protocol messages, since typically a large number of BSs are associated with a few CN nodes.

RAN sharing: The alternative in Fig. 1c refers to an architecture where two or more network operators share the BSs [8]. In other words, a single baseband unit serves users associated with the different network operators in the sharing agreement. As resource allocation and scheduling decisions are made by a single unit, Fig. 1c is an effective way to implement real-time centralized coordination, as the coordination-related processing is handled by a single physical entity.

CN sharing: The alternative in Fig. 1d refers to an architecture where two or more network operators share the CN. In the same way as Fig. 1c, Fig. 1d allows centralized coordination. However, in this case real-time spectrum pooling is not possible for the same reasons discussed for Fig. 1b.

Spectrum broker: The alternative in Fig. 1e refers to an architecture where coordination is implemented by means of a spectrum broker. A spectrum broker is a central resource management entity that grants spectrum resources on an exclusive basis during some time window [10].

Uncoordinated spectrum pooling: The alternative in Fig. 1f refers to the case where the network operators do not coordinate. When the number of networks in the pool is not limited, uncoordinated spectrum pooling is reminiscent of a license-exempt regime. For example, in wireless LANs (e.g., Wi-Fi) real-time spectrum pooling is realized through uncoordinated operation. As mentioned above, common rules need to be in place.

The spectrum pooling mechanisms, specifically the architectural solutions of Fig. 1, may impact the RAN sharing mechanisms and deployment options of currently evolving networks operating in bands below 6 GHz [9]. This is because future mmWave networks will tightly interwork with networks deployed for operation below 6 GHz to boost the capacity of congested macrocells while supporting seamless mobility both locally and with the overlaid cellular network [13].

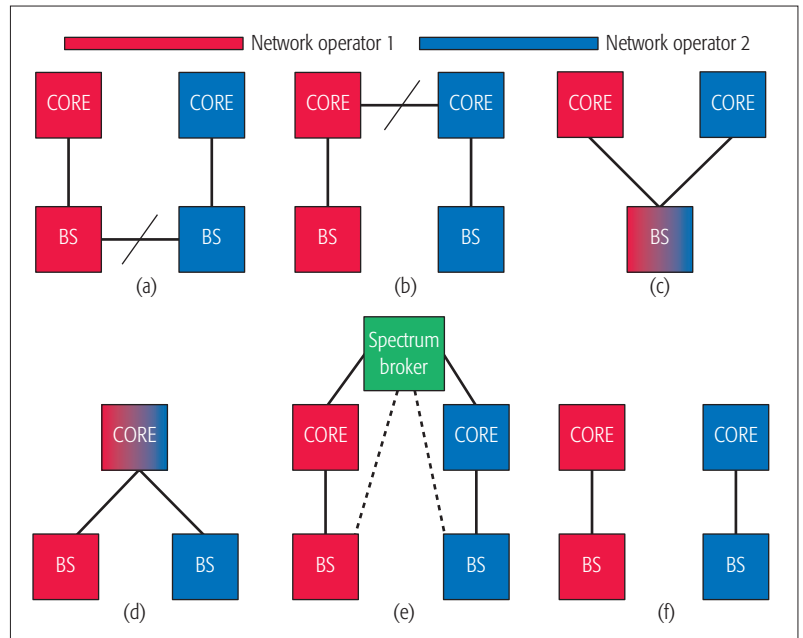


Figure 1. Architectural solutions supporting spectrum pooling between two different network operators: a) interface at the RAN (BS); b) interface at the CN; c) RAN sharing; d) CN sharing; e) via a spectrum broker; f) uncoordinated.

SUPPORTING FUNCTIONS

Spectrum pooling mechanisms require different supporting functions depending on the type of coordination and the architectural solution. In the following we discuss the most prominent supporting functions.

Spectrum sensing: Spectrum sensing and dynamic frequency/channel selection (DFS/DCS) are solutions in which systems participating in spectrum pooling dynamically select their operating frequency range based on measurement results [10, 11]. These measurements can be overall energy or reference signal detection. DFS/DCS are typically not considered as a reliable method due to the well-known hidden node problem. However, spectrum sensing and DFS/DCS can have a supporting role, for example, to identify spectrum subbands that have the least instantaneous traffic load, such that sharing overhead is minimized. Spectrum sensing may be supported by new radio interface capabilities such as the capability at the UE of sensing interference originated by a non-serving network that participates in the spectrum pool. We note that the lower the level of coordination between different network operators, the more important the role of spectrum sensing and DFS/DCS.

Enhanced channel state information (CSI) acquisition and exchange techniques that help learn the channel at both transmitters and receivers. Given the importance of accurate narrow beamforming as an enabler of spectrum pooling and the sensitivity of precoding to reference signal contamination, participating networks may create clean subbands used for training signals. For example, networks operating in time-division duplexing (TDD) mode and relying on channel reciprocity to acquire CSI both for UL reception and DL precoding may reserve their own (not shared) pilot sequences

Spectrum sensing may be supported by new radio interface capabilities such as the capability at the UE of sensing interference originated by a non-serving network that participates in the spectrum pool. We note that the lower the level of coordination between different network operators, the more important the role of spectrum sensing and DFS/DCS.

	Type of coordination	Time resolution	Supporting function required	Information exchange overhead
Interface at the RAN	Distributed	Real-time	Enhanced CSI, distributed synchronization	High
Interface at the CN	Distributed	Long-term	Enhanced CSI	Low
RAN sharing	Centralized	Real-time	Enhanced CSI	–
CN sharing	Centralized	Long-term	Enhanced CSI	–
Spectrum broker	Centralized	Long-term	Enhanced CSI	Low
Uncoordinated	Uncoordinated	–	Enhanced CSI, spectrum sensing and DFS/DCS	–

Table 1. Summary of the characteristics of the different architectural solutions.

that ensure code-domain orthogonality between participating operators. The different networks in the coordination pool might need to exchange some of the CSI acquired from user equipments (UEs). In general, there is a trade-off between the CSI accuracy and the required level of coordination, as a less accurate CSI requires tighter coordination.

Mechanisms to synchronize the BSs of participating operators within a geographical area. As noted in [1], similarly to coordinated multipoint (CoMP) systems, maintaining time and frequency synchronization across the BSs is beneficial in spectrum pooling systems. This is because a common notion of time and frequency enables the use of the same subcarrier spacing and symbol timing, which, in turn, are essential for the creation of clean subbands for the training signals mentioned above.

In Table 1, we summarize the characteristics of the different architectural solutions and link them to the type of coordination, time resolution, required supporting functions, and information exchange overhead.

PERFORMANCE ASSESSMENT

This section presents our initial assessments of spectrum pooling performance at mmWave, in terms of UE rate enhancement. The analysis is based on ideal assumptions on the channel estimation (no error) and coordination (no delay) and is aimed at unveiling the potential of spectrum pooling, rather than quantifying the gains in a realistic setup.

As a starting point, we note that with spectrum pooling each UE has access to a larger bandwidth at the expense of a potentially lower signal-to-interference-plus-noise ratio (SINR), due to higher inter-operator interference and noise power. Inter-operator interference can be tackled in two complementary ways: either narrower beams or inter-operator coordination. Under the assumption of a constant array size, the use of higher frequencies makes it possible to deploy more antenna elements per array at both BSs and UEs, hence increasing the beamforming gain. The gains of inter-operator coordination are more complex and are a function of the supporting architecture (Table 1). In this section, we consider a centralized implementation based on joint beamforming and user association. Later on we discuss the impact of more practical coordination schemes.

SIMULATION SCENARIOS

We consider an mmWave system where antenna and channel models are as in [14], and BSs and UEs are randomly distributed on the plane as in [7]. In particular, we consider a clustered channel model, characterized by a random number of clusters of multiple paths between a transmitter and a receiver. The parameters of the path loss of every cluster depend on whether or not line-of-sight propagation is available between the transmitter and the receiver, and are as in [14]. Moreover, locations of the BSs and UEs follow independent homogenous Poisson point processes with given densities. Without loss of generality, we assume analog beamforming at both the BSs and the UEs, four operators, and a total bandwidth of 1200 MHz at 32 GHz and 73 GHz. We assume TDD and perfect channel estimation at the BS. We consider three spectrum pooling scenarios:

- Exclusive: Each operator uses a 300 MHz exclusive bandwidth.
- Partial pooling: Operators 1 and 2 share the first 600 MHz, and operators 3 and 4 share the second 600 MHz.
- Full pooling: All operators share the whole 1200 MHz bandwidth.

We also consider two coordination scenarios:

- Baseline, without inter-operator coordination: Only BSs belonging to the same operator coordinate to perform joint user association and beamforming.
- Inter-operator coordination: Coordination is extended to BSs belonging to different operators.

We note that we use an ideal coordination scheme with the aim of providing a performance upper bound. In particular, we consider real-time centralized coordination (Table 1) where a central entity using a brute force algorithm jointly selects users and calculates analog beams so as to maximize the user rates based on a proportionally fair criterion. Power is equally allocated among the different RF chains. The central entity is assumed to have perfect knowledge of the long-term channel parameters for each user and each BS in the network, and of the load of each BS. We assume that different BSs are synchronized, and that there is no delay in the interface between the BSs and the central entity. The details of various coordination strategies and corresponding association optimization problems are formally presented in [15].

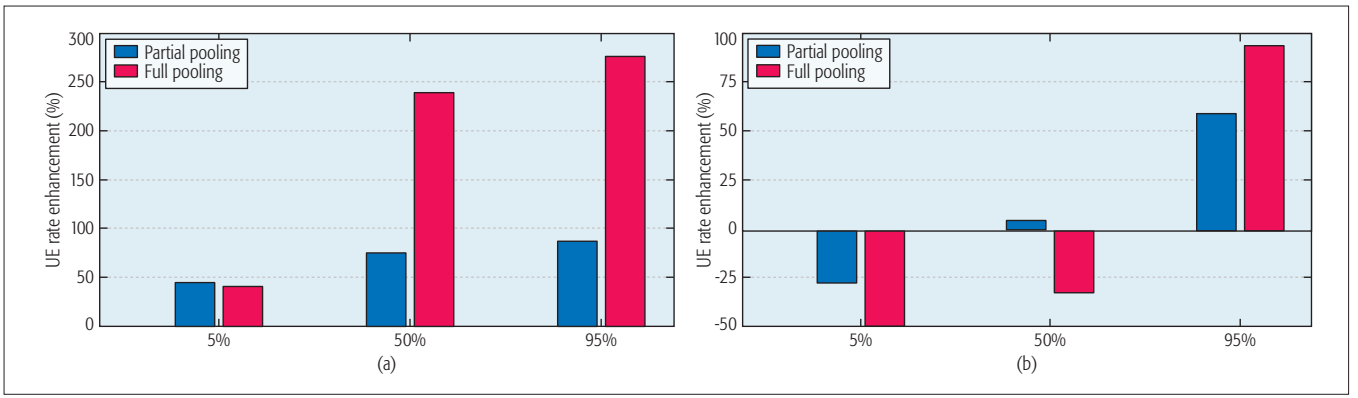


Figure 2. Pooling performance at 32 GHz, under the assumption of no inter-operator coordination. The baseline is a system with exclusive spectrum allocation. x-label indicates the 5th, 50th, and 95th percentiles of the UE rate: a) 4×4 UPA at the UE; b) omnidirectional antenna at the UE.

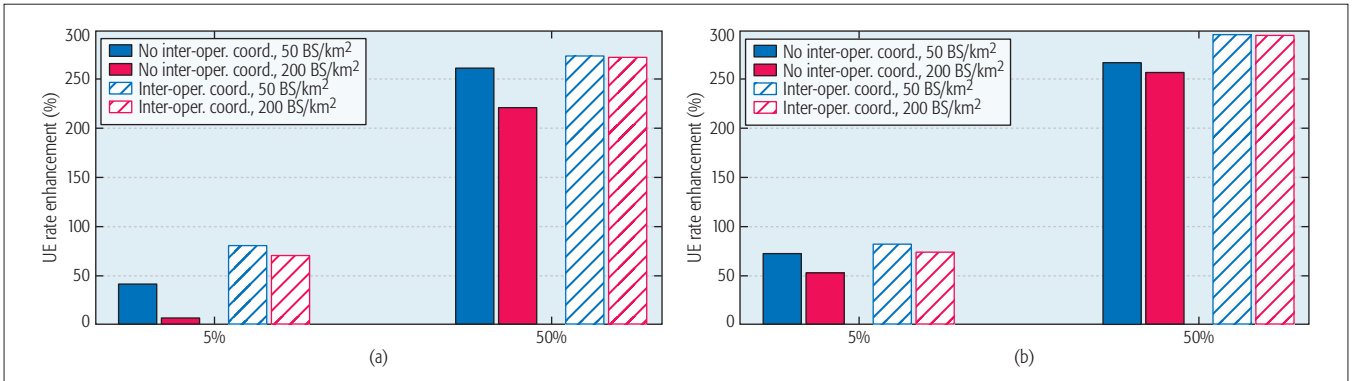


Figure 3. Full pooling performance, w/o and with inter-operator coordination. x-label is the 5th and 50th percentiles of the UE rate: a) 32 GHz, 4×4 UPA at the UE, 32×32 UPA at the BS; b) 73 GHz, 8×8 UPA at the UE, 64×64 at the BS.

SIMULATION RESULTS

Figure 2 illustrates the gain of partial and full pooling with respect to exclusive spectrum allocation, under the assumption of no inter-operator coordination. We show the 5th, 50th, and 95th percentiles of the UE downlink rates at 32 GHz, assuming a BS density of 100 BSs/km² and a user density of 800 UEs/km². These results also apply to 28 GHz. Note that all pooling scenarios of Fig. 2 have the same association and beamforming schemes, and the only differences are the available bandwidth to each operator and the number of operators sharing a channel, as specified in the next section.

In Fig. 2a, we assume a 32×32 uniform planar array (UPA) at each BS and a 4×4 UPA at each UE. Moreover, we assume that each BS uses six RF chains, such that it can create simultaneously up to six analog beams. We observe in Fig. 2a that most UEs benefit from partial and full pooling compared to the baseline (i.e., exclusive access). In Fig. 2b, we repeat the previous comparison under the assumption of a single omnidirectional antenna at the UE, as a way to study the effect of a less directive beamforming. In this case, partial and full pooling lead to worse performance for the 5th percentile UEs and (for the case of full pooling) for the 50th percentile UEs, due to the increased inter-operator interference. However, the top 5th percentile UE rates that experience the highest SINR values still benefit from spectrum sharing, as

these users are less affected by interference and they can benefit of the wider available bandwidths.

Figure 3 shows the impact of the operating frequency and the BS density on the performance. We consider transmissions at 32 GHz and 73 GHz. We keep the size of the antenna array constant as a function of the frequency, that is, at 73 GHz we consider twice the antenna elements in each dimension with respect to 32 GHz at both BS and UE. We plot the gain of full pooling compared to an exclusive spectrum allocation for a BS density of 50 and 200 BSs/km² (corresponding to a cell radius of 80 and 39 m, respectively). We note that without inter-operator coordination, increasing the BS density of individual operators exacerbates the inter-operator interference and reduces the benefits of spectrum pooling. For example, at 32 GHz when going from 50 BSs/km² to 200 BSs/km², spectrum pooling at the 5th percentile users is reduced from 50 percent to almost 0 percent. This effect is less pronounced at 73 GHz due to the higher directionality of the beams. Figure 3 shows that inter-operator coordination is very effective in very dense deployments (200 BS/km²) and for the weakest UEs (5th percentile UEs). Moreover, full coordination is more critical at 32 GHz than at 73 GHz, due to the fact that beamforming by itself is not sufficient to protect the weakest users from inter-network interference.

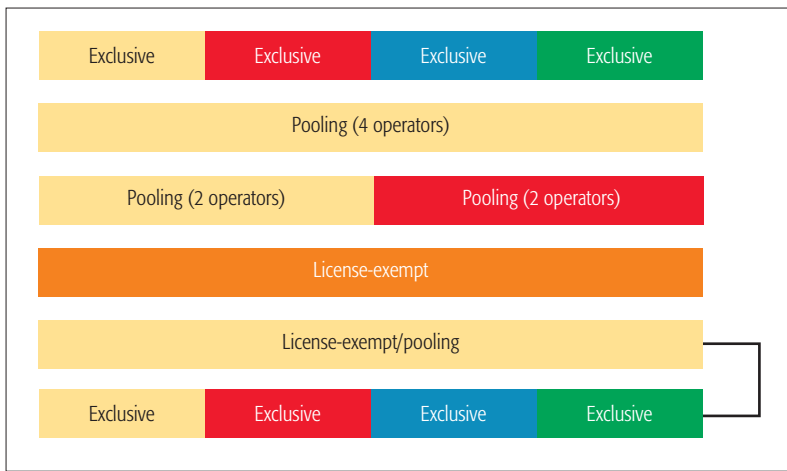


Figure 4. Examples of spectrum sharing regimes.

DISCUSSION

The results presented above clearly indicate that under ideal assumptions, spectrum pooling is beneficial, and it provides improved performance to the 5th, 50th, and 95th percentile users. However, more work is required to assess the impact of real-world effects. Our analysis indicates that beam directionality is a critical enabler, which is consistent with the results in [7]. For example, when using a single omnidirectional antenna at the UE, pooling performance drastically decreases. Moreover, real-world effects (e.g., imperfect channel estimation, pilot contamination, and mobility) may significantly reduce the beamforming gains. The BS density also impacts the performance: in very dense deployments, even under ideal assumptions about beamforming, the weakest users suffer from inter-operator interference. Different from [7], we found that inter-operator coordination is required, especially for the weakest users. Moreover, our results show that inter-operator coordination is more critical at 32 GHz than at 73 GHz. We note that, while in this study we consider a centralized coordination scheme, the impact of more realistic (e.g., distributed) coordination approaches should be further studied. Therefore, another critical enabler of spectrum pooling is a reliable control channel for exchanging coordination information [16].

FUTURE WORKS

In addition to spectrum pooling, we may explore other spectrum access regimes (Fig. 4). From the top to the bottom, the first access scheme refers to the case where a band is licensed to a single operator (e.g., cellular bands at 800 MHz or 2.1 GHz). The second and third access schemes refer to the case where different operators are pooling the same spectrum band. The fourth refers to the case where a band is license-exempt for a given application (e.g., wireless LAN at 5.150–5.250 GHz). The fifth access scheme refers to a hybrid spectrum regime.

Under a license-exemption regime, there is no limitation to the number of operators sharing the spectrum. In other words, for the scenarios considered in this regime, the main difference between spectrum pooling and license-exempt is that in the first case the fixed number of operators sharing the spectrum allows a tighter control of the inter-operator interference.

There are recent technologies that aggregate carriers in both licensed and license-exempt spectrum, to route the different information pipes to the carrier that best matches their requirements; for example, licensed assisted access (LAA) and LTE/WiFi link aggregation. We refer to this as a hybrid spectrum regime. A similar approach, based on a hybrid use of pooling (or license-exempt) at mmWave and exclusive spectrum allocation at traditional cellular frequencies could also be exploited (Fig. 4). For example, carrier aggregation could be used to transmit more critical information (e.g., control and synchronization signals [16]) over the licensed spectrum, while sending less critical information over pooled spectrum at mmWave. A recent contribution in this area can be found in [17], but additional work is needed to compare the different options.

Furthermore, the results discussed in this article have been obtained under ideal assumptions. There is clearly a need for further work to better understand the impact of real-world factors, including imperfect CSI (which critically affects the beamforming accuracy), realistic antennas, backhaul latency, BS synchronization, and distributed coordination.

Finally, the choice of the spectrum authorization type does not only depend on technology factors, which are the focus of this article. Other factors include, for example, the desirability of promoting competition, encouraging investments and innovation, and achieving widespread availability of services across rural and urban areas. A recent contribution on economic aspects of spectrum and RAN sharing in mmWave cellular networks can be found in [18]. Additional work is needed to further study these and other non-technical factors.

CONCLUSIONS

mmWave communications have recently emerged as a solution to the spectrum scarcity in bands traditionally used for cellular communications. However, even at mmWave frequencies, the spectrum is not unlimited, which means it is essential to achieve an efficient use of the spectrum. In this article, we discuss the technical enablers that are required to make spectrum pooling work under realistic assumptions and constraints, including the type of supporting architecture, the type of coordination, the amount and type of information exchange required, along with new functionalities. We also demonstrate that spectrum pooling at mmWave could allow more efficient use of the spectrum than a traditional regime where exclusive spectrum is allocated to individual operators. In particular, we assess the benefit of coordination among the networks of different operators, study the impact of beamforming at both the base stations and the user terminals, and analyze the pooling performance at different carrier frequencies.

REFERENCES

- [1] E. A. Jorswieck *et al.*, "Spectrum Sharing Improves the Network Efficiency for Cellular Operators," *IEEE Commun. Mag.*, vol. 52, no. 3, Mar. 2014, pp. 129–36.
- [2] A. Khan *et al.*, "Network Sharing in the Next Mobile Network: TCO Reduction, Management Flexibility, and Operational Independence," *IEEE Commun. Mag.*, vol. 49, no. 10, Oct. 2011, pp. 134–42.

- [3] W. C. Cheung, T. Quek, and M. Kountouris, "Throughput Optimization, Spectrum Allocation, and Access Control in Two-Tier Femtocell Networks," *IEEE JSAC*, vol. 30, no. 3, Apr. 2012, pp.561–74.
- [4] L. Badia *et al.*, "A Tunable Framework for Performance Evaluation of Spectrum Sharing in LTE Networks," *Proc. IEEE Int'l. Symp. and Wksp. on a World of Wireless, Mobile and Multimedia Networks*, June 2013, pp. 1–3.
- [5] W. Feng *et al.*, "Inter-Network Spatial Sharing with Interference Mitigation Based on IEEE 802.11ad WLAN system," *Proc. IEEE GLOBECOM Wksp.*, Dec. 2014, pp. 752–58.
- [6] G. Li, T. Irnich, and C. Shi, "Coordination Context-Based Spectrum Sharing for 5G Millimeter-Wave Networks," *Proc. Int'l. Conf. Cognitive Radio Oriented Wireless Networks and Communications*, June 2014, pp. 32–38.
- [7] A. K. Gupta, J. G. Andrews, and R. W. Heath, "On the Feasibility of Sharing Spectrum Licenses in mmWave Cellular Systems," arXiv preprint arXiv:1512.01290, Dec. 2015.
- [8] M. Rebato *et al.*, "Resource Sharing in 5G mmWave Cellular Networks," *IEEE INFOCOM*, San Francisco, CA, Apr. 2016.
- [9] 3GPP TR 22.852, 3GPP System Architecture Working Group 1 (SA1) RAN Sharing Enhancements Study Item.
- [10] T. Irnich *et al.*, "Spectrum Sharing Scenarios and Resulting Technical Requirements for 5G Systems," *Proc. IEEE Int'l. Symp. Personal, Indoor and Mobile Radio Commun.*, 8–9 Sept. 2013, pp. 127–32.
- [11] A. M. Akhtar, X. Wang, and L. Hanzo, "Synergistic Spectrum Sharing in 5G HetNets: A Harmonized SDB-Enabled Approach," *IEEE Commun. Mag.*, Jan. 2016, pp. 40–47.
- [12] 3GPP TS 36.420, "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 General Aspects and Principles (Release 8)."
- [13] R. Baldemair *et al.*, "Ultra-Dense Networks in Millimeter-Wave Frequencies," *IEEE Commun. Mag.*, Jan. 2015, pp. 202–08.
- [14] M. Akdeniz, *et al.*, "Millimeter Wave Channel Modeling and Cellular Capacity Evaluation," *IEEE JSAC*, vol. 32, no. 6, June 2014, pp. 1164–79.
- [15] H. Shokri-Ghadikolaei *et al.*, "Spectrum Sharing in mmWave Cellular Networks via Cell Association, Coordination, and Beamforming," to appear, *IEEE JSAC*.
- [16] H. Shokri-Ghadikolaei *et al.*, "Millimeter Wave Cellular Networks: A MAC Layer Perspective," *IEEE Trans. Commun.*, vol. 63, no. 10, Oct. 2015, pp. 3437–58.
- [17] M. Rebato *et al.*, "Hybrid Spectrum Access for mmWave Networks," *Proc. 15th IFIP Annual Mediterranean Ad Hoc Networking Wksp.*, Barcelona, Spain, June 20–21, 2016.
- [18] F. Fund *et al.*, "Spectrum and Infrastructure Sharing in Millimeter Wave Cellular Networks: An Economic Perspective," 2016; <http://arxiv.org/abs/1605.04602>.

BIOGRAPHIES

FEDERICO BOCCARDI (federico.boccardi@ieee.org) is a principal technology advisor at Ofcom (the U.K. communication regulator), where he is leading the technical work to release new spectrum for 5G and WiFi. He received his M.Sc. and Ph.D. degrees in telecommunication engineering from the University of Padova, Italy, in 2002 and 2007, respectively, and his postgraduate diploma in strategy and innovation from the Oxford Saïd Business School in 2014. Before joining Ofcom, he was with Bell Labs (Alcatel-Lucent) from 2006 to 2013 and with Vodafone R&D in 2014. During his career he held leadership positions in the 3GPP standardization activity for LTE and LTE-Advanced, and received several research awards including the 2014 IEEE GLOBECOM Best Paper Award and the 2016 IEEE Communication Society's Fred W. Ellersick Prize. He is an Associate Editor for *IEEE Transactions on Cognitive Communications and Networking*. His current interests fall at the intersection between technology and strategy.

HOSSEIN SHOKRI-GHADIKOLAEI (hshokri@kth.se) is a Ph.D. student at KTH Royal Institute of Technology, Stockholm, Sweden. He received his B.Sc. and M.Sc. degrees in communication systems from Iran University of Science and Technology and Sharif University of Technology, Tehran, Iran, in 2009 and 2011, respectively. He is a member of Working Group 1900.1 in the IEEE Dynamic Spectrum Access Networks Standards Committee (DySPAN-SC). His research interests include wireless communications, with applications in cellular, ad hoc, and cognitive networks.

GABOR FODOR (gabor.fodor@ericsson.com) received his Ph.D. degree in teletraffic theory from the Budapest University of Technology and Economics in 1998. He has been with Ericsson Research, Kista, Sweden, since 1998. He has been a visiting researcher with the Automatic Control Laboratory, Royal Institute of Technology, Stockholm, Sweden, since 2013. He is currently a

master researcher at Ericsson, specializing in modeling, performance analysis, and protocol development for wireless access networks.

ELZA ERKIP (elza@poly.edu) received her B.S. degree in electrical and electronics engineering from Middle East Technical University, Turkey, and her M.S. and Ph.D. degrees in electrical engineering from Stanford University, California. Currently, she is a professor of electrical and computer engineering with New York University Tandon School of Engineering, Brooklyn, New York. Her research interests are in information theory, communication theory, and wireless communications. She is a member of the Science Academy Society of Turkey, and is among the 2014 and 2015 Thomson Reuters Highly Cited Researchers. She was the recipient of the NSF CAREER Award in 2001, the IEEE Communications Society Stephen O. Rice Paper Prize in 2004, the IEEE ICC Communication Theory Symposium Best Paper Award in 2007, and the IEEE Communications Society Award for Advances in Communication in 2013. She coauthored a paper that received the IEEE International Symposium on Information Theory Student Paper Award in 2007. She has been a member of the Board of Governors of the IEEE Information Theory Society since 2012, where she is currently the Second Vice President. She was a Distinguished Lecturer of the IEEE Information Theory Society from 2013 to 2014. She has had many editorial and conference organization responsibilities. Most recently, she was a Guest Editor of the *IEEE Journal on Selected Areas in Communications* in 2015, the General Chair of the IEEE International Symposium on Information Theory in 2013, and an Associate Editor of *IEEE Transactions on Information Theory* from 2009 to 2011. She will be a Technical Chair of the IEEE Wireless Communications and Networking Conference in 2017.

CARLO FISCHIONE (carlofi@kth.se) is a tenured associate professor at KTH Royal Institute of Technology. He received his Ph.D. degree in electrical and information engineering and Laurea degree in electronic engineering from the University of L'Aquila, Italy. He has held research positions at Massachusetts Institute of Technology (2015), Harvard University (2015), and the University of California at Berkeley (2004 and 2007–2008). His research interests include optimization with applications to wireless networks and cyber-physical systems.

MARIOS KOUNTOURIS (marios.kountouris@huawei.com) received his Diploma in electrical and computer engineering from the National Technical University of Athens, Greece, in 2002, and his M.S. and Ph.D. from Télécom ParisTech, France in 2004 and 2008, respectively. From February 2008 to May 2009, he was with the University of Texas at Austin as a research associate. From June 2009 to July 2016, he was with SUPELEC, France, as an assistant and associate professor. Since January 2015, he has been a principal researcher at Huawei Technologies, France. He has received several awards including the 2013 IEEE ComSoc Outstanding Young Researcher Award and the 2012 IEEE SPS Signal Processing Magazine Award.

PETAR POPOVSKI (petarp@es.aau.dk) received his Dipl.-Ing. (1997) and Magister Ing. (2000) in communication engineering from Sts. Cyril and Methodius University, Skopje, Macedonia, and his Ph.D. from Aalborg University, Denmark (2004). He is currently a professor at Aalborg University. He is an Editor for *IEEE Transactions on Communications* and has served in the past as an Editor for *IEEE Communications Letters*, the *IEEE JSAC Cognitive Radio Series*, and *IEEE Transactions on Wireless Communications*. He is a Steering Committee member for the *IEEE Internet of Things Journal*. His research interests are in communication theory, wireless communications, and networking.

MICHELE ZORZI [F'07] (zorzi@dei.unipd.it) received his Laurea and Ph.D. degrees in electrical engineering from the University of Padova in 1990 and 1994, respectively. During academic year 1992–1993 he was on leave at the University of California, San Diego (UCSD). After being affiliated with the Dipartimento di Elettronica e Informazione, Politecnico di Milano, Italy, the Center for Wireless Communications at UCSD, and the University of Ferrara, in November 2003 he joined the faculty of the Information Engineering Department of the University of Padova, where he is a professor. His present research interests include performance evaluation in mobile communications systems, mobile radio networks, ad hoc and sensor networks, energy constrained communications protocols, and underwater communications and networking. He was Editor-in-Chief of *IEEE Wireless Communications* from 2003 to 2005, Editor-in-Chief of *IEEE Transactions on Communications* from 2008 to 2011, a Guest Editor for several Special Issues in *IEEE Personal Communications*, *IEEE Wireless Communications*, *IEEE Network*, and *IEEE JSAC*, and the founding Editor-in-Chief of *IEEE Transactions on Cognitive Communications and Networking*. He served as a Member-at-Large of the Board of Governors of the IEEE Communications Society from 2009 to 2011, and as its Director of Education in 2014 and 2015.

Initial Access in 5G mmWave Cellular Networks

Marco Giordani, Marco Mezzavilla, and Michele Zorzi

The massive amounts of bandwidth available at millimeter-wave frequencies (above 10 GHz) have the potential to greatly increase the capacity of fifth generation cellular wireless systems. However, to overcome the high isotropic propagation loss experienced at these frequencies, highly directional antennas will be required both at the base station and at the mobile terminal, to achieve sufficient link budget in wide area networks.

ABSTRACT

The massive amounts of bandwidth available at millimeter-wave frequencies (above 10 GHz) have the potential to greatly increase the capacity of fifth generation cellular wireless systems. However, to overcome the high isotropic propagation loss experienced at these frequencies, highly directional antennas will be required at both the base station and the mobile terminal to achieve sufficient link budget in wide area networks. This reliance on directionality has important implications for control layer procedures. In particular, initial access can be significantly delayed due to the need for the base station and the user to find the proper alignment for directional transmission and reception. This article provides a survey of several recently proposed techniques for this purpose. A coverage and delay analysis is performed to compare various techniques including exhaustive and iterative search, and context-information-based algorithms. We show that the best strategy depends on the target SNR regime, and provide guidelines to characterize the optimal choice as a function of the system parameters.

INTRODUCTION

The fifth generation (5G) of cellular systems is positioned to address the user demands and business contexts of 2020 and beyond.

The subscribers' growing demand for a better mobile broadband experience calls for significant performance improvements, such as:¹

- Much greater throughput (1 Gb/s or higher) to support ultra-high definition video and virtual reality applications;
- Much lower latency (less than 1 ms) to support real-time mobile control and device-to-device (D2D) applications
- Ultra-high reliability and much higher connectivity to provide seamless service everywhere

In order to deal with these requirements, some key aspects have been identified to make this future network a reality. Since current Long Term Evolution (LTE) spectrum under 6 GHz is fragmented and crowded, there has been significant interest in the millimeter-wave (mmWave) bands,² where the vast amount of largely unused spectrum available can support the higher data rates required in future mobile broadband access networks. Moreover, small mmWave wave-

lengths make it practical to build very large antenna arrays (e.g., with 32 or more elements) to provide further gains from spatial isolation and multiplexing. However, the increased carrier frequency makes the propagation conditions at mmWave more demanding. For example, blockage becomes an important issue, as mmWave signals do not penetrate most solid materials (e.g., buildings made of brick) and are subject to very high signal attenuation [1]. Another pillar of 5G will be the use of many more base stations, deployed according to a heterogeneous network (HetNet) paradigm, combining macro sites with smaller base stations and using a wide range of radio technologies. These will include LTE, Wi-Fi, and any future 5G technologies, integrated flexibly in any combination.

In this context, the definition of new control layer procedures is critical, in particular initial access (IA), which allows a mobile user equipment (UE) to establish a physical link connection with a base station (BS), a necessary step to access the network. In current LTE systems, IA is performed on omnidirectional channels, whereas beamforming (BF) or other directional transmissions can only be performed after a physical link is established. On the other hand, in order to overcome the increased isotropic path loss experienced at higher frequencies, in 5G mmWave cellular systems the IA procedure must provide a mechanism by which the BS and the UE can determine suitable initial directions of transmission. However, directionality can significantly delay the cell search and access procedures, which is a particularly sensitive issue in 5G networks, and thus motivated us to identify and study some performance trade-offs in terms of *delay, coverage, and overhead*.

This article reviews recent directional IA techniques for mmWave cellular systems. As an extension of our previous contribution in [2], here we investigate various search schemes, including exhaustive search, an iterative scheme that successively narrows the search beamwidth, and context information (CI)-based algorithms, where users are informed about the geolocations of surrounding mmWave BSs through an LTE link. We compare the performance of these approaches in terms of both misdetection probability and discovery time, under overhead constraints and different channel conditions. Our results show that the optimal strategy depends

¹ Nokia White Paper "Looking Ahead to 5G," May 2014; <http://networks.nokia.com/file/28771/5g-white-paper>

² Although strictly speaking mmWave bands include frequencies between 30 and 300 GHz, THE industry has loosely defined it to include any frequency above 10 GHz.

on the target signal-to-noise ratio (SNR) regime, and provide some guidance about the best scheme to use in each scenario.

4G-LTE: INITIAL ACCESS LIMITATIONS

In all mobile communication systems, a terminal transitioning from IDLE to CONNECTED mode must perform the following steps: cell search (CS), extraction of system information, and random access (RA). In this section, we discuss the main factors that make 4G-LTE procedures unsuitable for use in a 5G mmWave context.

Discovery Range Mismatch: In LTE systems, acquiring time-frequency synchronization during CS is facilitated, as signals are transmitted omnidirectionally in the downlink, and BF is used only after a physical link has been established. In mmWave bands, it may instead be essential to exploit the BF gains even during the CS phase, since omnidirectional signaling may generate a mismatch between the relatively short range at which a cell can be detected (C-plane range), and the much longer range at which a user could directionally send and receive data (U-plane range) [3, 4].

Multi-Connectivity: To ensure sufficient coverage, mmWave networks will be much denser. Each user is expected to simultaneously detect multiple potential serving stations, including at least a macro BS operating in the LTE spectrum. Consequently, the IA procedures have to be redesigned in order to capture this fundamental new feature. We refer to the previous section and [5] for further details.

Deafness and Blockage: In mmWave cellular networks, IA messages may not be received due to deafness or blockage phenomena. Deafness refers to a situation where the transmit-receive beams do not point to each other, whereas blockage causes a failed message delivery due to a channel drop, which may be related to obstacles, hand rotations, and other mmWave-sensitive events. Neither increasing the transmission power nor waiting for a random backoff time (as done in LTE) is a suitable approach in mmWave networks. Hence, to discriminate among different reasons for access failure, new adaptive techniques have to be introduced.

Dynamics-Aware Access: Due to denser topologies, association schemes based on reference signal received power (RSRP) would be highly inefficient in mmWave cellular networks, an issue already encountered in HetNets [6]. However, the challenge with higher frequencies is the need to also account for dynamics such as directionality and intermittency.

Hence, there is an urge to extend current LTE procedures and adapt them to the upcoming mmWave related challenges in order to overcome such limitations, or come up with new algorithms and new methods. A natural (and practical) solution is to use BF even in the first stages of the initial access procedure, keeping in mind that a fully directional data plane requires a directional IA procedure in the new frequency band. On the other hand, when considering an analog multi-antenna architecture, directionality means that only one direction can be considered at a time, thereby

losing the broadcast property of the wireless medium, with important implications for protocol design and delay performance that must be carefully taken into consideration.

The technical issues described in this section call for new initial access procedures and for a detailed assessment of their performance in realistic 5G mmWave scenarios. Such a comparison is the main goal of this article.

RELATED WORK

Papers on IA in mmWave 5G cellular systems are very recent. Most of the literature refers to challenges that were analyzed in the past at lower frequencies in ad hoc wireless network scenarios or, more recently, in the 60 GHz IEEE 802.11ad WLAN and WPAN scenarios. However, most of the proposed solutions are unsuitable for future 5G mmWave network requirements, since they present many limitations (e.g., they are appropriate for short-range, static, and indoor scenarios, which do not match well the requirements of 5G systems). Therefore, new specifically designed solutions for cellular networks need to be found.

IA in mmWave cellular networks was considered in [7], which proposes an exhaustive method to sequentially scan the 360° angular space. In [3], a directional cell discovery procedure was proposed, where BSs periodically transmit synchronization signals, potentially in time-varying random directions, to scan the angular space. IA design options were also compared in [8], considering different scanning and signaling procedures, to evaluate access delay and system overhead; the analysis demonstrated significant benefits of low-resolution fully digital architectures in comparison to single-stream analog beamforming. Additionally, in order to alleviate the exhaustive search delay, a two-phase hierarchical procedure based on a faster user discovery technique was proposed in [9].

CI-based procedures aim at exploiting knowledge about user and/or BS positions, which are provided by a separate control plane, in order to improve the cell discovery mechanism and minimize the delay [10]. In the scheme proposed in [11], booster cells (operating at mmWave) are deployed under the coverage of an anchor cell (operating at LTE frequencies). The anchor BS gets control over IA, informing the booster BS about user locations in order to enable mmWave cells to directly steer toward the user position. Furthermore, an evolution of [10] was presented in [12], showing how to capture the effects of position inaccuracy and obstacles. Finally, in [13], the authors studied how the performance of analog BF degrades in the presence of angular errors in the available CI during the initial cell search.

In [2], we presented a comparison between the exhaustive and iterative techniques. In this work, we expand the analysis to a CI-based algorithm and describe a proposed enhancement thereof. Our goal is to compare multiple IA procedures under an overhead constraint and to derive the best trade-offs, in terms of both misdetection probability and discovery delay, when considering a realistic dense, urban, multi-path scenario.

Due to denser topologies, association schemes based on reference signal received power would be highly inefficient in mmWave cellular networks, an issue already encountered in HetNets. However, the challenge with higher frequencies is the need to also account for dynamics such as directionality and intermittency.

At this stage, the UE does not have any resource or channel available to inform the network about its desire to connect to it; RA provides a means to set up this connection. Both the UE and the BS know, from the previous CS phase, the best directions through which they should steer their beams, and therefore they will exchange the following RA messages in one step.

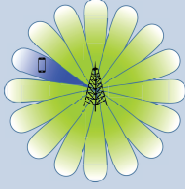
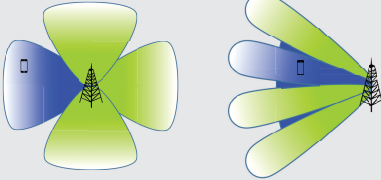
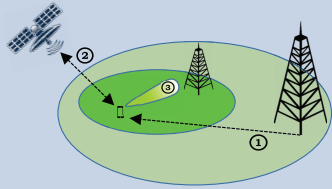
Exhaustive search	Iterative search	Pure CI-based search
		
<ul style="list-style-type: none"> • Good coverage • Suitable for edge users • High discovery delay 	<ul style="list-style-type: none"> • Low discovery delay • Not suitable for edge users • Bad coverage 	<ul style="list-style-type: none"> • Low discovery delay • Cost of getting GPS coordinated • Only line-of-sight (LOS) scenarios

Table 1. Summary of the three IA cell search algorithms compared in this work.

INITIAL ACCESS IN 5G MMWAVE NETWORKS

For mmWave, we list in the following the initial access steps that enable both the UE and the BS to determine their initial BF directions, in addition to detecting the presence of the BS and of the access request from the UE.

CELL SEARCH FOR INITIAL ACCESS

As summarized in Table 1, we evaluate three IA-CS procedures. In this study we focus on *analog* BF techniques (considered to be more energy-efficient than their digital counterparts [4]), where each transceiver can look at only one direction per slot.³ Hybrid and fully digital BF architectures are left for future work.

Exhaustive search: A brute-force sequential beam searching technique [7]. Both users and base stations have a predefined codebook of N directions (each identified by a BF vector) that cover the whole angular space and are used sequentially to transmit/receive.⁴

Iterative search: A two-stage scanning of the angular space [9]. In the *first phase*, the BS transmits pilots over wider sectors, while in the *second phase* it refines its search within the best such sector by steering narrower beams.

CI-based search: This algorithm is articulated into three main stages [10].⁵

1. The macro BS (operating at LTE frequencies) spreads the GPS coordinates of all the mmWave stations within its range omnidirectionally.
2. Each UE gets its own GPS coordinates and orientation (this will require a certain energy cost).
3. According to the information obtained in steps 1 and 2, each UE geometrically selects the closest BS to connect to and steers a beam toward it. Meanwhile, each mmWave BS performs an exhaustive search to detect the best transmit-receive direction.

We refer to [2, 7, 9] for further procedural details on exhaustive and iterative techniques, and to [10, 13] for CI-based algorithms.

In our study, we focus on the cell search phase described above, which determines whether or not a UE is able to be detected by the BS, and the operation of which dominates the overall delay performance.

RANDOM ACCESS

At this stage, the UE does not have any resource or channel available to inform the network about its desire to connect to it; RA provides a means

to set up this connection. Both the UE and the BS know, from the previous CS phase, the best directions through which they should steer their beams, and therefore they will exchange the following RA messages in one step. The RA is composed of four stages:

1. *RA preamble transmission*, when the UE randomly selects one contention-based signature and sends it to the BS
2. *RA response*, sent to the UE from the BS and containing an initial timing and power correction as well as some cell-specific medium access control (MAC) layer identifier to uniquely identify the UE in the cell
3. *Connection request message*, sent by the UE requesting initial access that includes, among other data, some authentication and identification information
4. *Contention resolution phase*, if needed

All subsequent communications can finally occur on scheduled channels.

PERFORMANCE EVALUATION

In the simulations, we assume a static deployment where no users are moving, so no handover management or UE motion tracking is required. The parameters are based on realistic system design considerations and are summarized in Table 2. To conduct our performance analysis, we assume a slot structure similar to the one described in [8]. The primary synchronization signal (PSS) is transmitted periodically in the downlink direction, once every T_{per} s, for a duration of T_{sig} s.

The channel model is based on recent real-world measurements at 28 GHz in New York City to provide a realistic assessment of mmWave micro and picocellular networks in a dense urban deployment. Statistical models are derived for key channel parameters including path loss, number of spatial clusters, angular dispersion, and outage. Further details of this model and its parameters can be found in [14].

We adopt analog beamforming, implemented through a uniform planar array (UPA), which allows steering toward one direction at a time. The total system bandwidth is taken to be 1 GHz, the transmission powers of BS and UE are set to 30 dBm and 23 dBm, respectively, and a noise figure of 5 dB is assumed.

In order to detect a PSS signal, the SNR received at the UE has to lie above a certain threshold τ , taken to be -5 dB in our simula-

³ A slot (either downlink or uplink) is defined as the interval of time (of length T_{sig}) in which a transceiver transmits a synchronization signal through a certain portion of the angular space.

⁴ In this work, we consider a fixed codebook of steering directions. A more complete analysis on how to optimally design the codebook to maximize the performance of the procedure is beyond the scope of this article and is left as future work.

⁵ We note that the procedure we consider in this article is different from [10] in that the CI is available at the UE and concerns the location of the BS. As also done in [13], this is a more natural way of using CI and assigns the burden of beam-scanning to the BS (which would have to do it anyway in the presence of multiple users) rather than to the UE (which can in this case save energy). A more detailed discussion and comparison between the two paradigms is beyond the scope of this article.

tions. Decreasing τ would allow finding more users at the cost of designing more complex (and expensive) receiving schemes, able to detect the intended signal in noisier channels.

Each PSS signal has a minimum duration $T_{\text{sig}} = 10 \mu\text{s}$, which is deemed to be sufficient to allow proper channel estimation at the receiver.

Simulations are conducted increasing the distance of the UE from the BS. At each iteration, the user is deployed within an annulus having outer radius R_1 and inner radius $R_2 < R_1$ according to a uniform distribution. In order to obtain statistically significant results through a Monte Carlo estimation, each simulation is independently repeated 10^6 times.

In this study, we evaluate the performance in terms of:

- *Discovery delay*, which is the time required by the BS and the UE to complete their angular scans (sending and receiving PSS control messages, respectively), to determine and select the best directions through which to steer their beams
- *Misdetection probability (PMD)*, which is the probability that a UE within the cell is not detected by the BS in the cell search phase, perceiving an SNR below threshold

We compare two sequential-based IA schemes: the exhaustive search and the iterative technique. The main conclusion of this study is that exhaustive search is likely to be the best IA configuration when performing CS, especially if we want to provide good coverage with high probability at relatively large distances (e.g., as in the case of edge users in large cells), whereas iterative search may be preferred otherwise (e.g., in case of smaller cells). However, the best technique generally depends on the target SNR and the considered scenario.

We compare the CI-based initial access technique with the two sequential schemes. The key finding is that pure CI algorithms may not be suitable for urban scenarios, where links are often non-line-of-sight (NLOS). Nonetheless, we show that more sophisticated directional procedures have the potential to reduce the discovery delay and grant good coverage.

SEQUENTIAL APPROACH

In the exhaustive technique, the BS is equipped with 64 (8×8) antennas and can steer beams in $N = 16$ directions. The UE has a set of combining vectors that also cover the whole angular space, and receives PSSs through 4 wide beams (using only 4 antennas) or through 8 narrower beams (using all 16 antennas). In the iterative first phase, the BS sends PSS messages in 4 macro directions through 4 wide beams, using 4 antennas, while in the second phase it sends the refining PSSs through 4 narrow beams, using 64 antennas.

Discovery delay: We consider a minimum signal duration $T_{\text{sig}} = 10 \mu\text{s}$ and a target *overhead* of $\phi_{\text{ov}} = 5$ percent, which results in a time between two consecutive signal transmissions of at least $T_{\text{per}} = T_{\text{sig}}/\phi_{\text{ov}} = 200 \mu\text{s}$. Given that cell search requires N_s slots (both DL and UL transmitted signals, one per T_{per}), the discovery delay can be computed as $N_s T_{\text{per}} = N_s T_{\text{sig}}/\phi_{\text{ov}}$.

For example, in the *exhaustive* 64×16 case,

Parameter	Value	Description
W_{tot}	1 GHz	Total system bandwidth
DL P_{TX}	30 dBm	Downlink transmission power
UL P_{TX}	23 dBm	Uplink transmission power
NF	5 dB	Noise figure
f_c	28 GHz	Carrier frequency
min. τ	-5 dB	Minimum SNR threshold
BS antennas	8×8	BS UPA MIMO array size
UE antennas	4×4 or 2×2	UE UPA MIMO array size
min. T_{sig}	10 ms	Minimum signal duration
ϕ_{ov}	5%	Overhead
T_{per}	$T_{\text{sig}}/\phi_{\text{ov}}$	Period between transmissions

Table 2. Simulation parameters.

the BS sends 8 DL pilots in each direction, while the UE sequentially scans its 8 receiving beams, and this is repeated for all the BS's 16 angular directions; finally, in each of the BS's 16 directions, the UE sends a UL signal through its best direction, to feed back to the BS the index of the optimal beam it should use. Therefore, the procedure requires $N_s = (16 \cdot 8) + 16 = 144$ signals to be sent. By analogy, and referring to our previous work [2], the reader can similarly determine how N_s was obtained for all the remaining IA schemes. The final values of N_s are reported in Table 3.

The time required for RA (the last phase of IA) can be neglected since here messages are sent through already set steering directions, without the need to scan the angular space again. Therefore, we argue that the CS latency is the dominant factor when determining the overall IA discovery delay.

Generally, the iterative approach requires fewer slots (and consequently presents lower discovery delay if all slots have the same length) with respect to an exhaustive technique, because BSs do not use narrow beams to scan the whole 360° angular space, but just need to refine a macro sector.

Misdetection probability: Since BSs in iterative search transmit over wider beams in the first phase, which results in a reduced BF gain, the misdetection probability is higher compared to exhaustive search, where BSs always use narrower beams that provide higher gains. Moreover, when the UE, in reception mode, exploits just four antennas (receiving PSSs through four beams), the BF gain is reduced as well, also resulting in a higher misdetection probability.

Figure 1 shows that in small cells the misdetection probability is small, and an iterative procedure may be preferred because of its lower delay. Conversely, at mid-range distances, iterative methods are not sufficiently reliable, and exhaustive techniques would be preferred, whereas for farther users almost all algorithms perform poorly due to the high probability of channel outage [14]. The ability of current tech-

The time required for RA (the last phase of IA) can be neglected since here messages are sent through already-set steering directions, without the need to scan again the angular space. Therefore, we argue that the CS latency is the dominant factor when determining the overall IA discovery delay.

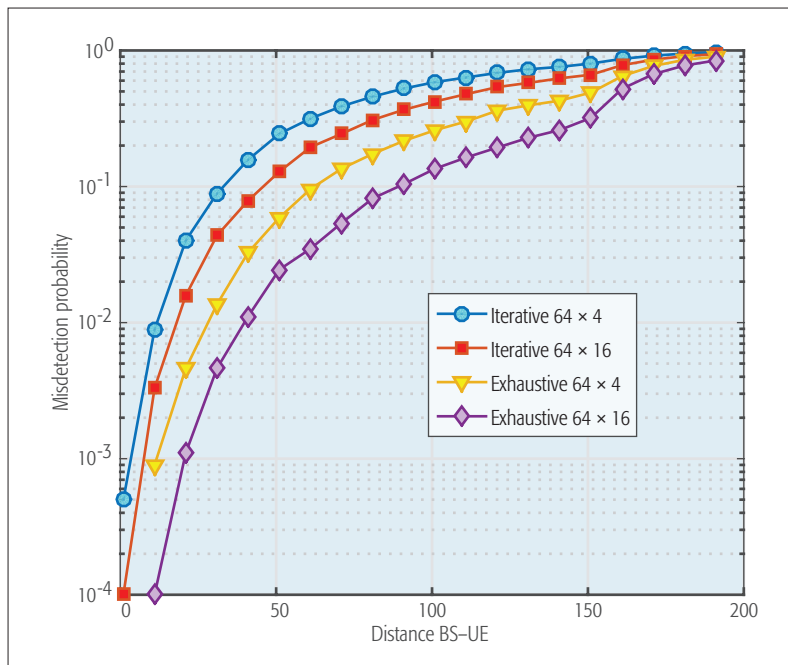


Figure 1. PMD for exhaustive and iterative techniques, when UE receives in 4 or 8 directions, vs. the BS-UE distance. SNR threshold $\tau = -5$ dB. $T_{\text{sig}} = 10\mu\text{s}$ and $T_{\text{per}} = 200\mu\text{s}$.

nologies to adequately support IA only at relatively small distances motivates the search for better IA methods.

Trade-off between delay and PMD: In order to keep the misdetection probability below a certain threshold (taken here to be 0.01), we can increase the signal duration: if T_{sig} is increased, the BS transmits its PSS for a longer time in the same sector, so UEs belonging to that sector can accumulate more energy, resulting in higher SNR and correspondingly reduced PMD. Figure 2 shows the PMD as a function of T_{sig} for the different schemes, considering a UE at a distance of 95 and 35 m from the BS, respectively. From these results we can derive the minimum signal length to meet a certain PMD requirement by reading the abscissa of the intersection of each curve with the horizontal line.

The *discovery delay* in Table 3a captures both the required number of slots and the PMD specifications. According to the specific T_{sig} values obtained from Fig. 2, the corresponding $T_{\text{per}} = T_{\text{sig}}/\phi_{\text{ov}}$ is selected in order to have a constant overhead $\phi_{\text{ov}} = 5$ percent.

For example, for the exhaustive 64×16 case in Fig. 2a, we see that the corresponding curve intercepts the PMD threshold when the signal duration is around $125\mu\text{s}$, while when considering the iterative 64×16 case, a signal duration of around $1580\mu\text{s}$ must be used to meet the misdetection requirements. Moreover, considering closer users in Fig. 2b, we show that exhaustive schemes reach the threshold even when adopting the minimum allowed signal duration $T_{\text{sig}} = 10\mu\text{s}$. Note that in these examples the iterative search, which was the best approach when keeping T_{sig} constant to $10\mu\text{s}$ for all IA schemes, is outperformed by the exhaustive approach if a PMD target is imposed, as its gain in terms of needing fewer slots (roughly a factor of 3 in this

example) is outweighed by their longer duration (about one order of magnitude).⁶

In general, iterative techniques try to compensate for the lower BF gain in the first phase by collecting enough energy to obtain a sufficiently high SNR, and for this reason require a longer signal duration, despite the lower number of slots required. Conversely, exhaustive searches can operate with shorter slots but need more of them. From the results of Table 3a, if the goal is to grant good coverage levels to users 95 m from the BS, an iterative approach is not preferred, as it requires a much longer discovery delay with respect to that of the exhaustive scheme. On the other hand, if we want to guarantee good PMD just for *closer users* (e.g., in smaller cells), even iterative techniques lead to a sufficiently low discovery delay, despite the longer slots required.

CI-BASED INITIAL ACCESS

We first consider the pure CI-based technique reported in Table 1, which requires $N_s = 32$ slots if the BS is equipped with 64 antennas, which provide 16 beams.⁷ However, if the direct path does not correspond to good channel conditions (e.g., as in NLOS or multi-path scenarios), the beam chosen by the UE may actually be sub-optimal. We then propose a more sophisticated scheme, where a successive *beam refinement* is performed: first, the UE points a beam over the direct path inferred via CI. Then the UE forms additional beams in adjacent directions, looking for a stronger path.

This requires a total of $N_s = 64$ slots if three beams are used (the one corresponding to the direct path plus one on each side).⁸ In both algorithms, we assume that CI is not affected by any GPS error and that the time required to collect them is negligible (e.g., the UE already has this information available for some other purposes).

In Fig. 3, we determine the minimum signal duration to meet the usual requirements of PMD (<0.01) for users at 95 and 35 m from the BS, while in Table 3b we compute the discovery delay $N_s T_{\text{sig}}/\phi_{\text{ov}}$, as in the analysis of an earlier section.

At 95 m, the discovery delay of the pure CI-based technique is higher than that of the exhaustive approach. The reason is that, as mentioned earlier, in a very dense urban environment the channel propagation is affected by many factors. As a consequence, the direct path obtained through CI-based GPS coordinates may be a poorer choice with respect to what will be selected by the exhaustive procedure in which, instead, the best beam is surely found through a complete scan. To meet the PMD requirements, in this case the CI-based algorithm may require very long signals to collect enough energy, with respect to the exhaustive search, thereby greatly increasing the discovery delay. On the other hand, the enhanced CI-based technique, where the refinement is performed, has the potential to drastically reduce the PMD, which almost overlaps with the results obtained through exhaustive search. According to Table 3b, the discovery delay is almost 50 percent lower than in the exhaustive case.

When considering instead users in good propagation conditions (i.e., at 35 m from the BS as in Fig. 3b), the LOS probability is about 60 percent

⁶ We remark that these computations can also be applied to evaluate the energy consumption performance, although this is not addressed explicitly in this study.

⁷ The BS transmits the PSS once per beam (16 downlink slots), and listens for the UE's response once per beam (16 uplink slots), whereas the UE beamforms in a single direction as provided by the CI, hence $N_s = 32$.

⁸ Unlike in the pure CI-based technique, for each beam formed by the BS the UE has to scan through three directions for downlink PSS transmissions, and replies to the BS using the best beam. This requires 3 downlink slots and 1 uplink slot per beam, for a total of $N_s = (16 \cdot 3) + 16 = 64$. (We found that using additional refining beams increases the IA delay without significantly improving the misdetection probability.)

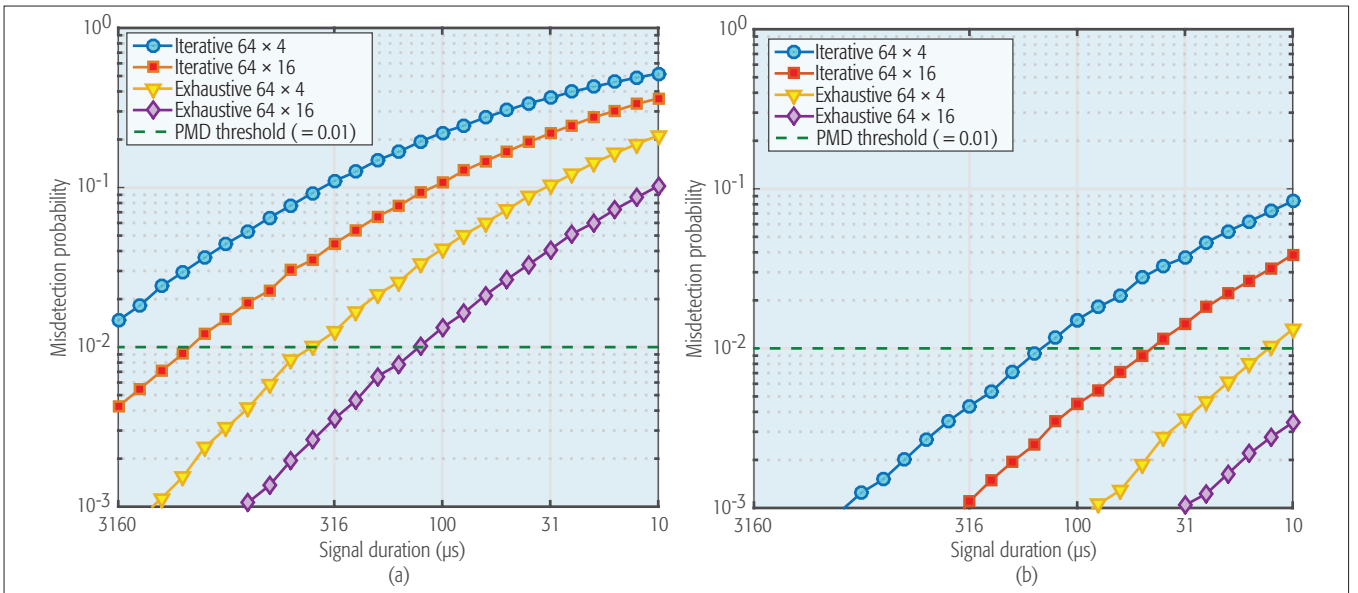


Figure 2. Trade-off between delay and PMD: a) PMD for exhaustive and iterative searches vs. signal duration T_{sig} . BS-UE distance $d = 95$ m; b) PMD for exhaustive and iterative searches vs. signal duration T_{sig} . BS-UE distance $d = 35$ m.

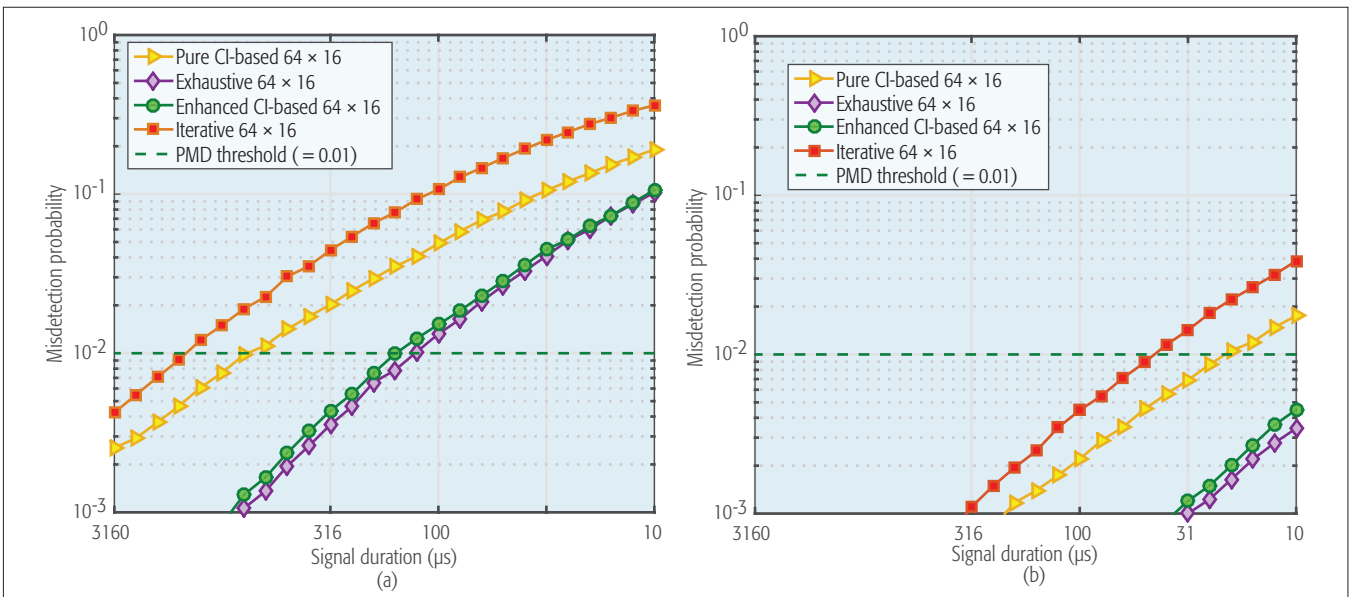


Figure 3. PMD analysis for CI-based initial access: a) PMD for exhaustive, iterative, and CI-based searches vs. signal duration T_{sig} . BS-UE distance $d = 95$ m; b) PMD for exhaustive, iterative, and CI-based searches vs. signal duration T_{sig} . BS-UE distance $d = 35$ m.

(according to [14]), resulting in sufficiently good performance of the pure-CI procedure in terms of discovery delay (about 9.6 ms) without any further improvement (which would in turn require sending many more control messages). Therefore, the use of enhanced CI is justified only to increase the coverage of edge users in large cells.

INITIAL ACCESS FUTURE CHALLENGES

In this section, we present some important challenges for IA in mmWave bands.

BEAM TRACKING

After one of the IA procedures described earlier is performed, the best BS-UE beam pair is determined. However, the movement of obstacles and

reflectors, or even changes in the orientation of a handset relative to a body or a hand, can cause the channel to rapidly appear or disappear [15]. Hence, the selected beam could rapidly change, thus requiring the UE to constantly monitor each potential directional link. *Beam tracking* can therefore introduce some latency, which lowers the rate at which the network can adapt, and can be a major obstacle in providing robust service in the face of variable link quality.

In [5], we proposed a novel multi-cell measurement reporting system to periodically monitor and update the beam direction at both the UE and the BS. A study of the effects of the channel variability and the UE motion over the beam tracking mechanism is left for future work.

Periodical signaling for beam tracking, besides increasing the system complexity, results in additional resource and energy consumption. Additionally, GPS coordinates and orientation need to be acquired over a certain amount of time when CI-based initial access schemes are performed and, consequently, energy needs to be consumed.

(a) Sequential approach. The minimum signal duration T_{sig} is determined according to Fig. 2 for each IA scheme.							
Procedure	Antennas at the BS	Antennas at the UE	N_s	User at 95 meters		User at 35 meters	
				Min. T_{sig}	Discovery delay: $N_s T_{sig}/\phi_{ov}$	Min. T_{sig}	Discovery delay: $N_s T_{sig}/\phi_{ov}$
Exhaustive 64×4	64	4	80	400 μ s	640 ms	13 μ s	20.8 ms
Exhaustive 64×16	64	16	144	125 μ s	360 ms	10 μ s*	28.8 ms
Iterative 64×4	4 in 1st phase 64 in 2nd phase	4	28	> 3160 μ s	> 1760 ms	160 μ s	89.6 ms
Iterative 64×16	4 in 1st phase 64 in 2nd phase	16	44	1580 μ s	1390 ms	50 μ s	44 ms
* Minimum allowed signal duration, which is deemed to be sufficient to allow proper channel estimation at the receiver.							
(b) CI-based approach. The minimum signal duration T_{sig} is determined according to Fig. 3 for each CI-based IA scheme.							
Procedure	Antennas at the BS	Antennas at the UE	N_s	User at 95 meters		User at 35 meters	
				Min. T_{sig}	Discovery delay: $N_s T_{sig}/\phi_{ov}$	Min. T_{sig}	Discovery delay: $N_s T_{sig}/\phi_{ov}$
Pure-CI 64×16	64	16	32	630 μ s	403 ms	15 μ s	9.6 ms
Enhanced-CI 64×16	64	16	64	150 μ s	192 ms	10 μ s*	12.8 ms
* Minimum allowed signal duration, which is deemed to be sufficient to allow proper channel estimation at the receiver.							

Table 3. Discovery delay that guarantees PMD < 0.01 for users at 95 and 35 m from the BS. The table also includes the number of slots (N_s) required to implement each IA technique. The transceiver can steer through 16, 8, or 4 directions if it is equipped with 64, 16 or 4 antenna elements, respectively [2].

MULTI-CONNECTIVITY

One of the challenges in designing mmWave cellular networks is robustness. A likely key feature to meet this requirement is multi-connectivity, where both 5G cells operating at mmWave (offering much higher rates) and traditional 4G cells below 6 GHz (providing much more robust operation) are employed [1].

The use of both LTE and mmWave control planes is a key functionality for the IA technique as well. In [5, 16], we proposed an approach where the UE is connected to a conventional 4G cell able to perform association and handoff decisions among mmWave cells. Unlike in traditional LTE, the proposed IA system is based on the UE periodically sending uplink (UL) sounding pilots. We argue that this has several key benefits:

- The use of UL signals eliminates the need for the UE to send measurement reports back to the network and thereby removes a point of failure in the control signaling path.
- If digital BF or BF with multiple analog streams is available at the mmWave cell, the directional scan time can be dramatically reduced when using UL-based measurements.⁹

ENERGY EFFICIENCY

Periodical signaling for beam tracking, besides increasing the system complexity, results in additional resource and energy consumption. Additionally, GPS coordinates and orientation need to be acquired over a certain amount of time when CI-based initial access schemes are performed and, consequently, energy needs to be consumed.

Finally, the use of digital/hybrid BF architectures can reduce the IA discovery time and meet the low-latency 5G requirements, but at the expense of requiring much higher energy consumption to feed multiple RF chains. As described in [8], low-resolution digital architectures can be a viable solution. In [5], we showed how a UL-based IA framework, where the BS receives sounding reference signals broadcast by the UE, can better exploit the digital BF delay improvements due to less stringent power constraints.

CONCLUSIONS AND FUTURE WORK

In this work, we have studied, analyzed, and compared some possible implementations of initial access techniques for 5G mmWave cellular networks, where we argued that directionality should also be used in the initial synchronization-access phase. Our analysis has indeed demonstrated the following key findings:

- There is a trade-off between IA delay and misdetection probability: on one hand, compared to exhaustive algorithms, iterative techniques require less time to perform the angular search; on the other hand, iterative schemes exhibit higher misdetection probabilities in general, as wider beams provide reduced gains.
- To guarantee a minimum coverage level at relatively large distances (around 100 m from the BS) in realistic channel environments, exhaustive procedures incur smaller discovery delay and therefore are to be preferred. Otherwise, for closer users (e.g., in ultra-dense cell deployments), iterative techniques still present acceptable delays.

⁹ Since the base station is less power constrained than a mobile device, digital or hybrid BF will likely be more feasible at the BS side.

- The misdetection probability of pure CI-based approaches is higher than for exhaustive search when considering cell edge users. However, when implementing a simple refinement of the direct link, steering beams through adjacent directions, the PMD drops to acceptable levels, making it possible to reduce the discovery delay.

As part of our future work, the simultaneous steering of narrow beams in multiple directions through digital and hybrid beamforming architectures will be considered. Furthermore, we will study the cost of obtaining CI as well as the performance implications of CI inaccuracy. Finally, we will consider methods that, through historical data about past initial access, can better capture the dynamics of the channel and drive the selection strategy toward more robust cells.

REFERENCES

- [1] L. Wei *et al.*, "Key Elements to Enable Millimeter Wave Communications for 5G Wireless Systems," *IEEE Wireless Commun.*, vol. 21, no. 6, Dec. 2014, pp. 136–43.
- [2] M. Giordani *et al.*, "Comparative Analysis of initial Access Techniques in 5G mmWave Cellular Networks," *Proc. 50th Annual Conf. Info. Sciences and Systems*, 2016.
- [3] C. Barati *et al.*, "Directional Cell Discovery in Millimeter Wave Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, Dec 2015, pp. 6664–78.
- [4] H. Shokri-Ghadikolaei *et al.*, "Millimeter Wave Cellular Networks: A MAC Layer Perspective," *IEEE Trans. Commun.*, vol. 63, no. 10, Oct 2015, pp. 3437–58.
- [5] M. Giordani *et al.*, "Multi-Connectivity in 5G mmWave Cellular Networks," *2016 15th Annual Mediterranean Ad Hoc Networking Wksp.*, Vilanova i la Geltrú, Barcelona, Spain, June 2016.
- [6] Q. Ye *et al.*, "User Association for Load Balancing in Heterogeneous Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, June 2013, pp. 2706–16.
- [7] C. Jeong, J. Park, and H. Yu, "Random Access in Millimeter-Wave Beamforming Cellular Networks: Issues and Approaches," *IEEE Commun. Mag.*, vol. 53, no. 1, Jan. 2015, pp. 180–85.
- [8] C. N. Barati *et al.*, "Directional Initial Access for Millimeter Wave Cellular Systems," *CoRR*, vol. abs/1511.06483, 2015; <http://arxiv.org/abs/1511.06483>
- [9] V. Desai *et al.*, "Initial Beamforming for mmWave Communications," *48th Asilomar Conf. Signals, Systems and Computers*, Nov 2014, pp. 1926–30.
- [10] A. Capone, I. Filippini, and V. Sciancalepore, "Context Information for Fast Cell Discovery in mmWave 5G Networks," *Proc. 21th Euro. Wireless Conf.*, May 2015.
- [11] Q. Li *et al.*, "Anchor-Booster Based Heterogeneous Networks with mmWave Capable Booster Cells," *IEEE GLOBECOM Wksp.*, Dec 2013, pp. 93–98.
- [12] A. Capone *et al.*, "Obstacle Avoidance Cell Discovery Using mmWaves Directive Antennas in 5G Networks," *IEEE 26th Annual Int'l. Symp. Personal, Indoor, and Mobile Radio Commun.*, Aug 2015, pp. 2349–53.

- [13] W. B. Abbas and M. Zorzi, "Context Information Based Initial Cell Search for Millimeter Wave 5G Cellular Networks," *Proc. 25th Euro. Conf. Networks and Commun.*, 2016.
- [14] M. Akdeniz *et al.*, "Millimeter Wave Channel Modeling and Cellular Capacity Evaluation," *IEEE JSAC*, vol. 32, no. 6, June 2014, pp. 1164–79.
- [15] M. Giordani *et al.*, "Channel Dynamics and SNR Tracking in Millimeter Wave Cellular Systems," *Euro. Wireless 2016*, Oulu, Finland, May 2016.
- [16] M. Polese, M. Mezzavilla, and M. Zorzi, "Performance Comparison of Dual Connectivity and Hard Handover for LTE-5G Tight Integration," *Proc. 9th EAI SIMUtools Conf.*, Prague, Czech Republic, Aug 2016, <http://arxiv.org/abs/1607.05425>

BIOGRAPHIES

MARCO GIORDANI (giordani@dei.unipd.it) received his Bachelor's degree in information engineering in 2013 and Master's degree in telecommunication Engineering in 2015, both from the University of Padova, Italy. Since October 2015 he has been a postgraduate researcher at the Department of Information Engineering of the University of Padova under the supervision of Prof. Michele Zorzi. From January to April 2016 he spent a period abroad at New York University (NYU) as a visiting research scholar. His research interests include design and validation of protocols and applications to the next generation of cellular networks (5G) and in particular millimeter-wave communication, heterogeneous networks, initial access, and multi-connectivity.

MARCO MEZZAVILLA (mezzavilla@nyu.edu) is a postdoctoral researcher at NYU Tandon School of Engineering, where he leads various mmWave-related research projects, mainly focusing on 5G PHY/MAC design. He received his B.Sc. (2007) and M.Sc. (2010) in telecommunications engineering from the University of Padova, and his Ph.D. (2013) in information engineering from the same university, under the supervision of Prof. M. Zorzi. He held visiting research positions at the NEC Network Laboratories in Heidelberg, Germany (2009), at the Telematics Department at Polytechnic University of Catalonia in Barcelona, Spain (2010), and at Qualcomm Research in San Diego, California (2012). He has authored and co-authored multiple publications in conferences, journals, and patent applications. He serves as a reviewer for many IEEE conferences, journals, and magazines. His research interests include design and validation of communication protocols and applications to 4G broadband wireless technologies, millimeter-wave communications for 5G networks, multimedia traffic optimization, radio resource management, spectrum sharing, convex optimization, cognitive networks, and experimental analysis.

MICHELE ZORZI [F'07] (zorzi@dei.unipd.it) received his Laurea and Ph.D. degrees in electrical engineering from the University of Padova in 1990 and 1994, respectively. During academic year 1992–1993 he was on leave at the University of California San Diego (UCSD). After being affiliated with the Dipartimento di Elettronica e Informazione, Politecnico di Milano, Italy, the Center for Wireless Communications at UCSD, and the University of Ferrara, in November 2003 he joined the faculty of the Information Engineering Department of the University of Padova, where he is currently a professor. His present research interests include performance evaluation in mobile communications systems, random access in mobile radio networks, ad hoc and sensor networks and IoT, energy constrained communications protocols, 5G millimeter-wave cellular systems, and underwater communications and networking. He was Editor-in-Chief of *IEEE Wireless Communications* from 2003 to 2005, Editor-in-Chief of *IEEE Transactions on Communications* from 2008 to 2011, and is currently the founding Editor-in-Chief of *IEEE Transactions on Cognitive Communications and Networking*. He was Guest Editor for several Special Issues in *IEEE Personal Communications*, *IEEE Wireless Communications*, *IEEE Network*, and *IEEE JSAC*. He served as a Member-at-Large on the Board of Governors of the IEEE Communications Society from 2009 to 2011, and as its Director of Education from 2014 to 2015.

As part of our future work, the simultaneous steering of narrow beams in multiple directions through digital and hybrid beamforming architectures will be considered. Furthermore, we will study the cost of obtaining CI as well as the performance implications of CI inaccuracy.

The New Frontier in RAN Heterogeneity: Multi-Tier Drone-Cells

Irem Bor-Yaliniz and Halim Yanikomeroglu

The authors study the opportunistic utilization of low-altitude unmanned aerial platforms equipped with base stations (i.e., drone-BSs) in future wireless networks. In particular, they envision a multi-tier drone-cell network complementing the terrestrial HetNets. They investigate the advancements promised by drone-cells, and discuss the challenges associated with their operation and management.

ABSTRACT

In cellular networks, the locations of the RAN elements are determined mainly based on the long-term traffic behavior. However, when the random and hard-to-predict spatio-temporal distribution of the traffic (load, demand) does not fully match the fixed locations of the RAN elements (supply), some performance degradation becomes inevitable. The concept of multi-tier cells (heterogeneous networks, HetNets) has been introduced in 4G networks to alleviate this mismatch. However, as the traffic distribution deviates more and more from the long-term average, even the HetNet architecture will have difficulty in coping with the erratic supply-demand mismatch, unless the RAN is grossly over-engineered (which is a financially non-viable solution). In this article, we study the opportunistic utilization of low-altitude unmanned aerial platforms equipped with BSs (i.e., *drone-BSs*) in future wireless networks. In particular, we envisage a *multi-tier drone-cell* network complementing the terrestrial HetNets. The variety of equipment and non-rigid placement options allow utilizing multi-tier drone-cell networks to serve diversified demands. Hence, drone-cells bring the supply to where the demand is, which sets new frontiers for the heterogeneity in 5G networks. We investigate the advancements promised by drone-cells and discuss the challenges associated with their operation and management. We propose a drone-cell management framework (DMF) benefiting from the synergy among SDN, network functions virtualization, and cloud computing. We demonstrate DMF mechanisms via a case study, and numerically show that it can reduce the cost of utilizing drone-cells in multi-tenancy cellular networks.

INTRODUCTION

Transportation and communication technologies are major contributors to our lifestyles. Combining the state-of-the-art advancements in these two technologies, drone-assisted mobile communications has gained momentum rapidly. Drones equipped with transceivers, that is, drone base stations (drone-BSs) forming drone-cells, can help satisfy the demands of future wireless networks [1].¹ Moreover, they can utilize the latest radio access technologies (RATs), such as millimeter-wave (mmWave) and free-space optical communication (FSO). Miscellaneous assets of

drones and placement options provide the opportunity to create *multi-tier drone-cell networks* to enhance connectivity whenever, wherever, and however needed. Therefore, the main advantage of drone-cells is the radical flexibility they create.

The phenomenon of providing ubiquitous connectivity to diversified user and device types is the key challenge for fifth generation (5G) and beyond 5G wireless networks. The Achilles' heel of the proposed technologies, such as decreasing cell size, cloud radio access networks (C-RANs), distributed antenna systems (DASs), and heterogeneous network (HetNet) deployments, is their rather rigid design based on long-term traffic behavior [2]. In case of unexpected and temporary events creating hard-to-predict inhomogeneous traffic demand [3], such as natural disasters, traffic congestions, and concerts, wireless networks may need additional support to maintain ubiquitous connections. Drone-cells address this need by increasing relevance between the distributions of supply (BSs) and demand (user traffic). They can be used opportunistically to leverage the heterogeneity, that is, by dynamically deploying BSs with different power levels and RATs.

Although discussions on utilizing drone-cells in cellular networks have flourished recently [1, 4], the readiness of cellular networks to employ such dynamic nodes has not been discussed. For instance, drone-cells require seamless integration to the network during their activity and seamless disintegration when their service duration is over. This requires the capability of configuring the network efficiently, for which configuration and management flexibilities, and self-organizing capabilities of the Third Generation Partnership Project (3GPP) Long Term Evolution (LTE) networks may not be adequate. Hence, updating the network, such as for adding new applications, tools, and technologies, is time and money consuming [5]. Also, massive amounts of granular information about users and networks must be continuously collected and analyzed by intelligent algorithms. Collecting, storing, and processing big data is challenging for existing wireless networks [2]. Moreover, it is not yet clear how to balance centralized (e.g., mobile cloud) and distributed (e.g., mobile edge computing) paradigms [5].

Recent proposals for future wireless network architectures aim to create a flexible network with improved agility and resilience. Cloud computing,

¹ Drone connectivity scenarios in recent 3GPP Release 14 documents (e.g., 3GPP TR 22.862 V14.0.0 (2016-06)) only include remote control of drones, which is different from the vision of drone-cells. Also, considering the limited time remaining until the development of 5G standards, we envision that drone-BSs can be utilized in beyond-5G/6G wireless networks (rather than 5G).

software-defined networking (SDN), and network functions virtualization (NFV) have been proposed to relax the entrenched structure of wireless networks, increase openness, ease configuration, and utilize cloud computing for storing and analyzing big data. At the same time, these technologies may decouple the roles in the business model into infrastructure providers (InPs), mobile virtual network operators (MVNOs), and service providers (SPs) [6], which also changes the owners and sources of information.

In order to utilize drone-cells in future wireless networks, we propose a drone-cell management framework (DMF), and discuss the related business and information models. The proposed framework relies on creating intelligence from big data in the cloud and reconfiguring the network accordingly by SDN and NFV. In the following section, we describe the drone-cells, the motivations for utilizing them in wireless networks, and the challenges. Then we introduce DMF, and discuss business and information models and challenges. Finally, we demonstrate the fundamental principles of DMF via a case study. The Conclusion section closes the article.

DESCRIPTIONS, OPPORTUNITIES, AND CHALLENGES

A drone-BS is a low-altitude² unmanned aerial vehicle equipped with transceivers to assist wireless networks [1], and a drone-cell is the corresponding coverage area. The size of a drone-cell varies based on the drone-BS's altitude, location, transmission power, RATs, antenna directivity, type of drone, and the characteristics of the environment. Hence, multi-tier drone-cell networks can be constructed by utilizing several drone types, which is similar to terrestrial HetNets with macro-, small-, femtocells, and relays. A multi-tier drone-cell network architecture, assisting the terrestrial HetNets in several cases, is depicted in Fig. 1.

Drone-cells are useful in scenarios requiring agility and resiliency of wireless networks because they can prevent over-engineering. These type of scenarios can be categorized as *temporary*, *unexpected*, and *critical*, as shown in Table 1, where relevant test cases of the METIS³ project are listed [8]. Based on the scenario, the benefit to the network from a drone-cell varies. For instance, in traffic jam, stadium, and dense urban information society scenarios, a drone-cell can help prevent unexpected or temporary congestion in the network. Alternatively, drone-cells can improve resilience of wireless networks by providing additional coverage in case of a natural disaster, or by enabling teleprotection for the smart grid.

Critical scenarios have challenging demands, such as very high data rates, high reliability, or low energy consumption. Beyond the benefits to the network, providing connectivity in some of these scenarios is important to prevent serious losses, for example, by saving lives in emergency communications, or increasing the lifetime of sensors and actuators in hard-to-reach areas. In case of emergency communications and tele-control applications, drone-cells can enable high data rates and reliability, especially for situations in which the conventional modes of wireless access are either not present or difficult to establish.

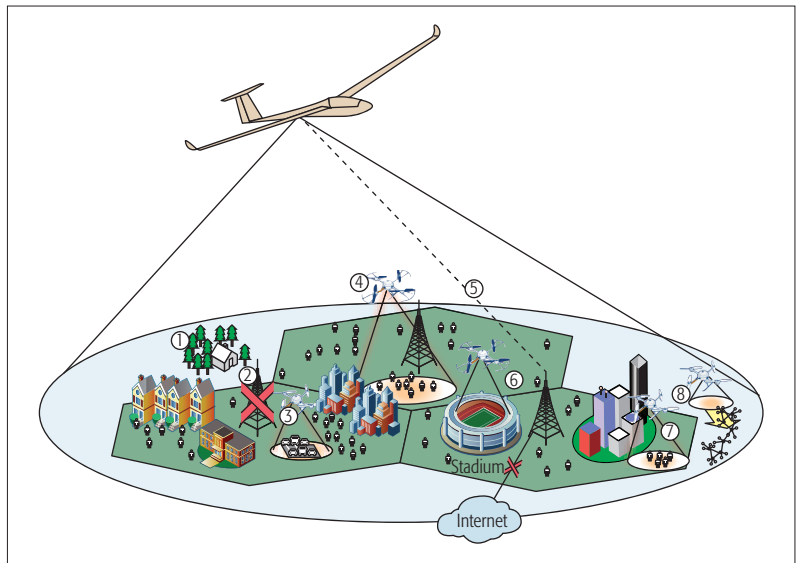


Figure 1. Multi-tier drone-cell networks can be used for many scenarios: ① providing service to rural areas (macro-drone-cell); ② deputizing for a malfunctioning BS (macro-drone-cell); ③ serving users with high mobility (femto-drone-cell); ④ assisting a macrocell in case of RAN congestion (pico-drone-cell); ⑤ assisting a macrocell in case of core network congestion (macro-drone-cell); ⑥ providing additional resources for temporary events (e.g., concerts and sports events); ⑦ providing coverage for temporary blind spots; ⑧ reducing energy dissipation of sensor networks by moving toward them (femto-drone-cell).

Mobility of drone-cells enables them to serve users with high mobility and data rate demand (e.g., for traffic efficiency and safety) [8]. Alternatively, sensor-type devices requiring low energy consumption can benefit from drone-cells. Instead of forcing low-power devices to transmit to farther BSs, or deploying small cells densely, mobile sinks can be used. A drone-BS can move toward clusters of devices and provide low-power communication due to its proximity and potential line-of-sight (LOS) connectivity. In particular, when unexpected events trigger massive sensor activity, drone-cells can reduce the overall stress on the network and increase the lifetime of sensors. Note that the critical scenarios, in which the conventional wireless access options are not feasible, may render them as the first applications of drone-cells in providing (almost) carrier-grade service.

Although the flexibility of drone-cells allows utilizing them in versatile scenarios, it creates significant design, operation, and management challenges, which are discussed next.

CHALLENGES OF DRONE-CELLS

Efficient Design: Drones have been utilized for military, surveillance, and reconnaissance applications for a long time. However, their usage in cellular communications as drone-BSs is a novel concept under investigation. For instance, a preliminary implementation of an LTE eNodeB-based drone operation was presented in [4], where a remote radio head (RRH) was deployed on an off-the-shelf helikite. The helikite was tethered to a truck carrying the baseband unit (BBU), and optical fiber is used for the fronthaul. This tethered helikite design is due to the nonexistence of drones that are specifically designed to operate as drone-BSs. Drones are

² The classification of drones is a rather involved task due to their variety [7, Ch. 5]. However, in this context, the term “low-altitude” is used to differentiate the drone-BSs from the high altitude platforms (HAPs) operating over 20 km.

³ Mobile and Wireless Communications Enablers for Twenty-Two (2020) Information Society.

Test case	Temporary	Unexpected	Critical
Stadium	X		
Teleprotection in smart grid		X	
Traffic jam	X	X	
Blind spots	X	X	
Open air festival	X		
Emergency communications	X	X	X
Traffic efficiency and safety			X
Dense urban information society	X		
Massive deployment of sensor-type devices	X	X	X

Table 1. An example of categorization of test cases of METIS requiring agility and resilience. An event can fall under one category or multiple categories and each combination may require different solutions. For instance, connectivity requirements in case of an only temporary event (e.g., stadium) may be addressed by over-engineering. Then, expenses of drone-BS operations may be compared to the expenses of over-engineering, including energy and maintenance costs. On the other hand, for both temporary and unexpected events, (e.g. traffic jam), drone-BSs may be utilized opportunistically. For temporary, unexpected and critical operations (e.g., emergency communications) drone-cells can provide much more than revenue, such as saving lives.

generally designed for their task, which is the reason for their great variety [7, Ch. 5].

Drone-BSs would have unique requirements that can benefit from special-purpose designs, such as long-time hovering, long endurance, robustness against turbulence, minimum wingspan allowing multiple-input multiple-output (MIMO), and provision of energy for transmission (in addition to flying). For instance, a hybrid-drone can be designed with vertical take-off capability of rotorcrafts and with collapsible wings (equipped with MIMO antenna elements and solar panels for energy harvesting), which can be unfolded for efficient gliding.

Designing the payload of drone-BSs is as important as determining their mechanics, such as size, aerodynamics, and maximum take-off weight [7, Ch. 9]. For efficient usage of the limited volume, weight, and energy of drone-BSs, the payload can vary according to the scenario. Several possible drone-cell configurations are listed below.

Drone-relay (Drolay): Compared to small- or macro-BSs, relays require less processing power, because their RRH may be relatively simple, and they may not require an onboard BBU. Hence, they operate with light payloads, that is, additional equipments to the ones required for a drone's own operation, and potentially consume less power. The size and weight of RAN nodes may not be critical for terrestrial HetNets; however, a lighter payload improves endurance and decreases capital and operational expenditure (CAPEX and OPEX) significantly in drone-cell operations.

Small-drone-BS: They resemble terrestrial small-BSs with wireless backhaul. If a reliable wireless fronthaul can be maintained despite the mobility of drone-BSs, its advantage is two-

fold: First, it alleviates the weight and processing power required for an onboard BBU. Second, if combined with C-RAN, it can allow cooperation. C-RAN is useful particularly for dense HetNets [2], or when a fleet of drone-BSs are deployed. Scenarios ③, ④, ⑦, and ⑧ in Fig. 1 exemplify potential usage.

Macro-drone-BS: They resemble terrestrial macro-BSs with wireless backhaul. They can be deployed for longer endurance, broader coverage, or increased reliability of the network, for example, ①, ⑤, and ⑥ (Fig. 1). BBU can be included if a reliable wireless backhaul exists. Since coverage is strongly related to altitude and power, macro-drone-BSs may have a larger size, which allows more payload (e.g., medium-altitude long-endurance drones) [7, Ch. 113].

In addition to the discussion above, efficient drone-cell design can be enhanced by advancements on low-cost and lightweight energy harvesting, high-efficiency power amplifiers, beyond visual LOS operations, and alternative fuels, to name a few.

Backhaul/Fronthaul Connection: In terrestrial networks, wireless backhaul/fronthaul is considered when fiber connectivity is unaffordable (e.g., dense HetNets or rural BSs). However, it is inevitable for multi-tier drone-cell networks. FSO and mmWave are promising for their high rate and low spectrum cost. However, their reliability and coverage are limited, especially for inclement weather conditions [9, 10]. Although mobility of drone-cells help maintains LOS, it necessitates robustness against rapid channel variations.

Placement: Terrestrial BSs are deployed based on long-term traffic behavior and over-engineering when necessary. However, drone-cells require quick and efficient placement. Therefore, it is of critical importance to determine the parameters affecting a drone-cell's performance, such as its altitude, location, and trajectory, based on the network demands [1, 11]. For instance, if a drone-cell is utilized to release congestion in RAN within a congested cell, the target benefit is to offload as many users as needed to the drone-cell [1]. Particularly, if the congestion is at the cell edge, the drone-cell can be placed right on top of the users there. On the other hand, if the congestion is at the backhaul, some of the most popular contents can be cached in a drone-cell for *content-centric placement*. Moreover, placement of multi-tier drone-cell networks requires integrated evaluation of many other challenges.

CHALLENGES OF MULTI-TIER DRONE-CELL NETWORKS

There are additional challenges of multi-tier drone-cell networks. Although these challenges are similar to those of terrestrial HetNets, the particular details related to drone-cells are discussed here.

Physical layer signal processing: The link between the drone-cell and terrestrial nodes (i.e., air-to-ground links) has different characteristics than terrestrial channels [1, 12]. However, the research on air-to-ground links is not mature, and the proposed channel models vary depending on factors such as temperature, wind, foliage, near-sea environments, urban environments, and the aircraft used for measurement campaigns, to name a few. For instance, higher ground speed

causes rapid variation of spatial diversity; users at different locations with respect to the drone-BS can have different channel characteristics simultaneously [12]. Therefore, designing robust signaling mechanisms with strict energy constraints of drone-BSs is challenging.

Interference dynamics: Drone-cells in proximity can suffer from co-channel interference for their air-to-ground links, and backhaul/fronthaul. Moreover, a drone-BS's mobility creates Doppler shift, which causes severe inter-carrier interference for RATs at high frequencies (e.g., mmWave). In HetNets, interference of terrestrial and air-to-ground-channels can decrease capacity. Therefore, advanced interference management schemes that consider the characteristics of air-to-ground links and mobility of drone-cells are required.

Cooperation among drone-cells: The dynamic nature of multi-tier drone-cell networks requires cooperation among drone-cells for efficiency in radio resource management. In addition to that, drone-cells can cooperate to adapt to the mobility of the users to decrease handover, optimize power and resource allocation, and avoid collisions.

Infrastructure decision and planning: The number and assets of drone-cells (e.g., access technology, memory, and speed) to be utilized for a multi-tier drone-cell network depend on circumstances, such as inclement weather conditions, size of the area to be served, type of service (e.g., virtual reality, the Internet of Things), target benefit of the network (e.g., congestion release, resilience, low latency), or service duration. Also, utilizing drone-cells with different access technologies can reduce interference and increase capacity of multi-tier drone-cell networks, for example, utilizing a macro-drone-cell with RF and small-drone-cells with mmWave to prevent frequency reuse. Hence, InPs must have a fleet that can respond to possible scenarios. To optimize the fleet and construct an efficient network, information sharing among all parties of the network (i.e., InPs, MVNOs, and SPs) is required.

Cost, lack of regulations, security, and air worthiness are among other challenges of drones. The vital point to consider is the effects of utilizing drones in highly sophisticated cellular communication networks, rather than using them for standalone applications such as aerial photography or inspection. Therefore, drone-cells require an equivalently sophisticated management system, which is discussed next.

THE DRONE-CELL MANAGEMENT FRAMEWORK

A drone-cell is not a one-size-fits-all solution; instead, it is tailored based on the target benefit. Along with the management of individual drone-cells, multi-tier drone-cell networks require active organization and monitoring (e.g., for nodes changing location or cells becoming congested). Three capabilities are required to integrate drone-cells with already sophisticated cellular networks.

Global information: The information gathered by BSs alone may be inadequate to generate intelligence for managing drone-cells. Global information, including location, type, and habits

of users, functionality of BSs, and contents to deliver must be stored and analyzed centrally. Big data and cloud computing can be effective solutions for that purpose.

Programmability: Both drone-cells and network tools need to be programmed based on network updates. Moreover, sharing the resources made available by a drone-cell can reduce CAPEX and OPEX. NFV can provide these capabilities.

Control: Wireless networks must be configured efficiently for seamless integration/disintegration of drone-cells, such as changing protocols and creating new paths. SDN can be useful to update the network automatically via a software-based control plane.

The current LTE architecture does not embody all of these abilities, but cloud, SDN, and NFV technologies can enable a more capable wireless communication system [2].

ENABLING TECHNOLOGIES FOR DMF

In this subsection, we briefly explain the technologies that increase capabilities of wireless networks and the interactions that are required to efficiently manage drone-cell-assisted wireless communications.

Cloud and Big Data: There are many ways to approach the problem of collecting and processing sufficient data (Table 1) in a timely manner for efficiently utilizing drone-cells. A cloud for drone-cells, consisting of computing power and data storage (Fig. 2), combined with big data analysis tools, can provide efficient and economic use of centralized resources for network-wide monitoring and decision making [5, 13]. If drone-BSs are owned by a traditional mobile network operator (MNO) (Fig. 2), the cloud is merely the data center of the MNO (similar to a private cloud), where the MNO as an administrator can choose to share its knowledge with some other players or use it for its own business purposes. Alternatively, if the drone-BSs are provided by an InP, the InP can use the cloud to collect information from MVNOs and SPs (Fig. 2 and Table 2). In this case, it is particularly important to guarantee security, latency, and privacy. The benefit of the cloud can be exploited better with a programmable (softwarized) network allowing dynamic updates based on big data processing, for which NFV and SDN can be enabling technologies.

Network Functions Virtualization: NFV alleviates the need for deploying specific network devices (e.g., packet and serving gateways, deep packet inspection modules, and firewalls) for the integration of drone-cells [5]. By virtualizing the above-network functions on general-purpose servers, standard storage devices, and switches, NFV allows a programmable network structure, which is particularly useful for drone-cells requiring seamless integration to the existing network (Ⓢ in Fig. 2). Furthermore, virtualization of drone-cells as shared resources among M(V)NOs can decrease OPEX for each party [6]. However, the control and interconnection of virtual network functions (VNFs) becomes complicated, for which SDN can be useful [5].

Software Defined Networking: By isolating the control and data planes of network devices, SDN provides centralized control, a global view of the

The vital point to consider is the effects of utilizing drones in highly sophisticated cellular communication networks, rather than using them for stand-alone applications, e.g., aerial photography or inspection. Therefore, drone-cells require an equivalently sophisticated management system, which is discussed next.

In traditional cellular networks, an MNO owns almost the entire cellular network, such as BSs and core network, and sharing among MNOs is limited. However, future cellular networks may be partitioned between InPs, MVNOs and SPs [6]. For instance, high sophistication of drone operations may result in the drone-cell operator becoming a separate business entity, such as a drone-InP.

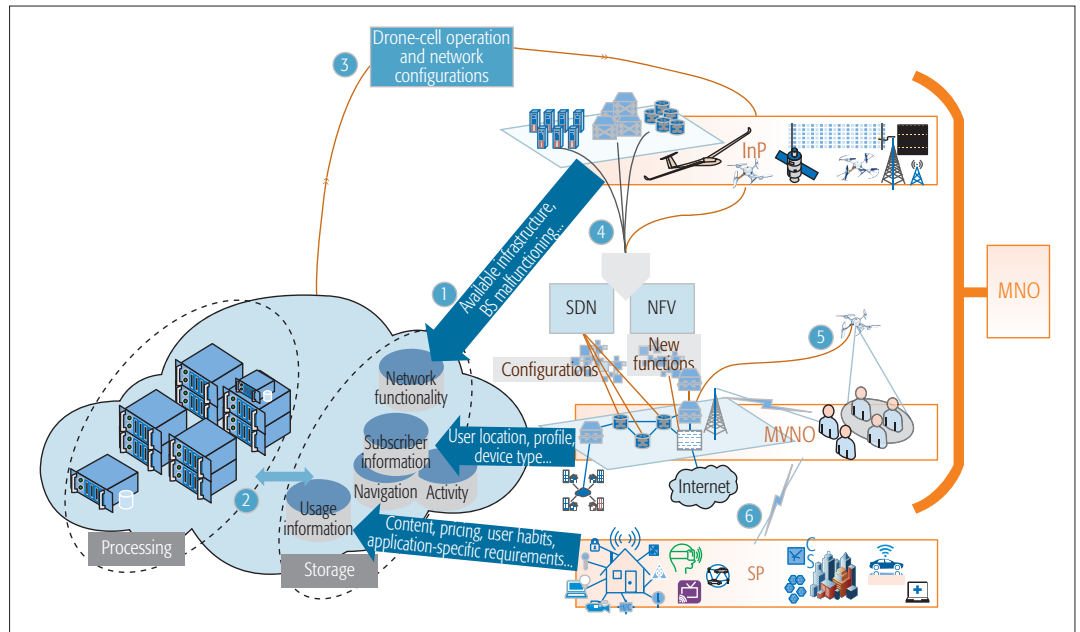


Figure 2. DMF mechanism and potential business and information model demonstrating partitioning of the traditional MNO into InP (cloud, server, drone-BS etc.) and MVNO: ① collect and store global data; ② process data for network monitoring and creating intelligence; ③ provide guidance for drone-cell's operation (placement, content to be loaded, access technology, service duration, coverage area, moving patterns); ④ reconfigure the virtual network of the MVNO for drone-cell integration by SDN and NFV technologies; for example, introduce another gateway to handle busy traffic and create new paths among the new and existing functions; ⑤ drone-cell assists the network; ⑥ SP can continue delivering services successfully.

network, easy reconfiguration, and orchestration of VNFs via flow-based networking (④ in Fig. 2). Specifically for cellular networks, a centralized SDN controller can enable efficient radio resource and mobility management [5], which is particularly important to exploit drone-cells. For instance, SDN-based load balancing, proposed in [5], can be useful for multi-tier drone-cell networks, such that the load of each drone-BS and terrestrial-BS is optimized precisely. An SDN controller can update routing such that the burst of traffic from the drone-cells is carried through the network without any bottlenecks [13]. Similarly, in the case of a natural disaster that causes the network to partially malfunction, network health information in the cloud can be utilized via SDN to route the traffic of drone-cells through undamaged parts of the network. Because SDN allows updating switches simultaneously (e.g., for new forwarding rules), it allows faster switching between RATs [14], which eases utilizing different RATs in multi-tier drone-cell networks. Furthermore, the architecture based on hierarchical SDN controllers for unified handoff and routing proposed in [14] can allow granular management of flows through drone-cells. For instance, the handoff strategy can be changed to a more complex proactive handoff for decreasing the latency of flows from drone-cells. Alternatively, DMF may collaborate with the mobility management entities for efficiency; for example, a drone-cell can follow high-mobility users on a highway (③ in Fig. 1) to reduce handover. For further exploitation of the new degree-of-freedom introduced by the mobility of drone-cells, the footprint of drone-cells can be adjusted to optimize paging and polling, and

location management parameters can be updated dynamically via the unified protocols of SDN.

BUSINESS AND INFORMATION MODELS OF DMF

In traditional cellular networks, an MNO owns almost the entire cellular network, such as BSs and core network, and sharing among MNOs is limited. However, future cellular networks may be partitioned between InPs, MVNOs, and SPs [6]. For instance, the high sophistication of drone operations may result in the drone-cell operator becoming a separate business entity, such as a drone-InP.

Figure 2 represents a DMF with potential business and information models, and shows what is owned by these parties, and what information flows from them to the cloud. According to the model, all physical resources of the cellular network, including drone-cells, BSs, spectrum, and core network, are owned by InPs. The MVNO is responsible for operating the virtual network efficiently such that the services of the SP are delivered to the users successfully. Note that in this model, perfect isolation and slicing is assumed such that an MVNO has a complete virtual cellular network [6].

Compared to traditional cellular networks, more granular data is available, but it is distributed unless collected in a cloud. A brief list of information, which can be critical for the operation of the DMF, is provided in Table 2 along with its type, source, and usage [5]. The results of the processing are then used to orchestrate SDN and NFV for the purpose of integrating drone-cells in the networks. This mechanism is demonstrated later.

Note that such isolated business roles may

not be realistic for the near future. Instead, the role of an MNO may get partitioned into three actors: InP, MVNO, and SP. Since it will mature in the long run, this partitioning should not be considered as siloing, but rather specialization. Accordingly, unique pricing strategies and QoS monitoring requirements will likely appear for drone-cell operations. Although complex and expensive, drone-cell operations can increase revenues in several ways, such as enabling a leaner terrestrial network, service to high-priority users (e.g., for public safety), and continuity of challenging services even in cases of unpredictable high density traffic in areas with relatively insufficient infrastructure.

CHALLENGES FOR DMF IMPLEMENTATION

Network management required for DMF involves the challenges of NFV and SDN. Slicing of drone-cells, isolation of the traffic of different MVNOs, migration of VNFs, virtual resource management, and scheduling can be listed among the major challenges related to NFV [6]. Regarding the SDN in DMF, the main challenges are providing a global view to the SDN controller, that is, scalability, efficiency in programming new paths, and communicating with different virtual network entities and application interfaces [15]. In particular, latency as a performance indicator is critical for drone-cells. Flow- and cloud-based networking are promising approaches to overcome these challenges [5, 13–15].

Flow-based networking requires advancements, such as developing new routing protocols, interfaces, and applications. The major difficulties associated with the cloud are centralizing the distributed data, providing security, determining the level of sharing while satisfying the regulations, and providing the power required for processing massive amounts of data [2, 13]. In this sense, real-time collection and processing of the data required to manage a drone's operation (e.g., tackling turbulence, avoiding collisions, tracking user mobility) is infeasible. Therefore, DMF is unlikely to alleviate the need for drones with high levels of autonomy [7, Ch. 70], but DMF can provide guidelines, as demonstrated in the following section.

A CASE STUDY: 3D PLACEMENT OF A DRONE-CELL VIA DMF

Efficient placement is a critical and challenging issue for drone-cells. In this section, we propose an objective for DMF, meeting various demands simultaneously. Then we numerically illustrate the benefit of using DMF by comparing the results with the efficient 3D placement⁴ method proposed in [1], and show that DMF can split costs among MVNOs without detracting from the network benefit in a multi-tenancy model.

Let us consider that a drone-cell, managed via DMF, is used to assist a terrestrial HetNet with the following considerations.

Congestion release in RAN: A set of users, \mathbb{U} , cannot be served by the BS because of congestion. The objective is to serve as many users from the set \mathbb{U} as possible by the drone-cell. Let u_i denote a binary variable indicating whether the i th user in \mathbb{U} is served by a drone-cell with

Information	Type	Source	Use
International Mobile Subscriber Identity (IMSI)	User	MNO	True identity of the user
User profile information	User	MVNO	Subscription type, activities
User's location	Network	MVNO	Location
Device type	Network	MVNO	Location, resource allocation provisioning, etc.
Functionality of the nodes	Network	InP	Location, coverage extension, energy saving, etc.
User's activity and navigation	Network	MVNO	Placement, consumption, lifestyle, etc.
Content	Usage	SP	Centers of interest, preferences, pricing, content delivery, etc.
Long-term historic data	Usage	SP	Content delivery, pricing, etc.

Table 2. Various information that can be gathered in the cloud.

orthogonal resources. Note that \mathbb{U} is determined by MVNOs based on the connection characteristics of each user [5] (Table 2).

Multi-tenancy: An InP owns the drone-cell and sends it to the congested macrocell according to the intelligence provided by the cloud (Fig. 2). This network structure allows sharing the drone-BS's resources, if desired, to maximize revenue and reduce OPEX. Assuming all users provide the same revenue (as in [1]), the number of users associated with an MVNO and served by the drone-cell can be a measure of the revenue provided to that MVNO. Hence, the objective becomes maximizing the number of served users, as well as forcing the drone-cell to serve the target number of users of each MVNO. Then, if the total number of MVNOs in the macrocell is J , a $J \times 1$ vector \mathbf{v} can be calculated, such that its j th element, v_j , denotes the ideal number of MVNO $_j$'s users to be served by the drone-cell. Also, the cloud must store the vector \mathbf{u} containing u_i , which indicates whether user i is served by the drone-BS, and the matrix \mathbf{S} , which denotes the user-MVNO associations. $S(i, j) \in \{0, 1\}$ indicates whether user i belongs to MVNO $_j$, which can be known from the subscriber information in the cloud (Table 2). Note that \mathbf{v} is derived by cloud computing based on several factors, such as agreements between the InP and MVNOs, pricing, user mobility, requested contents, and the scenario (Table 2, Fig. 2).

Green wireless communications: Λ represents the energy cost of users. Hence, the drone-cell can be placed close to energy-critical users, such as sensor-type devices or those in blind spots (\odot in Fig. 1). Device-type information is collected by the MVNO (Table 2).

Content-centric placement/congestion release at the backhaul: κ_i indicates if user i requests a popular and costly content (e.g., in terms of bandwidth or price), κ , which is cached in the drone-cell. Hence, the placement can be adjusted according to the content requirements of the users. Note that decisions about which contents are to be delivered depend on the short- and long-term data collected by SPs on usage, user habits, and so on (Table 2 and Fig. 2).

⁴ The 3D placement concept is introduced in [1] because the probability of having LOS connection increases with increasing altitude, while at the same time, path loss increases due to increased distance. Therefore, an optimum altitude is sought after, as well as an optimal area to cover in the horizontal domain.

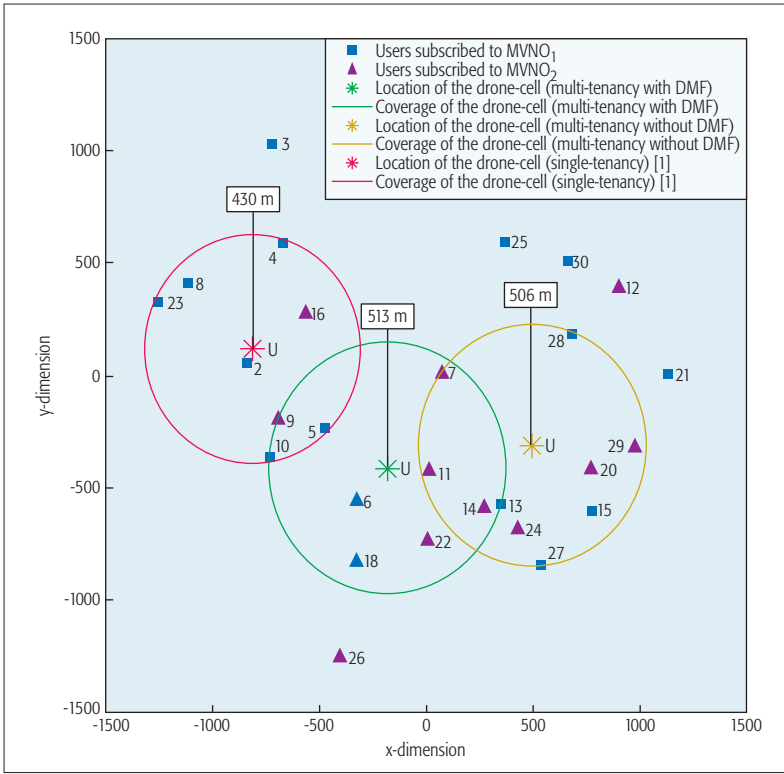


Figure 3. Effect of different policies on 3D placement of a drone-BS.

Then a comprehensive placement problem can be written as

$$\begin{aligned}
 & \max_{\mathbf{p}, \{u_i\}} \omega_1 \sum_{i \in U} u_i + \omega_2 \|\mathbf{S}\mathbf{u} - \mathbf{v}\| + \omega_3 \|\mathbf{S}\mathbf{u} - \Lambda\| + \omega_4 \sum_i \kappa_i \\
 & \text{s.t. } Q(\mathbf{p}, u_i) \geq \mathbf{q}_i, \forall i = 1, \dots, |U|, \\
 & \sum_{i \in U} u_i R_i \leq C, \forall i = 1, \dots, |U|, \\
 & u_i \in \{0, 1\}, \forall i = 1, \dots, |U|,
 \end{aligned} \tag{1}$$

where $|\cdot|$ and $\|\cdot\|$ represent the cardinality of a set and vector norm operation, respectively; ω represents the weight of each benefit; \mathbf{p} denotes the location of the drone-cell in 3-D space; $Q(\mathbf{p}, u_i)$, q_i , and R_i denote the QoS delivered to the i th user from the drone-cell at location \mathbf{p} , the minimum tolerable service quality such as signal-to-noise ratio (SNR), and the required resources to serve the i th user (e.g., bandwidth), respectively. C represents the capacity of the drone-cell, and \mathbb{P} denotes the set of allowable locations for placing the drone-cell, such as the allowed distance from buildings according to regulations, or the positions with LOS links to the backhaul/fronthaul node. Note that the weights among the benefits, ω_i , can be determined based on their importance to the owner of the drone-cells. Determining ω_i , \mathbf{v} , and κ_i , based on their importance to the owner of the drone-cells, is an interesting problem in itself.

The generic problem in Eq. 1 is mathematically formulated in [1] by assuming $\omega_1 = 1$, and the rest of the weights are 0. The air-to-ground channel model in [1] relates the size of a drone-cell to the altitude of the drone-BS. Therefore, both horizontal and vertical coordinates of a drone-BS must be determined simultaneously. Hence, an

efficient 3D placement algorithm is proposed to find the optimal altitude, as well as an optimal area to cover in the horizontal domain [1]. As a result, a maximum number of users are covered with a minimum required area. In this study, 3D placement of a drone-cell is improved over [1] to allow and regulate multi-tenancy by DMF.

If single-tenancy is considered, only users subscribed to MVNO₁ are served by the drone-BS (as in [1]). If multi-tenancy is allowed, users of both MVNO₁ and MVNO₂ can be served. In this case, it is important to regulate the amount of service delivered to each MVNO so that their agreements with the InP are not violated. Therefore, we assume that $\omega_1 = \omega_2 = 1$, and $\omega_3 = \omega_4 = 0$ corresponds to the case of regulated multi-tenancy with DMF, and only $\omega_1 = 1$ corresponds to either single-tenancy, or multi-tenancy without DMF.

For a numerical comparison, assume that there are 30 users that cannot be served by a terrestrial HetNet. They are distributed uniformly and arbitrarily subscribed to one of the two available MVNOs. The QoS requirement for all users is the minimum SNR (100 dB maximum tolerable path loss). Also, MVNOs are identical, for example, in terms of their agreements with InP, user priorities, and QoS requirements. Therefore, $v_1 = v_2 = 15$, which is in favor of providing an equal amount of service to each MVNO. Hence, they can share the cost of the drone-cell equivalently.

Figure 3 shows how the placement of a drone-cell changes with respect to policies, that is, single-tenancy and multi-tenancy with and without DMF. The circular areas indicate the coverage of the drone-cell, and enclosed users are served by the drone-BS (i.e., their QoS requirements are satisfied). In the single-tenancy case, the coverage area is shown by the red circle, and a total of 6 users of MVNO₁ (users shown with blue squares 2, 4, 5, 8, 10, 23) are covered. Note that users 9 and 16 belonging to MVNO₂ are not served in this case. On the other hand, 10 users are enclosed in both the green and orange drone-cells with multi-tenancy. In the orange drone-cell representing the placement without DMF, six users belong to MVNO₂, and four users belong to MVNO₁. Hence, the resources of the drone-BS are not equally distributed as suggested by the cloud. That may reduce the benefit of the network; for example, MVNO₁ may reject the drone-BS's services. However, when DMF is considered, five users of each MVNO are served in the green drone-cell. At the same time, there is no compromise in the network's benefit, since the total number of served users remains the same in both multi-tenancy scenarios. Note that not only single- or multi-tenancy (red vs. green and orange circles), but also regulating the service among MVNOs changes the placement (green vs. orange circles).

In order to clarify the advantage of DMF, we compared single-tenancy [1] with multi-tenancy regulated by DMF. In Fig. 4, 30 idle users in four different environments [1] are randomly distributed, and the results of 100 Monte Carlo simulations are averaged. It shows that MVNO₁ serves almost the same number of users (one to two users less in each case) when it shares the drone-cell with MVNO₂. In turn, the drone-cell's cost can be reduced by a factor of two. Moreover, the

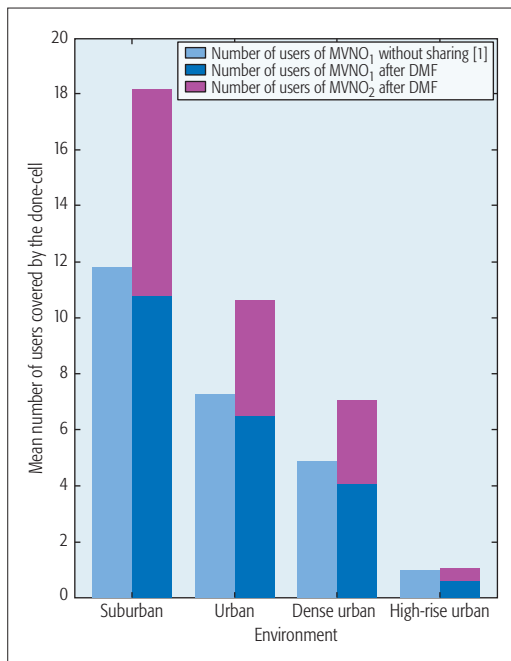


Figure 4. Mean number of users covered by the drone-cell with 3D placement in different environments.

total number of served users increases (approximately 1.5 times), which means that more congestion is released from the network.

Although it has remained implicit due to the limitations of this article, the number of covered users can also indicate the amount of injected capacity, enhanced coverage, and reduced retransmission time in a congested scenario. Moreover, we have demonstrated the 3D placement of one drone-cell, although multi-tier drone-cell networks require additional considerations, such as inter-cell interference, cell density, cooperation of drone-cells, and green networking. Therefore, collecting data to define the problem in Eq. 1 and then analyzing it efficiently requires a holistic and centralized cellular network rather than the existing distributed one. The better drone-cells are managed, the more the advantages of their flexibility can be exploited.

CONCLUSION

Ultra-dense small cell deployment has attracted significant attention in recent years as an advanced radio access architecture to cope with extreme traffic demands. However, the fact that such extreme demands can often be sporadic and hard to predict in space and time renders an ultra-dense deployment (which will end up being underutilized most of the time) highly inefficient and even prohibitive from a cost perspective. The multi-tier drone-cell network envisioned in this article is a new radio access paradigm that enables bringing the supply of wireless networks to where the demand is in space and time.

We discuss the potential advantages and challenges of integrating drone-cells in future wireless networks with a holistic and detailed approach from the mechanics of drone-BSs to potential applications of advanced networking technologies. Considering the fact that wireless networks are mainly designed for the mobility of the users but

not the BSs, and that the drone-cell operations can be highly complex, we propose a novel drone management framework for efficient operation. We demonstrate the proposed DMF and its benefits via a case study, where drone-cells are utilized in wireless networks with multi-tenancy. Although the effects of user mobility are not shown here, *dynamic 3D placement* strategies will be particularly important for access technologies at higher frequencies, such as mmWave.

ACKNOWLEDGMENTS

This work was supported in part by Huawei Technologies Canada and in part by the Ontario Ministry of Economic Development and Innovations Ontario Research Fund — Research Excellence Program.

REFERENCES

- [1] I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu, "Efficient 3-D Placement of an Aerial Base Station in Next Generation Cellular Networks," *IEEE ICC*, May 2016.
- [2] P. Demestichas *et al.*, "5G on the Horizon: Key Challenges for the Radio-Access Network," *IEEE Vehic. Tech. Mag.*, vol. 8, no. 3, Sept. 2013, pp. 47–53.
- [3] M. Mirahsan, R. Schoenen, and H. Yanikomeroglu, "HetHetNets: Heterogeneous Traffic Distribution in Heterogeneous Wireless Cellular Networks," *IEEE JSAC*, vol. 33, no. 10, Oct. 2015, pp. 2252–65.
- [4] S. Chandrasekharan *et al.*, "Designing and Implementing Future Aerial Communication Networks," *IEEE Commun. Mag.*, vol. 54, no. 5, May 2016, pp. 26–34; doi: 10.1109/MCOM.2016.7470932.
- [5] A. Bradai *et al.*, "Cellular Software Defined Networking: A Framework," *IEEE Commun. Mag.*, vol. 53, no. 6, 2015, pp. 36–43.
- [6] C. Liang and F. Yu, "Wireless Virtualization for Next Generation Mobile Cellular Networks," *IEEE Wireless Commun.*, vol. 22, no. 1, Feb. 2015, pp. 61–69.
- [7] K. P. Valavanis and G. J. Vachtsevanos, *Handbook of Unmanned Aerial Vehicles*, Springer, 2015.
- [8] ICT-317669 METIS Project, "Scenarios, Requirements and KPIs for 5G Mobile and Wireless System," Del. D1.1, tech. rep., May 2013, <https://www.metis2020.com/documents/deliverables>, accessed Jan. 11, 2016.
- [9] H. Kaushal and G. Kaddoum, "Optical Communication in Space: Challenges and Mitigation Techniques," to appear, *IEEE Commun. Surveys & Tutorials*; doi: 10.1109/COMST.2016.2603518
- [10] U. Siddique *et al.*, "Wireless Backhauling of 5G Small Cells: Challenges and Solution Approaches," *IEEE Wireless Commun.*, vol. 22, no. 5, Oct. 2015, pp. 22–31.
- [11] E. Kalantari, H. Yanikomeroglu, and A. Yongacoglu, "On the Number and 3D Placement of Drone Base Stations in Wireless Cellular Networks," *IEEE VTC-Fall*, 2016.
- [12] T. Willink *et al.*, "Measurement and Characterization of Low Altitude Air-to-Ground MIMO Channels," *IEEE Trans. Vehic. Tech.*, no. 99, 2015, pp. 1–1.
- [13] X. Zhou *et al.*, "Toward 5G: When Explosive Bursts Meet Soft Cloud," *IEEE Network*, vol. 28, no. 6, Nov. 2014, pp. 12–17.
- [14] V. Yazici, U. C. Kozat, and M. O. Sunay, "A New Control Plane for 5G Network Architecture with a Case Study on Unified Handoff, Mobility, and Routing Management," *IEEE Commun. Mag.*, vol. 52, no. 11, Nov. 2014, pp. 76–85.
- [15] S. Sezer *et al.*, "Are We Ready for SDN? Implementation Challenges for Software-Defined Networks," *IEEE Commun. Mag.*, vol. 51, no. 7, 2013, pp. 36–43.

BIOGRAPHIES

IREM BOR-YALINIZ (irembor@sce.carleton.ca) received her B.Sc. and M.Sc. degrees in electrical and electronics engineering from Bilkent University, Turkey, in 2009 and 2012, respectively. She worked at Aselsan, which is a leading defense company, where she was a design engineer for physical and data layer embedded coding of professional radio systems. She is currently pursuing her doctorate degree at Carleton University, Ottawa, Canada. She received scholarships through the Engage grant of the Natural Sciences and Engineering Research Council of Canada (NSERC) in 2014, and the Queen Elizabeth II Scholarship in Science and Technology in 2015.

HALIM YANIKOMEROGLU (halim@sce.carleton.ca) is a full professor in the Department of Systems and Computer Engineering at Carleton University. His research interests cover many aspects of wireless technologies with special emphasis on cellular networks. He has co-authored more than 85 IEEE journal papers on wireless technologies. His collaborative research with industry has resulted in about 25 patents (granted and applied). He is a Distinguished Lecturer for the IEEE Communications Society and a Distinguished Speaker for the IEEE Vehicular Technology Society.

We demonstrate the proposed DMF and its benefits via a case study, where drone-cells are utilized in wireless networks with multi-tenancy. Although the effects of user mobility are not shown here, *dynamic 3D placement* strategies will be particularly important for access technologies at higher frequencies, such as mmWave.

Energy Consumption Minimization for FiWi Enhanced LTE-A HetNets with UE Connection Constraint

Jijia Liu, Hongzhi Guo, Zubair Md. Fadlullah, and Nei Kato

The authors study the issues in energy consumption minimization with the UE connection constraint in FiWi enhanced LTE-A HetNets, and propose a heuristic greedy solution to find an optimal list of active BSs and their associated UEs.

ABSTRACT

The fiber-wireless (FiWi) enhanced LTE-A HetNet, which consists of fiber optic networks as its backhaul and LTE-A HetNets as its wireless front-end, is regarded as a promising technique for a 5G radio access network to host large-scale mobile data transmission. In order to reduce the energy consumption of a FiWi enhanced LTE-A HetNet, we should offload the traffic of lightly loaded BSs to other active BSs and turn them into sleep state, while in order to provide stable service to more user equipments (UEs), we should put as many BSs into active state as possible. Obviously, there is a trade-off between minimizing the energy consumption and maximizing the number of UEs associated with the BSs. However, previous research works either focused on energy consumption minimization or placed emphasis on UE connection maximization, and few of them integrate these two parts together and give an optimal solution. Toward this end, this article studies the issues in energy consumption minimization with the UE connection constraint in FiWi enhanced LTE-A HetNets, and propose a heuristic greedy solution to find an optimal list of active BSs and their associated UEs.

INTRODUCTION

To cope with the explosion of mobile data traffic in the future fifth generation (5G), heterogeneous networks (HetNets), which adopt a variety of radio access technologies, such as macrocells, WiFi access points (APs), as well as low-cost low-power small cells, to achieve high data rates are expected to be a paradigm shift from traditional cellular networks [1, 2]. However, the development of HetNets faces new challenges, such as cell association, interference coordination, resource partitioning, and backhaul bottleneck, including the delay and reliability of backhaul links, the importance of which has gradually been recognized over the last few years [3]. By leveraging the complementary advantages of high capacity and reliability of passive optical networks (PONs), and the high mobility and ubiquitous connectivity of wireless networks, fiber-wireless (FiWi) networking is believed to be a promising solution to enhancing Long Term

Evolution-Advanced (LTE-A) HetNets, which gives rise to so-called FiWi enhanced LTE-A HetNets [4, 5]. In order to provide support for a large number of connected devices in future 5G applications including the Internet of Things (IoT) and others, FiWi enhanced LTE-A HetNets have been regarded as a compelling solution among various technologies due to their high capacity, reliability, flexibility, and extremely low latency characteristics [5].

High energy efficiency is one of the most important design goals in future 5G networks so as to support various battery hungry services from mobile user equipments (UEs) in different 5G applications. To reduce energy consumption and prolong UE battery life in accessing networks, different kinds of energy saving (ES) schemes have been proposed by means of switching some of the devices into sleep state periodically [6, 7]. Specifically, wireless local area networks (WLANs) mostly adopt the power saving mode (PSM) standardized by IEEE 802.11 as their ES scheme, which powers off UEs' transmitters periodically according to the data traffic situation. For LTE/LTE-A networks, discontinuous reception (DRX) has been introduced as an effective mechanism to save UE battery, where UEs do not need to monitor the downlink channel in real time, and enter sleep state when there is no data addressed to them. Furthermore, to address the energy consumption problem in PONs, optical network unit (ONU) sleep mode has been specified to switch the ONU transceiver to sleep/active state cyclically, since the ONUs are the major energy consuming units in PONs [8].

Up to now, there have been plenty of research activities focusing on reducing the energy consumption of FiWi networks; some of them have placed emphasis on the wireless end to design device ES schemes, some of them focused on the optical backhaul to design efficient ES mechanisms, and some presented new energy efficiency improving schemes by cooperation between the ES mechanisms of both the optical and wireless ends. Note that when we perform ES schemes in FiWi networks, some quality of service (QoS) requirements should also be satisfied, for example, UE connection (i.e., the number of UEs associated with base stations [BSs]) and delay. Recently some research works have been pre-

sented on reducing energy consumption of FiWi networks while taking packet delay into account. To find the best trade-off between power saving and latency in the DRX mechanism of LTE-A networks, Koc *et al.* [9] proposed an analytical model and presented a trade-off scheme to maintain a balance between these two performance indicators via DRX configuration. Nishiyama *et al.* [10] proposed a cooperative ONU sleep scheme that dynamically controls the ONU sleep period according to mobile UEs' ES mechanisms, and reduced latency and energy consumption were reported there. To reduce the overall energy consumption in PON LTE-A converged networks supporting machine-to-machine (M2M) communications, Van *et al.* [11] analyzed the energy consumption and end-to-end delay in M2M scenarios via an M/G/1 queuing model (optical backhaul) and a semi-Markov process (LTE-A front end), and proposed an ES scheme by incorporating the ONU sleep mode with the DRX mechanism. Zhou *et al.* [12] studied the energy-efficient context-aware resource allocation problem in ultra-dense small cells, and proposed an energy-efficient matching algorithm based on the Gale-Shapley algorithm.

Moreover, there are also some research activities focusing on improving UE connection in HetNets. To achieve load balancing among multi-tier LTE-A HetNets and improve UE connection, Liu *et al.* [13] proposed a load balancing algorithm by adopting device-to-device (D2D) communications to offload the associated UEs of congested BSs to adjacent uncongested BSs. Xu *et al.* [14] proposed a novel D2D local area network architecture by introducing device freedom to improve network capacity as well as energy efficiency.

Obviously, there is a trade-off between energy consumption minimization and UE connection maximization in FiWi enhanced LTE-A HetNets, where UE connection maximization is to maximize the number of UEs associated with the BSs in the network. On one hand, to minimize the overall energy consumption, we close the ONUs, BSs, and UEs as long as possible, and thus few UEs can access the Internet. On the other hand, to maximize the UE connection, we turn on all equipment as long as possible, which results in very high energy consumption. Note that previous works mostly focused on reducing energy consumption in FiWi networks or improving UE connection in LTE-A HetNets. Little work can be found on minimizing the energy consumption of FiWi enhanced LTE-A HetNets, especially with the UE connection constraint being considered, where the UE connection constraint denotes the predefined ratio of the UEs that associate with the BSs vs. all the UEs in the network. Toward this end, we provide this article to discuss the challenging issues in minimizing energy consumption with the UE connection constraint in FiWi enhanced LTE-A HetNets. Two algorithms are presented as our solutions: a brute force algorithm, which adopts an enumeration strategy to find an optimal solution for an active BS list and their associated UEs, and a heuristic greedy algorithm, which attaches the UEs to each BS greedily in the order of WiFi APs, pico evolved NodeBs (eNBs), and macro eNBs.

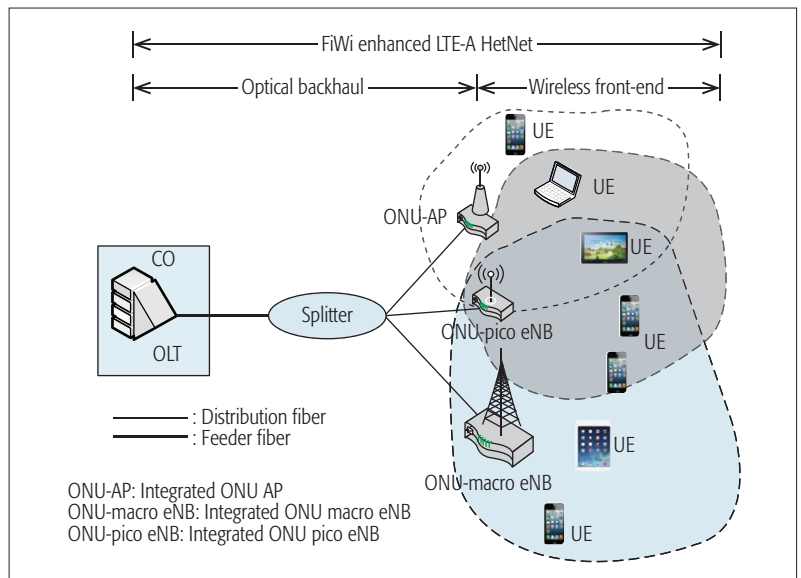


Figure 1. Illustration of a FiWi enhanced LTE-A HetNet.

The remainder of this article is organized as follows. An overview of FiWi enhanced LTE-A HetNets and some traditional ES schemes are illustrated. We define the problem of energy consumption minimization with the UE connection constraint, and present two solutions, a brute force algorithm and a heuristic greedy algorithm. Some numerical results are given to validate our algorithms, and we conclude the whole article.

OVERVIEW OF FIWI ENHANCED LTE-A HETNETS AND ENERGY SAVING SCHEMES

In this section, we present an architecture of FiWi enhanced LTE-A HetNets and review some classical ES schemes utilized in PONs, WLANs, and LTE-A networks: ONU sleep mode, PSM, and DRX.

FIWI ENHANCED LTE-A HETNETS

As depicted in Fig. 1, a FiWi enhanced LTE-A HetNet comprises two parts, the optical backhaul and the wireless front-end. In this article, we utilize cost-effective Ethernet PON (EPON), which has been widely deployed in lots of countries and areas for fiber optic access networks as the backhaul of the converged network. EPON can cover a range from 20 to 100 km and adopts a tree-based topology, where the optical line terminal (OLT) located at the central office (CO) forms the root, and a number of ONUs connected to the OLT via a 1:N splitter form the leaf nodes. In a FiWi enhanced LTE-A HetNet, an ONU can host an eNB or a WiFi AP to provide 4G LTE-A services or wireless access to UEs. WiFi APs and eNBs are distributed randomly with overlapping coverage in the wireless front-end. Nevertheless, it is noted that when a UE visits the network at any time, it can only utilize either an LTE-A network or a WiFi network.

ENERGY SAVING SCHEMES

ONU Sleep Mode in PONs: In ONU sleep mode, depending on whether the ONU has any data to transfer (send/receive), it switches to active state or sleep state periodically. During

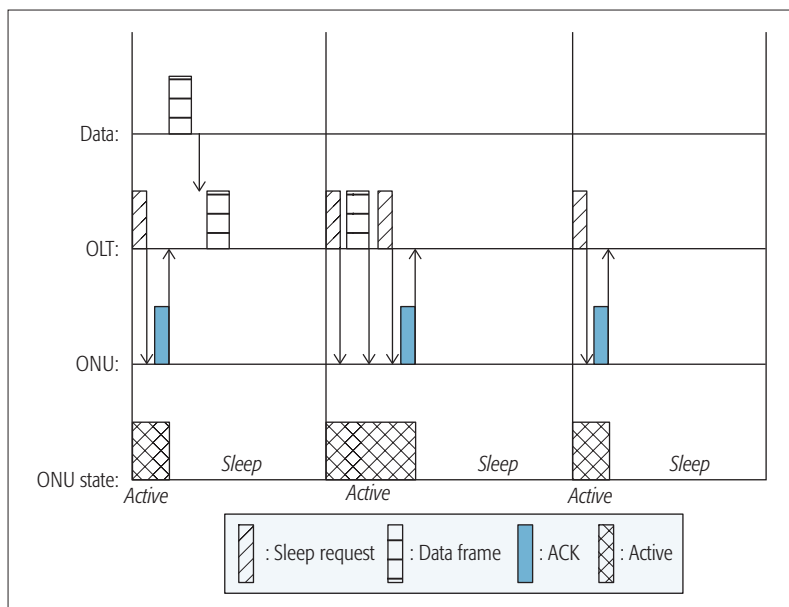


Figure 2. Illustration of the ONU sleep process in PONs.

active state, the ONU can receive and transmit data immediately, but consumes more energy. While in sleep state, it consumes little energy, but cannot send or receive data, and thus if there is incoming data from the OLT at this time, packet delay is unavoidable.

Figure 2 presents an example of the process of ONU sleep mode in PONs. When there is no incoming data at the OLT, it sends a sleep request (SR) message to the ONU, and then the ONU returns an acknowledge (ACK) message and enters sleep state for a sleep interval. During the sleep interval, no matter whether there is incoming data at the OLT, the ONU will not wake up unless the sleep period ends. After that, the ONU enters the active state and wakes up to check whether there is incoming data for it. If there is no arriving data, the OLT will send an SR message to the ONU with the predefined sleep interval, and the ONU switches to sleep state again. Otherwise, the OLT sends an SR message to the ONU with a sleep period of 0 ms, and then the data frames are transmitted to the ONU immediately. After that, the OLT resends an SR message to the ONU with the predefined sleep interval, and the ONU switches to the sleep state. From the analysis above, we can see that the additional packet delay caused by ONU sleep mode only appears when there is incoming data at the OLT and the ONU is in sleep state, where the data has to be buffered until the sleep period ends.

Power Saving Mode in WLANs: A UE in PSM has two states: active state, where the UE can send and receive data but consumes more energy, and sleep state, where the UE cannot receive/transmit any data but has lower energy consumption. A UE switches between these two states to reduce energy consumption in WLANs depending on whether there is any data to transfer.

Specifically, in the beginning of every beacon interval of PSM, the WiFi AP sends a beacon frame to the UE with a traffic indicator map (TIM). If the TIM indicates that there is no data

to transmit/receive, the UE enters sleep state until the beacon interval ends. Otherwise, the UE sends back a power save poll frame to request data from the AP, and then the AP transmits the buffered data to it. After that, the UE switches to sleep state until next beacon interval. It is noted that, during the whole process, if the data arrives at the AP when the UE is in sleep state, the incoming data has to first be buffered at the AP until the current beacon interval ends, and additional packet delay unavoidably occurs.

Discontinuous Reception in LTE-A Networks: A DRX mechanism configured by radio resource control (RRC) in LTE-A networks has two modes, RRC_CONNECTED and RRC_IDLE. In particular, if there is no data to transfer for a long period, the UE adopts the RRC_IDLE mode. Otherwise, it stays in the RRC_CONNECTED mode, where the UE is still hosted by an eNB during its sleep state.

Figure 3 illustrates a process of the DRX mechanism in LTE-A networks. Note that only a long DRX cycle is considered here since it represents most traffic scenarios. Generally, a UE using a DRX mechanism has four states: active state, listen state, sleep state, and sleep-to-active (S2A) state. The UE can send and receive data during the active state, which lasts for a predefined period. When the active state ends, if previous data transmission has not finished, the UE reenters the active state. Otherwise, it switches to the listen state to monitor the incoming data traffic. During the listen period, if there is any incoming data, the UE enters the active state immediately to perform data transmission. Otherwise, it goes to the sleep state to save energy after a configured period. Then the UE wakes up and proceeds to the S2A state, during which network synchronization is performed. After that the UE enters the active state and proceeds to another cycle. Depending on whether or not there is any incoming data, the UE switches to active or listen state accordingly.

ENERGY CONSUMPTION MINIMIZATION IN FiWi ENHANCED LTE-A HETNETS WITH THE UE CONNECTION CONSTRAINT

ENERGY CONSUMPTION MINIMIZATION WITH THE UE CONNECTION CONSTRAINT

The overall energy consumption in FiWi enhanced LTE-A HetNets mainly contains the energy consumed by ONUs in the optical backhaul, and the energy consumed by BSs and UEs in the wireless front-end. The ES schemes discussed can be adopted to reduce the energy consumption of ONUs and UEs. However, it is noted that among the equipment in the wireless front-end, BSs consume most of the energy [15]. Therefore, we present some analysis on energy consumption minimization in LTE-A HetNets in this section, where minimizing the energy consumption of BSs is our main focus. For LTE-A HetNets, due to a common assumption, the energy consumption of macro BSs and APs during downlink transmission includes the consumed transmit power and the non-transmit power consumption (the energy consumed in signal processing, equipment cooling, battery backup, etc.),

which is independent of data transmission. In this article, we only focus on the transmit powers related to data transmission, and non-transmit powers are outside the scope of this research.

Without loss of generality, we adopt a three-tier LTE-A HetNet as the wireless front-end of a FiWi enhanced LTE-A HetNet. In particular, the three-tier LTE-A HetNet consists of several macro eNBs and a number of pico eNBs, WiFi APs, and UEs randomly distributed in the coverage of these macro eNBs. Unless otherwise indicated, we use the general term BS to represent macro eNB, pico eNB, and WiFi AP in this article. Every WiFi AP is assumed to be open access. Moreover, each BS has a certain number of independent orthogonal channels to host a fixed amount of UEs, and the UEs within the BS coverage area can access the Internet by associating with any but one of them. For the case of multi-tier LTE-A HetNets including other types of small cells, the following algorithms can be followed similarly.

Assume that the backhaul of FiWi enhanced LTE-A HetNets is capable of handling the traffic demands in the wireless front-end. In order to reduce the overall energy consumption, we adopt cell load adaption in the wireless front-end, where lightly loaded BSs can be switched into sleep state via offloading their UE traffic to other active BSs. Besides, when we perform energy consumption reduction, a predefined UE connection ratio, which represents the ratio of UEs associated with the BSs vs. all UEs, should also be satisfied. Overall, the goal is to find an optimal solution that minimizes the overall energy consumption while satisfying a predefined UE connection constraint. Specifically, this problem can be defined as follows.

Energy Consumption Minimization with UE Connection Constraint (ECMCC): Given initial information, such as the topology of BSs and UEs, including their numbers, locations, channel numbers, and coverage radii, the numbers of OLTs and ONUs in the optical backhaul, the transmit powers of OLTs, ONUs, BSs, and UEs, and the UE connection constraint, find an optimal list of active BSs and their associated UEs that minimize the overall energy consumption in a FiWi enhanced LTE-A HetNet while satisfying a given UE connection constraint.

SOLUTIONS TO ECMCC IN FIWI ENHANCED LTE-A HETNETS

In the following, we present two schemes to solve the ECMCC problem in FiWi enhanced LTE-A HetNets. To simplify our algorithm presentation, some notations are predefined as follows. Let U , M , P , and A denote the sets of UEs, macro eNBs, pico eNBs, and WiFi APs, respectively, and G_U , G_M , G_P , and G_A separately denote their logical topology. S is the set of obtained active BSs, and U_i is the set of UEs associated with each active BS, $\forall i \in S$.

A Brute Force Algorithm: A brute force solution to ECMCC mainly contains three steps. First, we enumerate all possible combinations of active BSs; second, the UE connection ratios of all enumerated combinations are calculated; and finally, we can easily find an optimal combination

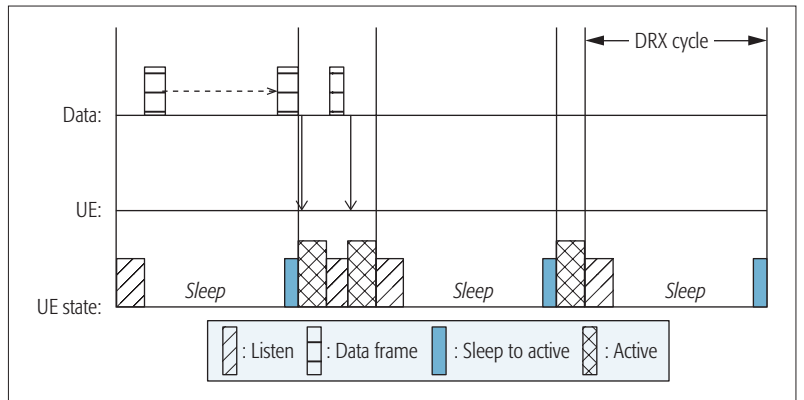


Figure 3. Illustration of the DRX process in LTE-A networks.

Input: $U, M, P, A, G_U, G_M, G_P, G_A; \alpha$.

Output: $S, U_i, \forall i \in S$, and overall energy consumption E ; otherwise, NIL.

```

1: Initialize  $E = 0, S = \emptyset$ ;
2: enumerate all combinations of active BSs and record into  $C$ ;
3: for all  $c \in C$ 
4:   calculate its maximum associated UEs,  $N_c$ ;
5:   if  $N_c/N_U \geq \alpha$  then
6:      $C' = C' \cup \{c\}$ ;
7:   else
8:     continue;
9:   end if
10: end for
11: if  $C' = \emptyset$  then
12:   return NIL;
13: else
14:   for all  $c \in C'$ 
15:     calculate the energy consumption,  $E_c$ ;
16:     if  $E_c < E$ 
17:        $E = E_c, S = c$ ;
18:     else
19:       continue;
20:     end if
21:   end for
22: end if
23: return  $S, E$ , and  $U_i, \forall i \in S$ .
```

Algorithm 1. A brute force algorithm for ECMCC.

of active BSs with the lowest energy consumption while satisfying the predefined UE connection constraint if an optimal solution exists. The details of the brute force algorithm (BFA) are described in Algorithm 1.

The brute force algorithm (BFA) presented as Algorithm 1 presents a simple method to solve the ECMCC problem. If optimal solutions to ECMCC exist, BFA is surely able to find one of them. However, BFA has the disadvantage of high computational complexity (i.e., $O(n \cdot 2^n)$). In particular, the running time of step 2, which enumerates all combinations of BSs, is $O(2^n)$ (n denotes the total number of BSs). The computational complexity of the for all loop from step 3 to step 10 is $O(n \cdot 2^n)$, where the outer loop runs 2^n times, and step 4 runs n times. Moreover, the running time of the for all loop of steps 14–21 is $O(2^n)$, where the outer loop runs 2^n times and steps 15–20 run in an $O(1)$ time. Finally, the computational complexity of Algorithm 1 can be calculated by adding them up (i.e., $O(n \cdot 2^n)$).

Obviously, to find an optimal solution to

For future 5G applications with a large number of connected devices, BFA and HGA provide energy-efficient solutions for network operators to obtain the optimal/near-optimal BS configurations with given UE connection constraints, i.e., a list of active BSs and their associated UEs with the highest overall energy efficiency.

Input: $U, M, P, A, G_U, G_M, G_P, G_A; \alpha$.
Output: $S, U_i, \forall i \in S$, and overall energy consumption E ; otherwise, NIL.
1: **Initialize** $E = 0, S = \emptyset$; N denotes the UE connection number, $N = 0$; $B = M \cup P \cup A$
2: **for all** $B_j \in B$
3: compute V_{B_j} ;
4: **end for**
5: **for all** $B_j \in B$
6: choose the BS B_k with the smallest V and record its associated UEs;
7: record its maximum associated UE number as N_{B_k} ;
8: $S = S \cup \{B_k\}, N+ = N_{B_k}$;
9: **if** $N/N_U \geq \alpha$ **then**
10: calculate the overall energy consumption, E ;
11: **return** S, E and $U_i, \forall i \in S$;
12: **else**
13: continue;
14: **end if**
15: **end for**
16: **return** NIL;

Algorithm 2. A heuristic greedy algorithm for ECMCC.

ECMCC, BFA may have to run for a very long time due to its exponential computational complexity, so that it can only be implemented with small number of BSs in practice. Therefore, BFA is just presented as a baseline for our following heuristic greedy solution.

A Heuristic Greedy Algorithm: It is noted that macro eNBs have the highest power consumption and the largest coverage in LTE-A HetNets so as to conserve as many UEs as possible, while both pico eNBs and WiFi APs have lower power consumption and smaller coverage. In order to minimize the overall energy consumption, we should turn on as few macro eNBs as possible. On the other side, to fulfill the UE connection constraint, more macro eNBs might have to be active. Therefore, in order to minimize the overall energy consumption with a given UE connection constraint, we introduce a variable V to denote the energy consumption/UE connection ratio of each BS, which can be computed via dividing the transmit power of a BS by its maximum associated UE number. After that a heuristic greedy solution to the ECMCC problem is proposed in the following. Note that if there are three kinds of BSs that host the same number of UEs (i.e., the same V s), it is better to adopt WiFi APs or pico eNBs due to their lower capital and operational expenditures (CAPEX and OPEX) than macro eNBs. Furthermore, WiFi APs have higher priority over pico eNBs since they transmit data traffic by adopting unlicensed spectrum. The details of our heuristic greedy algorithm (HGA) are shown in Algorithm 2.

Compared to BFA, HGA is a little more complex, but it can achieve better computational complexity (i.e., $O(n^2)$). In particular, the for all loop of steps 2–4 runs n times, and the computational complexity of steps 5–15 is $O(n^2)$, where the outer for all loop runs n times, and the running time of steps 4–12 is $O(n)$. Therefore, the overall computational complexity of HGA can easily be computed by adding them up (i.e., $O(n^2)$). Although HGA can only obtain a near-optimal solution, it is more practical and

can solve many scenarios with large numbers of BSs, which cannot be solved by BFA due to its high computational efficiency.

As discussed above, HGA is regarded as a workable solution to the ECMCC problem in the scenario of a FiWi enhanced LTE-A HetNet connecting to one OLT. For a large-scale FiWi enhanced HetNet, which consists of multiple zones connecting to different OLTs in the CO, our schemes can also be adopted. However, considering that the running time of solving the ECMCC problem in a large-scale network will be too long to be implemented, we can first divide the network into small independent zones according to the OLT to which they connect, and then execute HGA to optimize them separately.

Our proposed algorithms are mainly used to reduce the energy consumption of BSs in wireless front-ends. They not only hold for the ONU sleep mode, PSM, and DRX discussed earlier, but can also cooperate with other ES schemes for ONUs in PONs and those for UEs in wireless front-ends. For future 5G applications with a large number of connected devices, BFA and HGA provide energy-efficient solutions for network operators to obtain the optimal/near-optimal BS configurations with given UE connection constraints, that is, a list of active BSs and their associated UEs with the highest overall energy efficiency.

NUMERICAL RESULTS AND DISCUSSIONS

In this section, we present some comparison results in terms of overall energy consumption and running time between HGA and BFA, where BFA is adopted as our baseline.

EXPERIMENTAL SETTINGS

Without loss of generality, we adopt a FiWi enhanced LTE-A HetNet scenario in our experiments, where an EPON with 1 OLT and 64 ONUs is adopted as the optical backhaul, and a three-tier HetNet consisting of 1 macro eNB, 10 pico eNBs, and 10 WiFi APs is adopted as the wireless front-end. The transmit powers of OLT, ONU, macro eNB, pico eNB, and WiFi AP are set as 25 mW, 10 mW, 20 W, 200 mW, and 60 mW, respectively. The covering radii of macro eNB, pico eNB, and WiFi AP are set as 400 m, 100 m, and 50 m, respectively. Moreover, we use an Ubuntu 14 64-bit operating system in an Intel i5 core and 4 GB RAM computer, and a macro BS is assumed to have a frequency resource of 200 orthogonal channels, which means that it can provide services for at most 200 UEs at a time. Similarly, the maximum associated UE numbers of a pico eNB and a WiFi AP are assumed to be 60 and 20, respectively.

NUMERICAL RESULTS

Figure 4a illustrates the comparisons of energy consumption in 100 ms between BFA and HGA with different UE connection constraints, where the UE number is set as 2000. From the figure, one can easily see that HGA can find a near-optimal solution to the ECMCC problem compared to BFA. Moreover, with increasing UE connection constraints, the overall energy consumption increases accordingly, because more BSs have to be in active state to host more UEs.

Figure 4b shows the comparisons of running time between BFA and HGA with different number of BSs, where the UE number and UE connection constraint are set to 500 and 0.4, respectively. From the figure, we can observe that BFA has a much longer runtime than HGA, due to its exponential computational complexity. Furthermore, if the number of BSs continues to increase, the runtime of BFA will be so long that it cannot be adopted. Thus, BFA can only be implemented with a small number of BSs in practice.

Moreover, we further perform an evaluation on the energy consumption of HGA with different combinations of pico eNBs and WiFi APs. As illustrated in Fig. 4c, the UE number is set to 2000, and the total number of WiFi APs and pico eNBs is fixed as 50. It can easily be found that with the increase of WiFi APs, the overall energy consumption decreases, which corroborates the fact that WiFi APs have a greater advantage in power consumption than pico eNBs.

DISCUSSIONS

With a given topology of a FiWi enhanced LTE-A HetNet and the information of both BSs and UEs, the list of active BSs and their associated UEs should be provided as soon as possible, since the UE locations might change over time in practice. Doing so in real time is preferred, but it is challenging since a period of time is needed not only to obtain the topology of the whole network and the information of both BSs and UEs, but also to find an optimal solution to ECMCC. As our first step to study the ECMCC problem in FiWi enhanced LTE-A HetNets, we assume that the locations of both BSs and UEs are fixed. For more complex scenarios with moving UEs, our schemes could also work with slow moving devices by introducing continuous optimization and load-balancing techniques. However, for fast moving devices, the ECMCC problem should be redefined by taking multiple connectivity of UEs to different BSs into account, and new optimal solutions should be explored in future works.

CONCLUSION

To minimize the energy consumption in FiWi enhanced LTE-A HetNets with a given UE connection constraint, we study the issues in energy saving of LTE-A HetNets and define the ECMCC problem in this article. After that we present two solutions, that is, BFA, which adopts an enumeration strategy to find an optimal list of active BSs and their associated UEs, and HGA, which associates the UEs to the BSs greedily according to their energy consumption/UE connection ratios. Numerical results on the comparisons of energy consumption and running time between BFA and HGA are provided to show the promise of our proposed scheme. For future work, it should be meaningful to study the trade-off between energy consumption minimization and UE connection maximization, and take more QoS requirements into account while designing energy saving schemes for the overall network.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (61372073, 61373043, 61202394, 61472367, and 61432015),

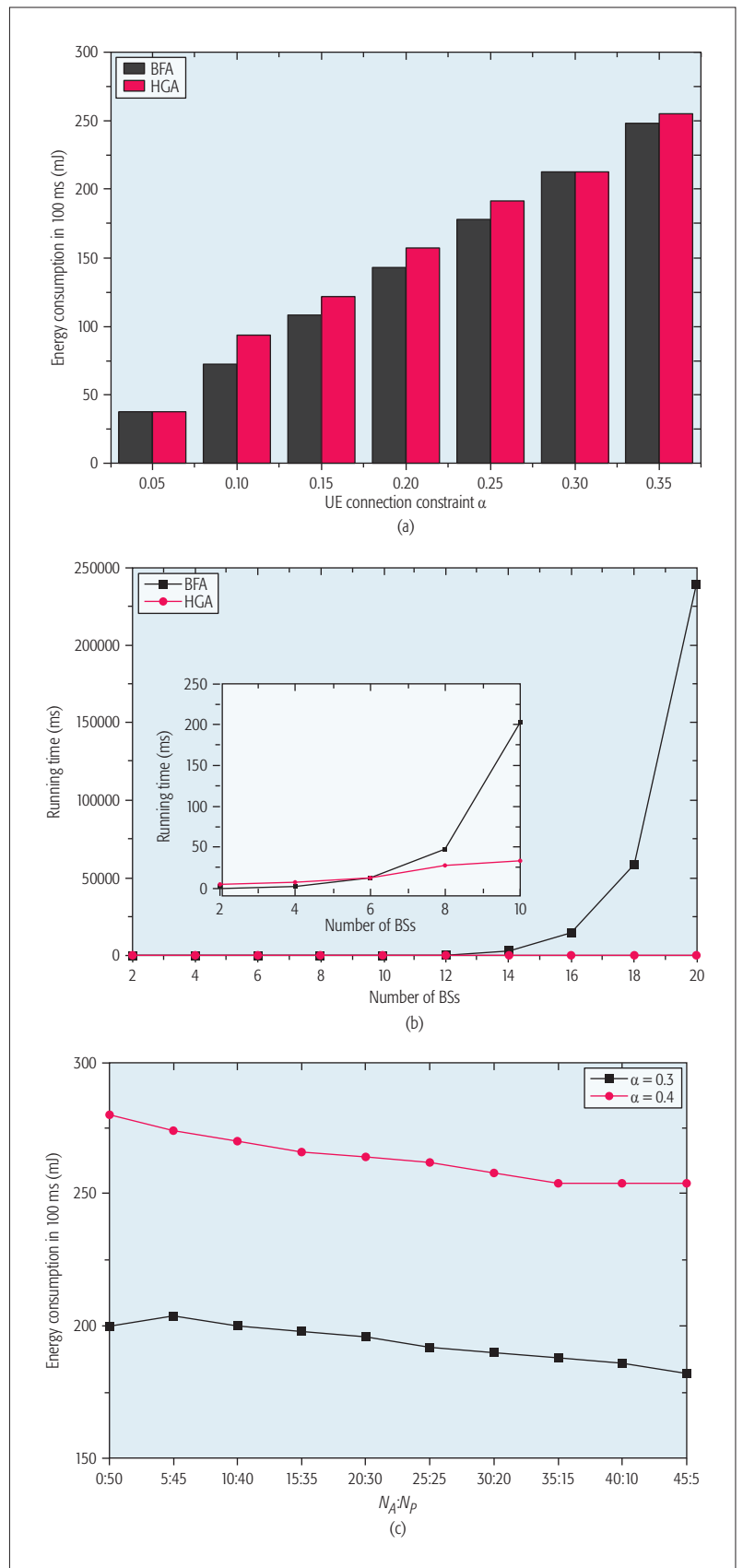


Figure 4. Illustration of numerical results of energy consumption and running time: a) comparisons of energy consumption in 100 ms between BFA and HGA with different UE connection constraints; b) comparisons of run-times between BFA and HGA; c) energy consumption of HGA in 100 ms with different numbers of WiFi APs and pico eNBs, where the total number of N_A and N_P is fixed as 50.

For future work, it should be meaningful to study the tradeoff between energy consumption minimization and UE connection maximization, and take more QoS requirements into account while designing energy saving schemes for the overall network.

in part by China 111 Project (B16037), and in part by the Fundamental Research Funds for the Central Universities (JB150311).

REFERENCES

- [1] A. Ghosh *et al.*, "Heterogeneous Cellular Networks: From Theory to Practice," *IEEE Commun. Mag.*, vol. 50, no. 6, June 2012, pp. 54–64.
- [2] R. Trivisonno *et al.*, "SDN-Based 5G Mobile Networks: Architecture, Functions, Procedures and Backward Compatibility," *Trans. Emerging Telecommun. Technologies*, vol. 26, no. 1, Jan. 2015, pp. 82–92.
- [3] J. Andrews, "Seven Ways that HetNets Are a Cellular Paradigm Shift," *IEEE Commun. Mag.*, vol. 51, no. 3, Mar. 2013, pp. 136–44.
- [4] N. Ghazisaidi, M. Maier, and C. Assi, "Fiber-Wireless (FiWi) Access Networks: A Survey," *IEEE Commun. Mag.*, vol. 47, no. 2, Feb. 2009, pp. 160–67.
- [5] H. Beyranvand *et al.*, "Backhaul-Aware User Association in FiWi Enhanced LTE-A Heterogeneous Networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, June 2015, pp. 2992–3003.
- [6] M. Hossain, K. S. Munasinghe, and A. Jamalipour, "Distributed Inter-BS Cooperation Aided Energy Efficient Load Balancing for Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, Nov. 2013, pp. 5929–39.
- [7] M. F. Hossain, K. S. Munasinghe, and A. Jamalipour, "On the eNB-Based Energy-Saving Cooperation Techniques for LTE Access Networks," *Wireless Commun. and Mobile Computing*, vol. 15, no. 3, Feb. 2015, pp. 401–20.
- [8] M. Hosseinabadi and N. Ansari, "Multi-Power-Level Energy Saving Management for Passive Optical Networks," *IEEE/OSA J. Optical Commun. and Networking*, vol. 6, no. 11, Nov. 2014, pp. 965–73.
- [9] A. Koc *et al.*, "Device Power Saving and Latency Optimization in LTE-A Networks through DRX Configuration," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, May 2014, pp. 2614–25.
- [10] H. Nishiyama *et al.*, "A Cooperative ONU Sleep Method for Reducing Latency and Energy Consumption of STA in Smart-FiWi Networks," *IEEE Trans. Parallel and Distrib. Sys.*, vol. 26, no. 10, Oct. 2014, pp. 2621–29.
- [11] D. Van, B. Rimal, and M. Maier, "Power-Saving Scheme for PON LTEA Converged Networks Supporting M2M Communications," *ICUWB*, 2015.
- [12] Z. Zhou *et al.*, "Energy-Efficient Context-Aware Matching for Resource Allocation in Ultra-Dense Small Cells," *IEEE Access*, vol. 3, Sept. 2015, pp. 1849–60.
- [13] J. Liu *et al.*, "Device-to-Device Communications Achieve Efficient Load Balancing in LTE-Advanced Networks," *IEEE Wireless Commun.*, vol. 21, no. 2, Apr. 2014, pp. 57–65.

[14] C. Xu *et al.*, "Efficiency Resource Allocation for Device-to-Device Underlay Communication Systems: A Reverse Iterative Combinatorial Auction Based Approach," *IEEE JSAC*, vol. 31, no. 9, Sept. 2013, pp. 348–58.

[15] S. Tombaz *et al.*, "Is Backhaul Becoming a Bottleneck for Green Wireless Access Networks?" *ICC*, 2014.

BIOGRAPHIES

JIAJIA LIU [S'11, M'12, SM'15] is currently a full professor at the School of Cyber Engineering, Xidian University. His research interests cover wireless mobile communications, FiWi, IoT, and so on. He has published more than 50 peer-reviewed papers in many prestigious IEEE journals and conferences, and currently serves as an Associate Editor for *IEEE Transactions on Communications* and *IEEE Transactions on Vehicular Technology*, an Editor for *IEEE Network*, and a Guest Editor of *IEEE TETC* and the *IEEE IoT Journal*. He is a Distinguished Lecturer of IEEE ComSoc.

HONGZHI GUO [S'08, M'16] received his B.S. degree in computer science and technology from Harbin Institute of Technology in 2004, and M.S. and Ph.D. degrees in computer application technology from Harbin Institute of Technology Shenzhen Graduate School, China, in 2006 and 2011, respectively. He is currently a lecturer with the School of Cyber Engineering, Xidian University. His research interests cover a wide range of areas including fiber-wireless networks, IoT, 5G, and smart grid.

ZUBAIR MD. FADLULLAH [S'06, M'11, SM'13] received his B.Sc. degree in computer sciences from Islamic University of Technology, Bangladesh, in 2003, and M.S. and Ph.D. degrees from GSIIS, Tohoku University, in 2008 and 2011, respectively. Currently, he is serving as an assistant professor at Tohoku University. His research interests cover smart grid, network security, intrusion detection, and more. He was a recipient of the prestigious Dean's and President's awards from Tohoku University in March 2011.

NEI KATO [A'03, M'04, SM'05, F'13] is currently a full professor at GSIIS, Tohoku University. He currently serves as a Member-at-Large on the Board of Governors, IEEE ComSoc, Chair of the IEEE Ad Hoc & Sensor Networks Technical Committee, Editor-in-Chief of *IEEE Network*, Associate Editor-in-Chief of the *IEEE IoT Journal*, and an Area Editor of *IEEE Transactions on Vehicular Technology*. He is a Distinguished Lecturer of IEEE ComSoc and the Vehicular Technology Society. He is a fellow of IEICE.

THE GLOBAL COMMUNITY OF COMMUNICATIONS PROFESSIONALS



Member Benefits

IEEE Communications Magazine
(electronic & digital delivery)

IEEE Communications Surveys and Tutorials
(electronic)

*Online access to IEEE Journal of Lightwave
Technology, IEEE OSA Journal of Optical
Communications and Networking and
IEEE RFID Virtual Journal*

Member Discounts

*Valuable discounts on conferences,
publications, IEEE WCET Certification program,
IEEE Training courses and other exclusive
member-only products.*

***These membership and exclusive benefits
will expand your technical community and
valuable networking opportunities!***

Join Now!
<http://bit.ly/1WH1tH5>



**If your technical interests are in communications, we encourage you to join
the IEEE Communications Society (IEEE ComSoc) to take advantage
of the numerous opportunities available to our members.**

www.comsoc.org

NEW WAVEFORMS FOR 5G NETWORKS



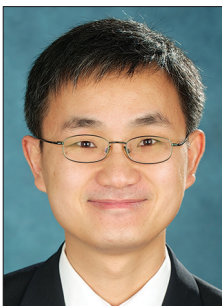
Charlie Jianzhong Zhang



Jianglei Ma



Geoffrey Ye Li



Wei Yu

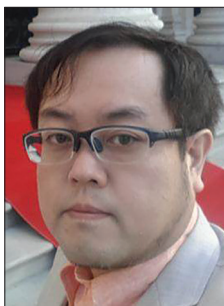


Nihar Jindal

With expected initial commercialization starting in 2018, fifth generation (5G) mobile communications is gathering increased interest and momentum around the world. The 5G vision and its associated key requirements, such as high data-rate in the multi-gigabit-per-second range, end-to-end latency of 1 ms, and massive capacity in terms of bits per second per square meter, as well in the number of devices per square kilometer, have been the focus of a lot of research in the last couple of years. There is also an increased interest in carrier frequencies above 6 GHz as a means to gain access to additional spectrum. While all the requirements need not be met simultaneously, the design of 5G networks and radio access should provide flexibility to more efficiently support various applications meeting part of the above requirements on a use case basis. The International Telecommunication Union (ITU) has categorized the usage scenarios for International Mobile Telecommunications (IMT) for 2020 and beyond into three main groups: enhanced mobile broadband, massive machine type communications, and ultra-reliable and low latency communications. In addition, they have specified target requirements to be fulfilled by IMT-2020-compliant radio access. Some of these targets can be met by Long Term Evolution (LTE), but there is also a need for a new 5G radio access technology to meet all the requirements.

Work on 5G technologies has been ongoing in academia and international projects for several years. Examples of the key technologies considered include massive antenna technologies (from legacy cellular frequency bands up to high frequencies) to provide beamforming gain and support increased capacity, new waveforms to flexibly accommodate various services/applications with different requirements, new multiple access schemes to support massive connections, and so on.

The Third Generation Partnership Project (3GPP) has initiated a study item on new 5G technologies in 2016, targeting completion by mid-2017. Based on these studies,



Yoshihisa Kishiyama



Stefan Parkvall

specification work will start, targeting initial specifications at the end of 2017 and fully IMT-2020-compliant specifications in mid-2019. One of the first agreements to be made in 3GPP is on the waveform. The waveform chosen must allow for a wide range of applications and usage scenarios, support large bandwidths (100 MHz and above), be able to also operate at high carrier frequencies, allow for high spectral efficiency, support various multi-antenna techniques, and allow for low-power and low-complexity design. Based on this, 3GPP has chosen orthogonal frequency-division multiplexing (OFDM) as the basic waveform, and the associated flexibility in numerology has been defined. However, waveforms for massive machine-type communication are still under discussion.

This Feature Topic contains six articles, giving an overview of some of the waveform and multiple access technologies discussed in recent research projects and considered for future 5G radio access technology.

The first article of the Feature Topic, “Introduction to QAM-FBMC: From Waveform Optimization to System Design” by C.-H. Kim, Y.-H. Yun, K.-Y. Kim, and J.-Y. Seol, discusses how to achieve better spectrum confinement, which is an important target of 5G waveform design. It addresses issues of CP-OFDM and OQAM-FBMC, and proposes a new waveform, QAM-FBMC, as a 5G waveform candidate. It also proposes a pulse shape optimization scheme as a way to improve waveform spectrum efficiency.

The second article, “On the Waveform for 5G” by X. Zhang, L. Chen, J. Qiu, and J. Abdoli, provides a comparative study of leading 5G waveform candidates, including CP-OFDM, W-OFDM, UFMC, f-OFDM, FBMC, and GFDM. The authors compared and contrasted these candidate waveforms according to several well selected criteria such as backward compatibility to existing transceivers, flexibility in supporting multiple numerology, spectrum efficiency, and out-of-band emission (OOBE) performance.

The third article, “Interference Management via Sliding-Window Coded Modulation for 5G Cellular Networks” by K.-T. Kim, S.-K. Ahn, J. Park, C.-Y. Chen, and Y.-H. Kim, discusses interference management of sliding-window coded modulation (SWCM). It describes the structure and principles of SWCM. Through extensive computer simulation, the article demonstrates that SWCM significantly outperforms conventional interference-aware schemes for multiple user networks.

The fourth article, “Waveform, Numerology, and Frame Structure to Support 5G Services and Requirements” by A. A. Zaidi, R. Baldemair, H. Tullberg, H. Björkegren, L. Sundström, J. Medbo, C. Kilinc, and I. Da Silva, outlines an OFDM-based 5G system, including the frame structure and multiple access scheme, capable of handling uplink, downlink, and sidelink transmissions. It is shown that OFDM is a good choice where the wide range of carrier frequencies, channel properties, and latency requirements calls for a scalable numerology.

The last two articles introduce a waveform design named DFT-s-OFDM. In the fifth article of the Feature Topic, “Generalized DFT-Spread-OFDM as 5G Waveform,” the authors introduce the generalized DFT-spread-OFDM (G-DFT-s-OFDM), a candidate 5G waveform that removes the need for explicit CP in the conventional DFT-s-OFDM, and in the meantime meets the enhanced 5G requirements of multi-service integration, asynchronous access support, and better scalability toward larger subcarrier spaces. The article also includes a good qualitative comparison between G-DFT-s-OFDM and other waveforms such as FBMC and UFMC.

The final article, “Flexible DFT-S-OFDM: Solutions and Challenges” by A. Sahin, R. Yang, E. Bala, M. C. Beluri, and R. L. Olesen, further explains several flavors of generalized DFT-s-OFDM waveforms, including GFDM and ZT-DFT-S-OFDM, as well as two new unique word (UW)-based waveforms. These new waveforms offer low OOB and peak-to-average power ratio, while achieving the same or better bit error rate performance compared to that of conventional CP DFT-S-OFDM.

We wish to thank all the authors for contributing to this Feature Topic, and wish to thank the reviewers for their valuable feedback. We hope that these articles provide the readers of the magazine a technical overview of the different waveforms and access schemes currently under consideration, as 5G becomes increasingly relevant in industry and academia, and as regulatory bodies, cellular operators, stan-

dardization bodies such as 3GPP, and chipset and device manufacturers increase their efforts to support new applications and requirements.

BIOGRAPHIES

CHARLIE JIANZHONG ZHANG [F] (jianzhong.z@samsung.com) is a VP and head of the Standards and Mobility Innovation Lab of Samsung Research America, where he leads research, prototyping, and standards for 5G cellular systems and future multimedia networks. He received his Ph.D. degree from the University of Wisconsin, Madison. From August 2009 to August 2013, he served as the Vice Chairman of the 3GPP RAN1 working group and led development of LTE and LTE-Advanced technologies such as 3D channel modeling, UL-MIMO and CoMP, Carrier Aggregation for TD-LTE, and so on. Before joining Samsung, he was with Motorola from 2006 to 2007 working on 3GPP HSPA standards, and with Nokia Research Center from 2001 to 2006 working on IEEE 802.16e (WiMAX) and EDGE/CDMA receivers.

JIANGLEI MA (jianglei.ma@huawei.com) is a Distinguished Engineer at the Huawei Canada R&D Center. Her research area is next generation wireless access technologies, and she is currently leading 5G air interface research in Huawei. She was a professor in the Department of Electronic Engineering at Southeast University, China, and a visiting associate professor in the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign. She received her Ph.D. degree from Southeast University.

GEOFFREY YE LI [F] (liye@ece.gatech.edu) is a professor with the School of ECE at Georgia Tech. His general research interests include wireless communications and statistical signal processing. In these areas, he has published around 400 articles with over 25,000 citations. He has been listed as a Highly-Cited Researcher by Thomson Reuters. He has received several awards from the IEEE Communications Society, the IEEE Vehicular Technology Society, and Georgia Tech.

WEI YU [F] (weiyu@ece.utoronto.ca) is a professor and Canada Research Chair in Information Theory and Wireless Communications at the University of Toronto, Canada. He obtained his Ph.D. from Stanford University in 2002. He currently serves on the IEEE Information Theory Society Board of Governors, and is an IEEE Communications Society Distinguished Lecturer. He received a Steacie Memorial Fellowship, IEEE Communications Society Best Tutorial Paper Award in 2015, and an IEEE Signal Processing Society Best Paper Award in 2008. He is a Highly Cited Researcher according to Thomson Reuters.

NIHAR JINDAL [F] (niharjindal@google.com) received a B.S. in electrical engineering/computer science from the University of California, Berkeley in 1999 and a Ph.D. in electrical engineering from Stanford University in 2004. He is currently a staff hardware engineer at Google, working on wireless projects within the Access & Energy area. From 2011 to 2014 he worked in WLAN systems design at Broadcom, on 802.11 standards and chip development. Prior to that he was an assistant/associate professor in the Department of Electrical Engineering at the University of Minnesota (2004–2010), where he taught courses and led a research lab focused on the analysis and development of innovative wireless technology. He has more than 30 publications in internationally recognized journals such as *IEEE Transactions on Wireless Communications*, and three of these publications have received best paper awards (IEEE Communications Society and Information Theory Society Joint Paper Award in 2005 and 2011, Best Paper Award for *IEEE Journal on Selected Areas in Communications* in 2009).

YOSHIHISA KISHIYAMA (kishiyama@nttdocomo.com) is a senior research engineer of the 5G Laboratory in NTT DoCoMo, INC., Japan. He received his B.E., M.E., and Dr.Eng. degrees from Hokkaido University, Sapporo, Japan, in 1998, 2000, and 2010, respectively. Since he joined NTT DoCoMo in 2000, he has been involved in research and development on 4G/5G radio access technologies and 3GPP standardization. In 2012, he received the ITU Association of Japan Award for contributions to LTE.

STEFAN PARKVALL [SM] (stefan.parkvall@ericsson.com) is a principal researcher at Ericsson Research, active in the area of 5G research and 3GPP standardization. He received his Ph.D. degree from the Royal Institute of Technology in 1996, served as an IEEE Distinguished Lecturer 2011–2012, and is co-author of several popular books such as *4G – LTE/LTE-Advanced for Mobile Broadband*. He has received the Ericsson Inventor of the Year award and the Swedish government’s Major Technical Award for his contributions to HSPA, and was nominated for the European Inventor Award for contributions to LTE.

Introduction to QAM-FBMC: From Waveform Optimization to System Design

Chanhong Kim, Yeo Hun Yun, Kyeongyeon Kim, and Ji-Yun Seol

The authors explain fundamental trade-offs in waveform design and show how they can be optimized for QAM-FBMC. They also introduce the transmission and reception procedures for QAM-FBMC including its similarities and differences compared to well-known waveforms such as CP-OFDM and OQAM-FBMC.

ABSTRACT

The asynchronous heterogeneous network scenario is becoming one of the key features for 5G mobile communications. Compared to CP-OFDM, which is 4G waveform, superior spectrum confinement as well as higher spectral efficiency has been taken into consideration for 5G waveform design. In this article, we introduce a new waveform called QAM-FBMC that provides superior spectrum confinement and higher spectral efficiency simultaneously against CP-OFDM. The article explains fundamental trade-offs in waveform design and shows how they can be optimized for QAM-FBMC. We also introduce the transmission and reception procedures for QAM-FBMC including its similarities and differences compared to well-known waveforms such as CP-OFDM and OQAM-FBMC. Performance evaluation results show where the spectral efficiency gain of QAM-FBMC over CP-OFDM comes from.

INTRODUCTION

Wireless communications and the Internet have fueled phenomenal growth in the demand for mobile data. An unprecedented number of devices now connect to the Internet using wireless technologies. The fourth generation (4G) technologies and network architecture, which were primarily conceived in the pre-smartphone age, are simply overwhelmed, and are unable to scale and keep up with this growth.

The fifth generation (5G) era has dawned. Industry efforts and investments to define, develop, and deliver the systems and specifications for 5G mobile communications and services are now well underway. 5G technologies are expected to deliver high-speed connectivity to support immersive applications, a fully realized Internet of Things (IoT) service framework, and lower latency with higher reliability in a spectrally efficient way.

To cope with the situation of a wireless data traffic crunch, a large increase in the supportable data traffic should be seriously considered by means of spectral efficiency improvement, bandwidth increase, or areal spectrum reuse. Apart from bandwidth increase and areal spectrum reuse, efforts to improve spectral efficiency are still untiring to cope with limited frequency

resources for wireless communications. Furthermore, since 5G use cases are too diverse, from enhanced mobile broadband to massive machine-type communications and ultra-reliable low latency communications, the main requirements for each use case are also very different each other, as shown in Fig. 1.

In wireless communications, the waveform can be defined as a time domain signaling that carries information data. The waveform also determines the spectral shape of a wireless communication system in the frequency domain according to its time domain pulse shape. Since it is becoming increasingly important that heterogeneous services with a diverse set of requirements can be supported simultaneously and efficiently within given frequency bands, well localized spectrum is a key requirement for the 5G waveform. Since the waveform also affects the design of the air interface, the Third Generation Partnership Project (3GPP) has started study on the waveform as an important physical layer technology for new radio [1].

Cyclic prefix orthogonal frequency-division multiplexing (CP-OFDM) has been adopted as the baseline waveform for the state-of-the-art radio access technologies (RATs) such as 4G Long Term Evolution (LTE) and WiFi. The main benefits of CP-OFDM are efficient sub-channelization for broadband multiple access, reduced equalization complexity from the use of a CP to handle multipath fading channels, and consequently efficient per-tone multiple-input multiple-output (MIMO) processing support. However, its good performance is valid only when synchronization and intersubcarrier orthogonality are strictly maintained. As such, research activities have been initiated to overcome the limitations of CP-OFDM in specific asynchronous signaling scenarios for 5G such as uplink multiple access without timing advance for massive machine type communications (mMTC), flexible duplex, and dynamic spectrum access for fragmented spectrum [2].

Although its theory has a long history close to OFDM [3], filter-bank multicarrier (FBMC) has recently been recognized as an enabling waveform to enhance fundamental spectral efficiency against CP-OFDM due to CP-less transmission and guardband reduction [4]. It is also considered as a good candidate for cognitive radio systems due to its well localized per-sub-

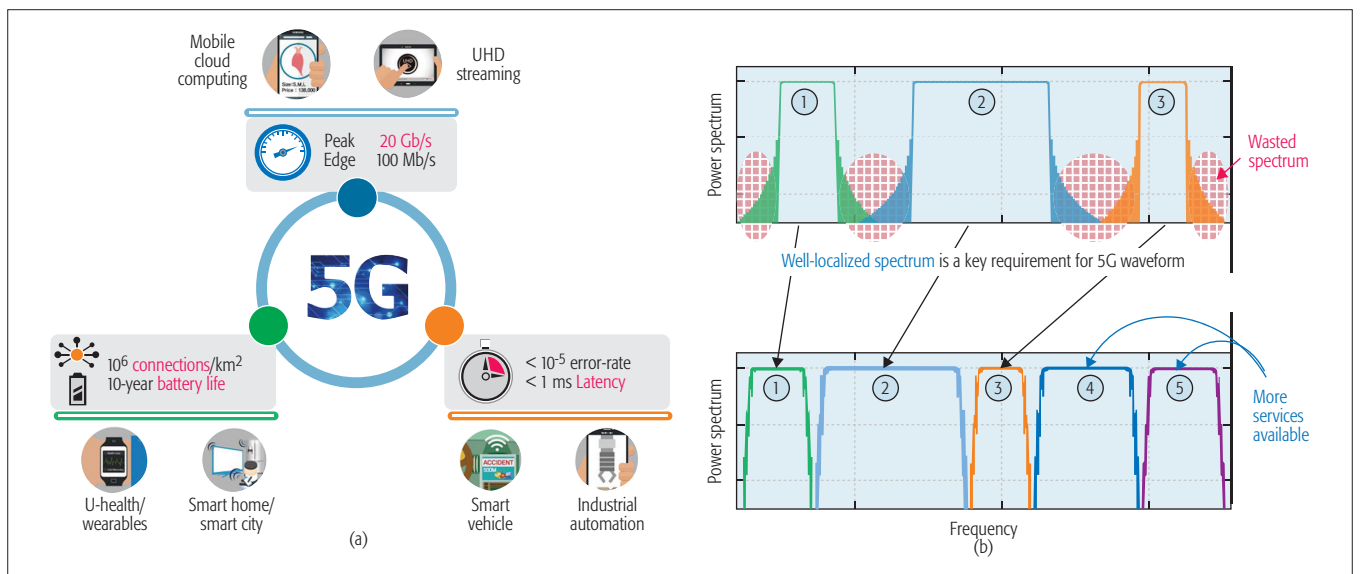


Figure 1. Frequency localization in waveform design becomes more important to meet diverse 5G requirements in spectrally efficient way: a) diverse 5G requirements; b) necessity of well-localized spectrum.

carrier spectrum [5]. Furthermore, with the rise of 5G communications, FBMC has been acknowledged as one of the post-OFDM waveform candidates that can utilize the spectrum efficiently under various environments, such as asynchronous heterogeneous networks (Het-Nets), fragmented spectrum, and multi-RAT coexistence [2]. However, not to compromise spectral efficiency, conventional FBMC systems generally double the transmission symbol rate compared to that of CP-less OFDM while adopting offset quadrature amplitude modulation (OQAM), in which in-phase and quadrature-phase symbols are modulated separately with the timing offset of half OFDM symbol duration, hence called OQAM-FBMC or staggered multitone (SMT) [3]. This causes intrinsic interference that makes it difficult to adopt low-overhead scattered pilot designs and apply trivial channel estimation algorithms. Furthermore, conventional per-tone MIMO schemes cannot be applied easily, unlike CP-OFDM systems [6].

It is important to know that the pros and cons of these multicarrier waveforms are inherent results in their design philosophies based on fundamental trade-offs among mutual orthogonality between information symbols, symbol transmission rate, and spectrum sharpness in waveform design for wireless communications. In this article, we introduce trade-off relationships among the three conditions and how these trade-offs are reflected to well-known waveforms such as CP-OFDM and FBMC. A new waveform design for 5G is proposed to achieve superior spectrum confinement and higher spectral efficiency against CP-OFDM while minimizing side-effects. The resultant waveform is called QAM-FBMC, and its practical transceiver structure is explained. From comparisons between CP-OFDM and OQAM-FBMC, it is shown that QAM-FBMC can be a good compromise solution between them from the viewpoints of frequency localization and spectral efficiency.

WAVEFORM DESIGN PRINCIPLE

FUNDAMENTAL LIMITATIONS IN WAVEFORM DESIGN

In digital communications, *waveform* means a time domain pulse to convey a modulation symbol including information data bits. Thus, we can roughly regard waveform design as pulse shaping in the time domain for easy understanding. Since it is well known that there is an inverse relationship between time domain and frequency domain descriptions, we may specify an arbitrary function of time or an arbitrary spectrum, but we cannot specify both of them together. As time domain pulse duration gets longer, its frequency bandwidth gets narrower. Furthermore, there is the uncertainty principle inequality of time-frequency dispersion product, so every waveform cannot be strictly limited in both time and frequency [7]. Since the ideally band-limited pulse is not realizable because of its infinite time duration to transmit, usually a waveform is designed as a time-limited pulse while allowing some amount of spectrum leakage in the frequency domain.

In communication theory, there is a fundamental lower bound on the bandwidth required to transmit information symbols contiguously without any intersymbol interference (ISI). The Nyquist criterion says that for baseband pulse signals with symbol transmission period T , the minimum signal bandwidth is $(1/2T)$ Hz. Since the Nyquist criterion is based on single-carrier waveforms, it can be generalized for multicarrier waveforms as follows: for $M(>1)$ baseband pulse signals with period T , the minimum signal bandwidth required for transmission without any ISI and intercarrier interference (ICI) simultaneously is $(M/2T)$ Hz. That is, the requirement that there be M orthogonal pulses in the symbol interval increases the minimum bandwidth requirement by M . This is called the generalized Nyquist criterion (GNC) [8].

The important thing is that there are many waveforms to satisfy the GNC, but they cannot satisfy the uncertainty principle equality of their time-frequency dispersion products (i.e., the best

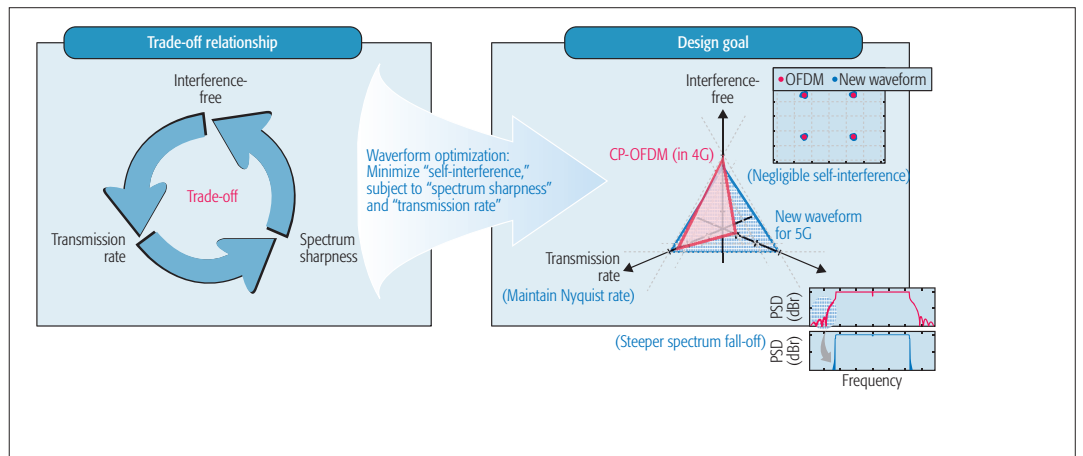


Figure 2. Waveform design principle for 5G: finding the best trade-off.

time-frequency localization property). Reversely, the Gaussian pulses can satisfy the uncertainty principle equality but they cannot satisfy the GNC. According to the Balian-Low theorem [9], there is no waveform that satisfies the following three conditions simultaneously:

1. Mutual orthogonality in the complex domain
2. Best time-frequency localization
3. Symbol density 1

Condition 1 means there are no ISI and ICI

between demodulated symbols in the ideal channel such that fading coefficient is fixed to be 1, no additive noise, and no synchronization errors. Condition 2 means the uncertainty principle equality of time-frequency dispersion product. In condition 3, the symbol density is defined as $(1/TF)$, where T is symbol transmission period and F is subcarrier spacing. It can be understood as number of complex modulation symbols available within unit time-frequency plane. Hence, the waveform design problem is how to trade these conditions off according to requirements for wireless communications.

QAM-FBMC: A NEW WAVEFORM OPTIMIZED FOR 5G

For spectral efficiency improvement, it is desirable to increase available time-frequency resources under given bandwidth as much as possible. This can be achieved by increasing symbol transmission rate in time and reducing guard-band size in frequency, but we have to consider broken mutual orthogonality between information symbols caused by their trade-off relationship. Therefore, our design principle for 5G waveform is to minimize self-interference caused by breaking the orthogonality rule under given target spectrum confinement while keeping the symbol transmission rate the same as the Nyquist rate, as shown in Fig. 2. This can be formulated mathematically and optimized numerically. For more details, see [10].

There are at least two prototype filters in QAM-FBMC. When we trade the orthogonality off frequency confinement in optimization, using a pair of two or more prototype filters that can have different coefficients from each other helps to maintain self-signal-to-interference ratio (self-SIR) of QAM-FBMC much higher than using a single prototype filter, thanks to the increased degree of freedom. A further trade-off between two prototype filters is also capable considering practical wireless communication system design issues such as channel estimation and equalization. Table 1 shows an asymmetrically optimized pair of prototype filter coefficients in frequency domain with overlapping factor $L = 4$ where the primary filter is designed to have better spectrum confinement and self-SIR than the secondary filter. By positioning pilot signals on time-frequency resources for the primary filter only, channel

Primary (4 taps)		Secondary (31 taps)			
Real	Imag	Real	Imag	Real	Imag
+1.00000		+1.00000			
-0.97196		-0.89608	+0.00163	-0.00707	-0.00106
+0.70711		+0.46096	+0.01723	-0.13109	-0.01284
-0.23515		-0.06264	-0.00799	+0.08418	+0.01578
		+0.00849	+0.00013	+0.02322	+0.00299
		+0.22399	-0.00018	+0.01099	+0.00194
		-0.44020	-0.01628	+0.04237	+0.00509
		+0.29270	+0.03426	-0.06274	-0.01173
		-0.00829	-0.00064	-0.05727	-0.00549
		-0.26687	-0.04119	-0.00791	-0.00168
		+0.15119	+0.01693	-0.00587	-0.00463
		+0.00716	+0.00415	+0.05230	+0.01420
		+0.01129	+0.00158	+0.00881	+0.00294
		+0.07419	+0.01120	+0.00898	+0.00278
		-0.12905	-0.01432	+0.01116	+0.00408
		+0.01275	-0.01138	-0.03217	-0.00812

Table 1. An example of optimized prototype filter coefficients for QAM-FBMC systems (frequency domain, one-sided).

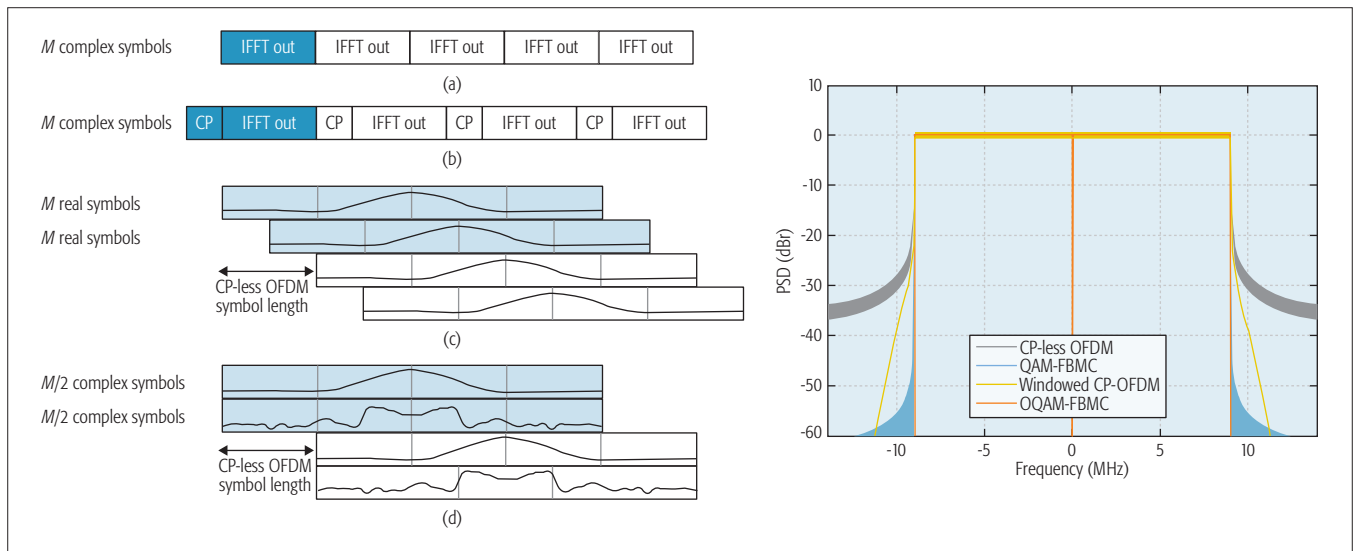


Figure 3. Comparison of time and frequency representations between well-known waveforms and QAM-FBMC: a) CP-less OFDM; b) CP-OFDM; c) OQAM-FBMC; d) QAM-FBMC.

estimation accuracy at the QAM-FBMC receiver can be increased. Later we describe more details of QAM-FBMC transceiver structure. Note that in Table 1, the power is not normalized to be 1 but is normalized relative to the value at zero frequency. Since both filters are conjugate symmetric in the frequency domain, we only noted zero and positive frequency parts, so the total number of filter coefficients is 7 and 61, respectively. Although the two filters have different numbers of taps from each other in the frequency domain, both have the same length in the time domain. Their time domain coefficients can easily be obtained by applying inverse discrete Fourier transform (IDFT) to their frequency domain coefficients.

TRADE-OFF RELATIONSHIP COMPARISON BETWEEN WELL-KNOWN WAVEFORMS

We introduce how the fundamental trade-off relationship of the three conditions is reflected to well-known waveforms such as CP-less OFDM (i.e., pure OFDM without any redundancy), CP-OFDM, and OQAM-FBMC. Their time and frequency representations are shown in Fig. 3.

First, CP-less OFDM satisfies condition 1 perfectly by using rectangular pulse (no shaping) at the cost of poor frequency localization caused by its *aliased* sinc-like spectrum (also known as Dirichlet kernel). CP-less OFDM also satisfies condition 3 in theory, but a large number of null subcarriers is required not to interfere with adjacent channels in practice.

Second, CP-OFDM does not satisfy condition 3 because time domain redundancy CP is added. However, CP is usually included to make OFDM systems robust and to enable simple per-tone channel equalization without any performance loss at the receiver in practical multipath fading channel environments. Additional pulse shaping such as windowing and filtering can also be applied within CP duration at the transmitter to reduce spectrum leakage. Let T_0 and T_{CP} be OFDM symbol duration and CP length, respectively. Then the symbol transmission period of CP-OFDM is given by $T_{CP-OFDM} = T_0 + T_{CP}$.

Since the subcarrier spacing of CP-OFDM is given by $F_{CP-OFDM} = (1/T_0)$, the symbol density is given by $(1/TF)_{CP-OFDM} = T_0/(T_0 + T_{CP})$, which is less than 1. If we assume the normal CP length in LTE (i.e., $T_{CP} = (1/14)T_0$), $(1/TF)_{CP-OFDM} \approx 0.93$. CP-OFDM also sacrifices frequency localization to retain orthogonality. Note that under practical regulation such as spectrum emission mask (SEM) or adjacent channel leakage ratio (ACLR), both CP-less OFDM and CP-OFDM require huge amounts of guardband due to slow decay of their frequency responses. This is also a cause of decreasing spectral efficiency. To reduce the guardband, additional pulse shaping methods such as filtering and windowing are used in most commercial OFDM systems. For example, the guardband-to-system-bandwidth ratio of LTE is 10 percent, so the effective symbol density of LTE after considering guardband can be calculated as $(14/15) \times (9/10) = 0.84$.

The third is OQAM-FBMC adopting the PHYDYAS prototype filter [6]. By relaxing orthogonality condition 1 to the real field only, OQAM-FBMC achieves near orthogonal transmission. This can also be represented as self-SIR in the real domain (e.g., 65 dB with the overlapping factor $L = 4$). To enable this, the complex data symbols are divided into its real and imaginary parts, and each purely real symbol has its neighbors in both time and frequency to be purely imaginary data symbols. Instead, frequency localization is excellent due to the prototype filter with length L times longer and a pulse shape designed to have faster fall-off than in CP-less OFDM. Thanks to this good confinement, almost no guardband is required to meet the SEM or ACLR requirement. Although the symbol transmission period of OQAM-FBMC is $T_{OQAM-FBMC} = (1/2F)$, since OQAM-FBMC can only achieve real-field orthogonality, its effective symbol density after considering reduced degree of freedom and almost no guardband can be approximated as 1. However, the drawback is its intrinsic interference, which is calculated in the complex domain. On the complex channel assumption, the self-SIR of OQAM-FBMC is 0 dB. In other

In FBMC systems, increased symbol length and smooth pulse shape in time can mitigate ISI and ICI under multipath fading channel environments even if CP is removed. However, to achieve enhanced performance, sophisticated receiver algorithms must be developed to cope with the residual interference.

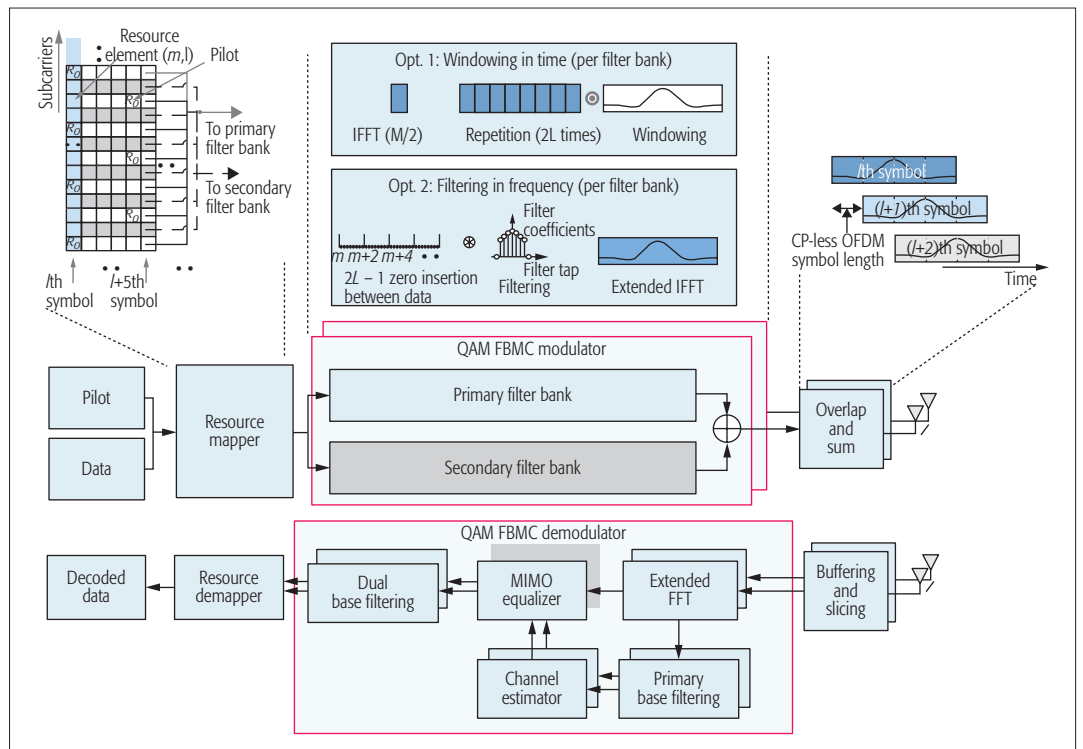


Figure 4. QAM-FBMC transceiver structure with dual filter banks.

words, OQAM-FBMC is very sensitive to residual interference after channel equalization. This implies that, unless adopting an ideal channel estimator and equalizer at the receiver that can turn the complex channel back to the real one, good performance cannot be guaranteed to be as promising as expected from the real field self-SIR of 65 dB.

OVERALL QAM-FBMC SYSTEM OPERATION

QAM-FBMC TRANSMIT STRUCTURE

As shown in Fig. 4, encoded and modulated QAM symbols are spatially multiplexed to a per-antenna FBMC modulator, where pilots are mapped on the primary filter bank. Each modulated QAM symbol is alternatively mapped on dual filter banks. When an OFDM symbol consists of M QAM symbols, and dual filter banks are used for QAM-FBMC systems, $(M/2)$ QAM symbols can be allocated to each filter bank on the l th symbol. Each filter bank can be implemented with at least two options; one is implemented in the time domain by windowing, and the other is implemented in the frequency domain by circular convolution. In the first option, the time domain data, that is, the output of an $(M/2)$ -point inverse fast Fourier transform (IFFT), is repeated $2L$ times, where L is the overlapping factor of FBMC. The repeated time domain data can be pulse modulated by the time domain window with point-wise multiplication. Note that since there are two filter banks in a QAM-FBMC system, this window operation should be performed per filter bank and then added. In the frequency domain approach, oversampling and filtering can be done by zero padding and its circular convolution with frequency domain filter coefficients. Also, in this case, circular

convolution should be performed per filter bank and then added. The oversampled and filtered data in the frequency domain are converted in the time domain with extended IFFT ($N = LM$ -points) as in the second option. The l th QAM-FBMC modulated symbol is given by summation of the outputs from dual filter banks. Since generic QAM-FBMC with $L \geq 2$ has symbol duration $N = LM$, L times longer than M , contiguous QAM-FBMC symbols partially overlap one another with M -sample shifting as shown in Fig. 4. Then the symbol transmission rate is maintained as the same as in CP-less OFDM. Due to the time domain overlap, there can be ISI. However, this ISI can be minimized through filter optimization considering GNC as explained in [10]. In terms of implementation complexity, the time domain process with data size IFFT, repetition, and windowing is preferred at the transmitter side.

There are similarities between other waveforms and QAM-FBMC as follows.

OQAM-FBMC: In the OQAM-FBMC transmitter, there is only one filter bank based on the so-called prototype filter, which is implemented by means of M -IFFT and a poly-phase network (PPN). Since a purely real or imaginary symbol is mapped to each resource element, the overlapping period is set to be not M but $(M/2)$ to keep the transmission symbol rate the same as in CP-less OFDM. Hence, one OQAM-FBMC symbol is overlapped with the former $(2L - 1)$ and latter $(2L - 1)$ contiguous symbols.

CP-Less OFDM: CP-less OFDM is a special case of QAM-FBMC that can be obtained by setting one filter bank and overlapping factor as 1 with a rectangular time domain pulse. However, QAM-FBMC with a single filter bank dif-

fers from OFDM in that non-rectangular pulse is applied to each subcarrier, which makes one QAM-FBMC symbol duration L times longer than one OFDM symbol duration of M . Thus, one simple way to make the same number of frequency/time resource elements convey QAM data symbols in a unit time as in CP-less OFDM is to overlap the former ($L - 1$) and the latter ($L - 1$) contiguous symbols.

Windowed CP-OFDM: In windowed CP-OFDM, a smooth window rather than a rectangular pulse has been considered to reduce spectrum leakage by allowing redundancy to be the same as CP length. Normally, CP length has been decided by considering both multipath delay spread for given deployment scenarios and window length. For windowing, postfix as long as window length is required in addition to prefix. Compared to QAM-FBMC transmitter implementation in time domain with windowing, the prefix and postfix can be interpreted as fractional repetition. Since only one filter bank is used for windowed CP-OFDM, M -point IFFT is considered. After M -point IFFT, each windowed OFDM symbol can be generated with fractional repetition and windowing. To keep the same redundancy as CP-OFDM, each windowed CP-OFDM symbol is overlapped with previous and next windowed CP-OFDM symbols as same as half of the window length. This overlap is similar to the overlap and sum in Fig. 4 with a different symbol rate, that is, the amount of shift in the overlap and sum, where the symbol transmission period is given by the symbol length as long as $M + N_{CP}$, where N_{CP} denotes CP length in terms of number of QAM symbols. It is different from QAM-FBMC, which has a symbol rate the same as the CP-less OFDM symbol length ($= M$). Therefore, windowed CP-OFDM data can be generated with M -point IFFT, fractional repetition, windowing, and the overlap and sum transmission with symbol transmission period as $M + N_{CP}$.

QAM-FBMC RECEIVER STRUCTURE

In FBMC systems, increased symbol length and smooth pulse shape in time can mitigate ISI and ICI under multipath fading channel environments even if CP is removed. However, to achieve enhanced performance, sophisticated receiver algorithms must be developed to cope with the residual interference. By considering the trade-off between performance and complexity, per-tone MIMO equalization in the oversampled frequency domain is introduced. We designed a receiver for both single-input single-output (SISO) and MIMO QAM-FBMC systems including channel estimation, equalization, and a soft demapper that work on the residual interference awareness by well exploiting the filter property and the overlap and sum transmission structures. The overview of our receiver structure is summarized as follows.

FFT in the Oversampled Domain: To reduce the frequency selective fading effect, equalization before filtering (i.e., oversampled domain equalization) is considered. This requires bigger size (N -point) FFT in the oversampled domain per receiving antenna as a part of the reverse procedure of filter bank implementation option

2 in Fig. 4. The reverse procedure of filtering in the frequency domain can be implemented with extended FFT, equalization, and filter-coefficient-weighted despreading in fading channels.

Channel Estimation Based on the Primary Filtered Pilots: In QAM-FBMC, any channel estimation can be applied in similar but slightly different ways as in the conventional OFDM. Since transmit/receive filtering processes are included in QAM-FBMC modulation/demodulation as shown in Fig. 4, and they help to maintain orthogonalities among time-frequency resource elements, least square (LS) channel estimation at the pilot position is done after receive filtering to obtain less contaminated pilot symbols. Since only receive filtering can manage interference at the channel estimation stage, the preferable pilot position will be the subcarriers in which a base filter with well confined spectrum and higher self-SIR (i.e., primary base filter) is allocated for better channel estimation. This can also reduce filtering complexity by utilizing one base filter in channel estimation. If DFT-based channel estimation is considered, IFFT with reduced size as (M/D_p) can be used to get channel impulse response (CIR), where D_p denotes the distance between pilots in the frequency domain. Based on the fact that the length of CIR is much less than FFT size and significant taps in CIR are sparse [11], denoising can be considered, that is, noise nulling dependent on maximum channel delay and noise reduction by thresholding based on estimated noise variance within the noise nulling window. For frequency domain channel interpolation, a bigger FFT is used. For oversampled domain equalization, N -point channel frequency response is obtained from the output of channel interpolation by N -point FFT. For MIMO systems, the same channel estimation method as in SISO systems can be used. To distinguish multiple transmit antennas, pilot positions are exclusively defined in the frequency or time domain for other transmit antennas similar to the cell-specific reference signal in LTE systems.

Per-Tone MIMO Equalization in the Oversampled Domain and Filtering: Separate channel equalization and filtering are considered due to the trade-off between complexity and performance. In addition, per-tone equalization is performed with only diagonal terms of the channel frequency response of the desired symbol. Even if per-tone equalization is used, the equalization in the oversampled domain before filtering can reduce channel induced ICI and ISI by equalization with higher resolution. Filtering after equalization can achieve the designed signal-to-interference-plus-noise ratios (SINRs) of the original filters in frequency selective channels, since filtering can be interpreted as the weighted combination of equalized samples in the L times oversampled frequency domain. However, the weighted combination of received signals can be affected more by frequency selectivity in channels when the number of filter taps increases.

Soft Demapper with Residual Interference Consideration: Even if the interantenna interference (IAI) can be suppressed with equalization and filtering, residual interference still remains. In the soft demapper, that is, log like-

For MIMO systems, the same channel estimation method as in single-input single-output (SISO) systems can be used. To distinguish multiple transmit antenna, pilot positions are exclusively defined in frequency or time domain for other transmit antenna similar to cell specific reference signal in LTE systems.

Separate channel equalization and filtering are considered due to the trade-off between complexity and performance. In addition, per-tone equalization is performed with only diagonal terms of the channel frequency response of the desired symbol.

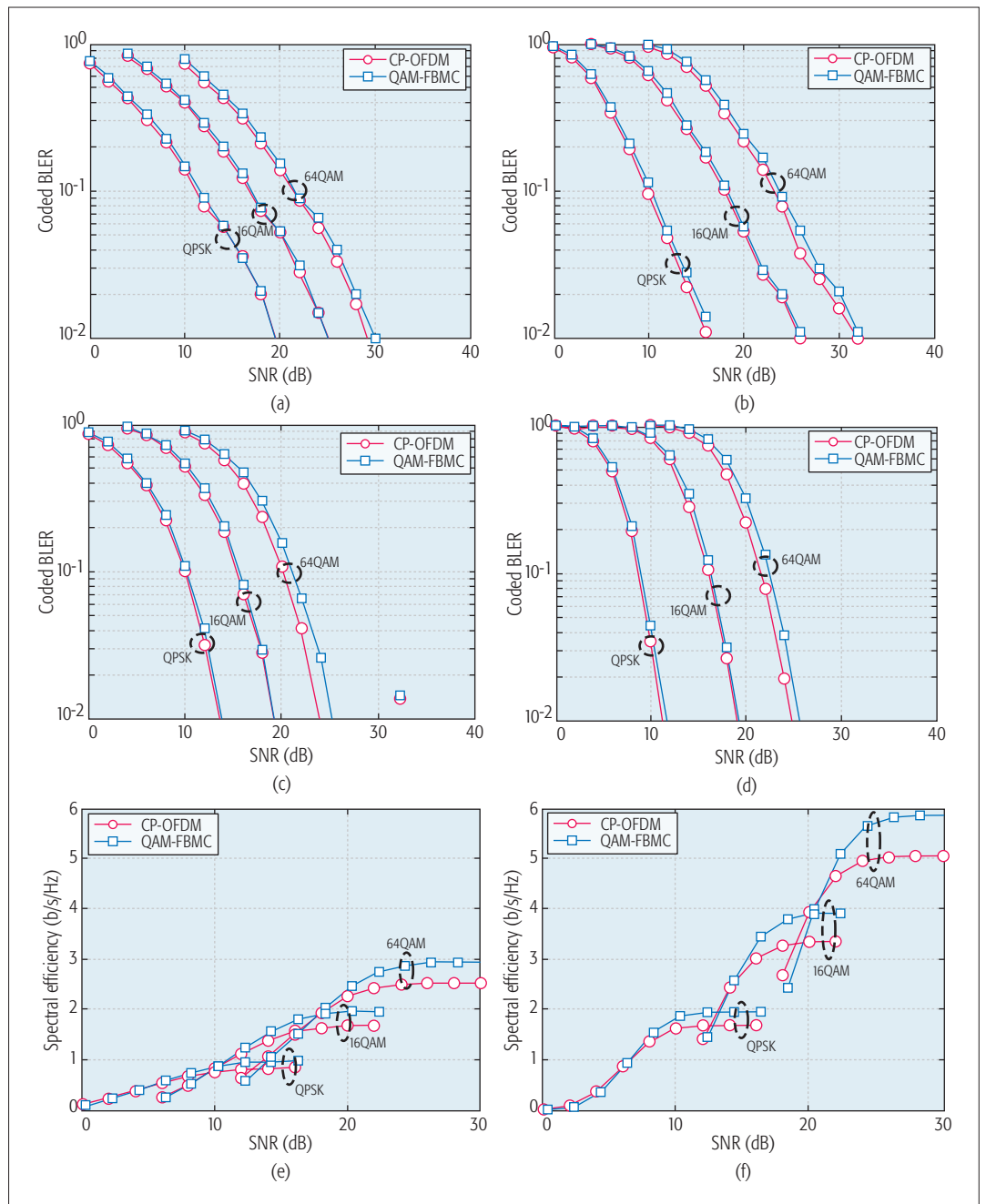


Figure 5. Performance comparison between CP-OFDM and QAM-FBMC: a) coded BLER in SISO EPA channel; b) coded BLER in 2×2 EPA channel; c) coded BLER in SISO ETU channel; d) coded BLER in 2×2 ETU channel; e) spectral efficiency in SISO ETU channel; f) spectral efficiency in 2×2 ETU channel.

likelihood ratio (LLR) calculation, the interference distribution and power of interference can be considered. As the simplest method, a Gaussian soft demapper can be taken into account with SINR recalculation after equalization and filtering with assumption on the distribution of residual IAI (i.e., Gaussian distribution). Further complexity reduction can be achieved by max-log approximation in the LLR calculation. The approximated LLR can be calculated from the distances between an equalized and filtered signal and all possible constellation points. In a single-antenna system, the same LLR calculation is done except for the effective SINR calculation part.

LINK PERFORMANCE EVALUATION RESULTS FOR QAM-FBMC

In order to evaluate the link performance of the proposed QAM-FBMC system, we used the pair of two prototype filters described in Table 1. We evaluated coded block error rate (BLER) performances of CP-OFDM and QAM-FBMC in conditions similar to LTE. For CP-OFDM, CP length is set to be (1/14) of symbol length based on LTE normal CP mode. For pilot-based channel estimation, we used LS and DFT-based interpolation with denoising. The scattered pilot pattern is set to be rectangular and equidistance with a 4-subcarrier grid in the frequency domain

and a 4-symbol grid in the time domain, respectively.

In Figs. 5a–5d, we plotted BLER performance comparison results between CP-OFDM and QAM-FBMC according to modulation levels with rate 0.5 LTE turbo code in SISO EPA, 2×2 MIMO EPA, SISO ETU, and 2×2 MIMO ETU channels, respectively. Even in MIMO environments as well as in SISO, we can see comparable performance of QAM-FBMC and CP-OFDM. Note that if more sophisticated equalization and soft demapping algorithms considering residual interference are used, especially in interference-limited environments, which could not be described in detail, the performance can be further improved [12].

As shown in Figs. 5e and 5f, QAM-FBMC has spectral efficiency gain against CP-OFDM, which comes from guard-band reduction within channel bandwidth in the frequency domain and CP-less transmission in the time domain, while coded BLER performance is kept the same as in CP-OFDM. For the spectral efficiency calculation in CP-OFDM, the guardband is set to be 10 percent of channel bandwidth and the CP ratio is set to be $(1/14)$; these configurations are being used for LTE with normal CP. We also use same total transmit power for both systems for fair comparison, which means that per-tone SNR in QAM-FBMC is less than that in CP-OFDM due to the increased transmission bandwidth. The resultant spectral efficiency gain of QAM-FBMC over CP-OFDM is about 16.4 percent $(= (0.978 \times 1)/(0.9 \times (14/15)) \times 100)$. Note that this gain changes according to the portion of frequency and time overheads in CP-OFDM.

CONCLUSION

We have introduced a waveform design principle for 5G and proposed a new waveform called QAM-FBMC that could radically reduce the inherent overheads in CP-OFDM. A new optimization method for a pair of base filters design and overall transceiver structures are introduced to support the conventional QAM in FBMC. The performance evaluation results show that the proposed QAM-FBMC system provides much lower spectrum leakage as well as enhanced spectral efficiency against the CP-OFDM with comparable link performance. Therefore, QAM-FBMC can be a good waveform solution for 5G mobile communications.

REFERENCES

- [1] 3GPP R1-162141, "Draft Agenda," RAN WG1 Meeting #84bis, Apr. 2016.
- [2] G. Wunder *et al.*, "5G-NOW: Non-Orthogonal, Asynchronous Waveforms for Future Mobile Applications," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 97–105.
- [3] B. Farhang-Boroujeny, "OFDM versus Filter Bank Multicarrier," *IEEE Signal Processing Mag.*, vol. 28, no. 3, May 2011, pp. 92–112.
- [4] F. Schaich, "Filterbank Based Multi Carrier Transmission (FBMC) – Evolving OFDM: FBMC in the Context of WiMAX," *Proc. Euro. Wireless Conf.*, Apr. 2010, pp. 1051–58.
- [5] PHYDYAS: <http://www.ict-phydyas.org/>
- [6] M. Bellanger *et al.*, "FBMC Physical Layer: A Primer," PHYDYAS, tech. rep., June 2010.
- [7] R. Haas and J.-C. Belfiore, "A Time-Frequency Well-Localized Pulse for Multiple Carrier Transmission," *Wireless Personal Commun.*, vol. 5, no. 1, 1997, pp. 1–18.
- [8] J. R. Barry, E. A. Lee, and D. G. Messerschmitt, *Digital Communication*, 3rd ed., Springer, 2004.
- [9] J. J. Benedetto, C. Heil, and D. F. Walnut, "Gabor Systems and the Balian-Low Theorem," *Gabor Analysis and Algorithms*, 1998, pp. 85–122.
- [10] Y. H. Yun *et al.*, "A New Waveform Enabling Enhanced QAM-FBMC Systems," *Proc. Int'l. Wksp. Signal Processing Adv. Wireless Commun.*, June 2015, pp. 116–20.
- [11] H. Minn and V. K. Bhargava, "An Investigation into Time-Domain Approach for OFDM Channel Estimation," *IEEE Trans. Broadcast.*, vol. 46, no. 4, Dec. 2000, pp. 240–48.
- [12] K. Kim *et al.*, "Pre-Processing Based Soft-Demapper for Per-Tone MIMO Operation in QAM-FBMC Systems," *Proc. IEEE Int'l. Symp. Personal Indoor and Mobile Radio Commun.*, Sept. 2015, pp. 609–13.

BIOGRAPHIES

CHANHONG KIM (ch12.kim@samsung.com) received his B.Sc. and Ph.D. degrees in electrical engineering from Seoul National University, Korea, in 2004 and 2011, respectively. After one year of postdoctoral research in 2011, he currently works as a senior engineer for Samsung Electronics Co., Ltd., Suwon, Korea. His research interests include multicarrier waveform, multiple access, and MIMO/beamforming for mobile communications, with current emphasis on 5G communication system design.

YEO HUN YUN (yeohunss.yun@samsung.com) received his B.Sc. degree (summa cum laude) in information, communication, and electronics engineering from Catholic University, Korea, in 2006. He received his M.Sc. degree in computer and communications engineering and Ph.D. degree in electrical engineering from Pohang University of Science and Technology, Korea, in 2009 and 2013, respectively. Since October 2013, he has been with Samsung Electronics Co., Ltd. His current research interests include multicarrier waveform and 5G mobile communication systems.

KYEONGYEON KIM [S'06, M'10] (kyeongyeon.kim@samsung.com) received her B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Yonsei University, Seoul, Korea, in 2001, 2003, and 2007, respectively. After graduation, she was a postdoctoral fellow at Purdue University from 2007 to 2008 and at the University of Illinois at Urbana-Champaign from 2008 to 2010. In 2010, she joined Samsung Electronics Co. Ltd., and has contributed to research on next-generation wireless system design and 5G standardization. Her research interests include statistical signal processing, wireless/underwater communication and broadcasting systems, adaptive filtering, and machine learning.

Ji-YUN SEOL (jijun.seol@samsung.com) received his B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Seoul National University in 1997, 1999, and 2005, respectively. He has been with Samsung Electronics Co., Ltd. since 2004. He has years of experience in development of modem algorithms and standardization for mobile WiMAX and 3GPP LTE. He is currently a director and head of the Advanced Communications Lab., Communications Research Group at Samsung Electronics. He has been in charge of research for the next generation (B4G/5G) mobile communications since 2011. His current fields of interest include research/development of next generation mobile communication systems and advanced PHY algorithms.

The performance evaluation results show that the proposed QAM-FBMC system provides much lower spectrum leakage as well as enhanced spectral efficiency against the CP-OFDM with comparable link performance. Therefore, QAM-FBMC can be a good waveform solution for 5G mobile communications.

On the Waveform for 5G

Xi Zhang, Lei Chen, Jing Qiu, and Javad Abdoli

The authors present an overview and in-depth analysis of the most discussed 5G waveform candidates. In addition to general requirements, the nature of each waveform is revealed including the motivation, the underlying methodology, and the associated advantages and disadvantages. These waveform candidates are categorized and compared both qualitatively and quantitatively. By doing all these, the study in this work offers not only design guidelines but also operational suggestions for 5G waveform.

ABSTRACT

In this article, an overview and an in-depth analysis of the most discussed 5G waveform candidates are presented. In addition to general requirements, the nature of each waveform is revealed including the motivation, the underlying methodology, and the associated advantages and disadvantages. Furthermore, these waveform candidates are categorized and compared both qualitatively and quantitatively. By doing all these, the study in this work offers not only design guidelines but also operational suggestions for the 5G waveform.

INTRODUCTION

In each generation of wireless communication technologies, the underlying waveform has always served as a cornerstone and a shaping component. As a recent example, in the fourth generation (4G) Long Term Evolution (LTE) networks, cyclically prefixed orthogonal frequency-division multiplexing (CP-OFDM) and its low peak-to-average-power ratio (PAPR) variant, discrete Fourier transform spread OFDM (DFT-s-OFDM) have provided efficient support for the widely popularized mobile broadband (MBB) service, enabling the commercial success of 4G LTE networks. Nevertheless, it should be noted that, mainly targeting MBB service, the network functionality of 4G LTE networks is still relatively simple.

After years of discussion [1], the expectations for 5G have been made clear by the International Telecommunication Union [2], and standards activities have also been initiated recently [3]. In general, three different types of services are to be supported in 5G: enhanced MBB (eMBB), massive machine-type communication (mMTC), and ultra-reliable and low-latency communication (URLLC), featuring 20 Gb/s peak data rate, $10^6/\text{km}^2$ device density, and less than 1 ms latency, respectively. To meet these goals, enormous research challenges have been placed in front of wireless engineers, the very first of which is the largely increased diversity in the services to be provided in 5G, together with the desired forward compatibility for future applications.

For eMBB services, a broad range of deployment scenarios have been identified, including indoor hotspot, urban macro, dense urban, rural, and high speed, some of which may be reused for the mMTC and URLLC services [3]. It is expected that each service and deployment scenario will prefer different physical signaling formats

for achieving the best possible performance. For instance, a smaller subcarrier spacing would help to support the massive connections required for mMTC, while a larger one with shorter symbol duration may help to reduce latency for URLLC services. In order to support all these services and the associated deployment scenarios efficiently, a flexible air interface framework would be required. Therefore, the waveform, as the shaping component of any air interface, has to be designed carefully to enable such flexibility.

In recent years, several waveforms have been proposed and discussed among academia and industry [4]. With the investigation going deeper, some of the proposed waveforms are ruled out quickly due to their inefficiency in practical systems, while others are still under active discussion. In this article, an overview and in-depth analysis of the most discussed 5G waveform candidates are provided. Specifically, the nature of each waveform is revealed including the motivation, the underlying methodology, and the associated advantages and disadvantages. Furthermore, these waveforms are categorized and compared both qualitatively and quantitatively. By doing all of this, the investigation in this work provides not only design guidelines but also operational suggestions for the 5G waveform.

The waveforms discussed in this article are classified into two categories: legacy OFDM transceivers supported and new transceivers required. The waveforms in the first category can be considered as straightforward but nontrivial enhancements of CP-OFDM, including windowed OFDM (W-OFDM) [5], universal filtered multi-carrier (UFMC) [6], and filtered OFDM (f-OFDM) [7–9]. The second category utilizes new transceiver structures that are different from CP-OFDM. The waveforms in this category include filter-band multi-carrier (FBMC) [10, 11] and generalized frequency-division multiplexing (GFDM) [12]. Detailed analysis, comparison, and evaluation of these waveforms are relegated to subsequent sections.

REQUIREMENT ANALYSIS

CP-OFDM offers several attractive properties such as simple channel estimation, low-complexity equalization, efficient hardware implementation, easy combination with multiple-input multiple-output (MIMO) transmission, and backward compatibility to 4G LTE. With these outstanding benefits, CP-OFDM will remain an important candidate for the 5G waveform.

On the other hand, in pursuit of orthogonality, a single and uniform parameter configuration was required by LTE CP-OFDM and its low-PAPR variant, DFT-s-OFDM. For this reason, a one-size-fits-all approach was adopted in LTE. To be specific, a fixed set of waveform parameters (a.k.a. numerology), including subcarrier spacing and length of cyclic prefix (CP), has been uniformly applied across the entire system bandwidth. While easy to implement, the main issue of such a fixed design is lack of flexibility to support mixed services with different waveform parameters within one carrier, which, as mentioned earlier, is a key requirement of the 5G air interface design.

Moreover, CP-OFDM and DFT-s-OFDM also suffer from high spectral side lobes resulting from the Sinc pulse shape in the frequency domain, which is equivalent to a rectangular pulse shape in the time domain, leading to the following issues.

Poor Spectrum Utilization: Large frequency guard bands are needed at both edges of the system bandwidth for the signal to reach enough attenuation to meet the requirements of spectrum mask and adjacent channel leakage ratio (ACLR). This implies a significant loss in spectrum efficiency. For instance, 10 percent of the system bandwidth was reserved as guard bands in LTE.

Stringent Synchronization Requirement: To achieve interference-free reception, the uplink signals from different users should arrive at the base station at the same time. Any timing misalignment larger than the CP duration, especially between the signals closely located in the frequency domain, would lead to severe inter-user interference and inter-carrier interference (ICI), and hence large performance degradations. Therefore, in LTE networks, frequent timing advance (TA) signals are transmitted from the base station to mobile users to maintain the orthogonality in reception.

To overcome the above-mentioned shortcomings, a new waveform should be considered for 5G. On one hand, the waveform candidate should inherit all the advantages of CP-OFDM. On the other hand, the new waveform should be able to provide the flexibility required for efficient support of diverse services and deployment scenarios in one carrier. Furthermore, the new waveform needs to have sufficiently good spectrum confinement, yielding higher spectrum efficiency. The following design criteria are essential in selecting and designing the 5G waveform.

High Flexibility:

Multi-Numerology Coexistence: The 5G waveform should be able to support flexibly configured numerology and multi-numerology coexistence, to enable tailored services for different applications in the associated deployment scenarios. To reduce scheduling constraints, frequency domain multiplexing of different numerologies is of interest. In addition, for smooth evolution and efficient resource utilization, dynamic allocation of bandwidth for different numerologies/services should also be supported.

Bandwidth Extension: The 5G waveform should enable straightforward and scalable

extension to support ultra-wideband operations (e.g., ≥ 100 MHz), with affordable complexity.

High Spectrum Efficiency:

Frequency-Domain Localization: Sufficient spectral confinement is essential for achieving high spectrum efficiency. It is also highly related to the guard band overhead for supporting neighboring but non-orthogonal transmissions.

Time-Domain Localization: To reduce the time-domain overhead due to tails of filtering or pulse shaping, and thus achieve a high spectrum efficiency, a time-domain localized waveform is preferred.

MIMO Friendliness: MIMO, especially massive MIMO, as one of the most recognized 5G technical components, has to be supported by the new waveform, without requiring much additional effort, in either open- or closed-loop operations.

High-Order Modulation: To boost the peak spectrum efficiency, high-order modulations, such as 256-quadrature amplitude modulation (QAM), are likely to be used in a wider range of scenarios in 5G. The new waveform should be able to support such high-order modulations.

High Symmetry: Having a symmetric waveform would be beneficial for the unification of access, sidelink, and backhauling functions. Although the design principles for the downlink (DL) and uplink (UL) might be different (e.g., UL transmission would be subject to stricter PAPR restrictions due to the nonlinearity of the power amplifiers), the waveforms for DL and UL are still preferred to be as similar as possible to facilitate interference cancellation. To form the desired symmetry while alleviating PAPR restrictions, PAPR reduction schemes that are transparent to standards (e.g., clipping, companding) can be applied if coverage or power efficiency becomes of greater relevance.

QUALITATIVE ANALYSIS

As mentioned before, this article focuses on the most discussed and promising candidates for the 5G waveform, that is, the variants of CP-OFDM, including W-OFDM, UFMC, and f-OFDM, and also FBMC and GFDM. A common motivation for these waveforms is to reduce the out-of-band emission (OOBE), thus reducing guard band overhead and improving spectrum utilization. In this section, we provide an overview of the mentioned waveforms, starting from the legacy OFDM transceiver supporting candidates (i.e., W-OFDM, UFMC, and f-OFDM) and then proceeding to the new transceiver required candidates (i.e., FBMC and GFDM). Along with this categorization, the most important and fundamental aspects of the shortlisted candidates and their implications for the 5G waveform are discussed.

LEGACY OFDM TRANSCEIVER SUPPORTABLE

To maintain the benefits of CP-OFDM while reducing the OOBE, straightforward enhancements can be made. Example approaches include windowing to smooth the time-domain symbol transitions and filtering to suppress spectrum leakage, both of which are simple but effective.

To maintain the benefits of CP-OFDM while reducing the OOBE, straightforward enhancements can be made. Example approaches include: windowing to smooth the time-domain symbol transitions, and filtering to suppress spectrum leakage, both of which are simple but effective.

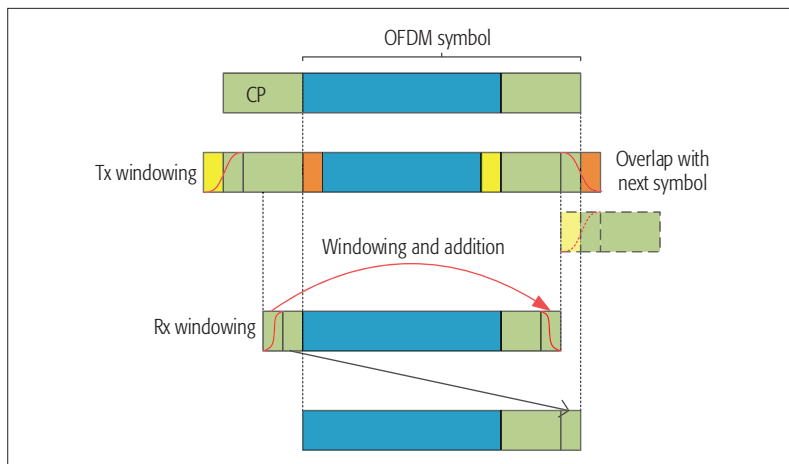


Figure 1. Illustration of windowing.

W-OFDM [5], UFMC [6], and f-OFDM [7] are the most discussed waveforms along this line of study. In this class of waveforms, the modulation process may largely reuse that of CP-OFDM.

A large part of the OOB of CP-OFDM emanates from the discontinuity between adjacent symbols. A natural and straightforward way to reduce the OOB is to introduce additional windowing to both edges of the generated symbols, and thus smooth the transition between adjacent symbols [5]. More specifically, the windowing operation can be performed on both the original cyclic prefix and the newly added guard interval (i.e., the extended part of the cyclic prefixes and the newly added cyclic suffixes). The windowed parts could overlap with each other to suppress/eliminate the time-domain overhead resulting from windowing, as illustrated in Fig. 1.

The most obvious advantage of W-OFDM is its simplicity, as only low-complexity time-domain multiplications and additions are involved. However, the largest disadvantage of W-OFDM is also due to its simplicity. That is, with a simple windowing operation, only limited OOB improvement can be achieved, and non-negligible guard bands are still needed at both edges of the system bandwidth. In addition, the OOB improvement of W-OFDM often comes with reduced effective CP length (i.e., a smaller portion of CP is indeed the cyclic extension of an OFDM symbol), losing the efficiency in multi-path channels. Furthermore, as illustrated in Fig. 1, to suppress the interference from adjacent non-orthogonal signals (e.g., asynchronous transmission), advanced receiver windowing can also be applied, which will further reduce the length of effective CP. As a straightforward enhancement to CP-OFDM, W-OFDM has been discussed for a long time, and variants have been proposed for 5G [5].

An alternative solution to reduce the OOB of CP-OFDM is to introduce subband filtering, as shown in Fig. 2. More specifically, the system bandwidth is divided into several subbands, and a conventional OFDM signal (or any other type of waveform) is contained in each subband. Different numerologies may be applied in different subbands to suit the needs of different services. Filtering operation is performed on each sub-

band to reduce its OOB and facilitate its coexistence with the adjacent subbands. With reduced/removed guard band, this mixed numerology feature can enable multi-service coexistence with improved spectrum efficiency, as compared with the traditional guard band based fragmented carrier operation. In addition, with reduced OOB, improved tolerance to interference from asynchronous transmissions can be provided, with which the requirement on synchronization can be relaxed. UFMC [6] and f-OFDM [7] are the two most discussed waveforms along this line.

For UFMC, the bandwidth of each subband is kept fixed (e.g., a resource block (RB) in 4G LTE, i.e., 180 kHz). In other words, UFMC performs filtering with fixed frequency-domain granularity. In order to prevent any inter-symbol interference (ISI) due to the filtering operation, zero prefix (ZP) is typically used in UFMC, instead of CP. The typical filter length of UFMC is less than or equal to the length of ZP. In this way, the filter tails extend into the ZPs without overlapping with each other.

As one may expect, the newly introduced subband filtering in UFMC leads to a reduced OOB and the filter tails do not pose any problem, as they are contained within the ZPs. Nevertheless, the OOB improvement of UFMC over CP-OFDM is quite limited, due to the relatively short filter length. Moreover, with fixed filtering granularity, the implementation complexity increases quickly with increasing the operating bandwidth, at both the transmitter and the receiver. Without CP and circular convolution, to account for the linear convolution of the signal with the channel impulse response, zero-padding and a double-sized fast Fourier transform (FFT) is typically used at the receiver, which further increases the complexity [6].

In contrast to UFMC, f-OFDM maintains the CP used in CP-OFDM. Moreover, the filtering granularity in f-OFDM can be configured to any value larger than the bandwidth of a physical RB in 4G LTE (i.e., 180 kHz). Filters with arbitrary bandwidth can be generated in a systematic manner, based on soft truncation of Sinc filters. More details on such a filter design which can support flexible filtering granularity can be found in [9]. A short comparison of different filters can also be found in [8]. Furthermore, the filter length in f-OFDM is not limited to the length of CP.

The main benefit of a customizable filtering granularity is a reasonably low complexity for wide-band operations. Moreover, with a relatively long filter length, the OOB can be suppressed to a desirable level, leading to an improved spectrum efficiency. The ISI incurred by filtering can also be confined into a very limited extent for two main reasons: First, the chosen filter has only a limited time-domain energy spread. Second, the filter's frequency response is flat over the entire subband and only a few edge subcarriers in the subband are affected by filtering and may contribute to the time-domain stretch of each f-OFDM symbol. Subband-wise filtering operation can be transparent to spec, and what needs to be indicated is more about subband bandwidth configuration, with which flexible multiplexing of different numerologies and services can be supported. The main draw-

back of f-OFDM, compared with W-OFDM, is its relatively high complexity due to the filtering operation.

NEW TRANSCIVER REQUIRED

In CP-OFDM, a rectangular time-domain pulse shape is used, which is equivalent to a Sinc function in the frequency domain with large side lobes, resulting in high OOB. For this reason, researchers have been looking for waveforms that support variable and customizable pulse shaping (instead of a simple rectangular window) to achieve a better trade-off between time-domain and frequency-domain localization. FBMC [10, 11] and GFDM [12] are the most mentioned representatives along this line of study. In this category of waveforms, pulse shaping/filtering is performed at the subcarrier level, and the modulator is generally more complicated than a single inverse FFT used in CP-OFDM.

By extending the pulse duration in the time domain to multiple times of a symbol duration and using properly designed pulse shaping filters, excellent frequency domain localization can be provided by FBMC [10]. In the meantime, to maintain Nyquist-rate transmission, adjacent FBMC symbols are overlapped with each other in the time domain. To minimize the interference among them, staggering of in-phase and quadrature-phase symbols in both time and frequency is applied, which is widely referred to as offset QAM (OQAM). The intrinsic/imaginary interference between symbols in the time-frequency grid is suppressed by the taking-real-part operation at the receiver side. In other words, unlike CP-OFDM, FBMC-OQAM is orthogonal in the real field only, not in the complex field.

With subcarrier filtering or pulse shaping, several desirable properties can indeed be achieved by FBMC-OQAM. With its excellent frequency localization, FBMC-OQAM requires no more than one subcarrier as guard band for neighboring but non-orthogonal transmissions. In addition, the edge-smoothed pulse shape in the time domain makes it less sensitive to multi-path spreading, which reduces the necessity for cyclic prefixing. The savings on both guard band and CP overhead enables FBMC-OQAM to achieve high spectrum efficiency. Even more, the filter bank structure of FBMC-OQAM enables flexible pulse shape adaptation for different channel conditions.

On the other hand, the disadvantages of FBMC-OQAM are also obvious. Note that the communication channels are inherently complex. For this reason, with its real-field orthogonality only, the application of FBMC-OQAM in practical systems is quite limited. The very first challenge for FBMC-OQAM is pilot design. More complicated pilot schemes such as data-dependent auxiliary pilots are often needed to take care of the interference from the complex field. As mentioned before, the 5G waveform will have to coexist with other recognized techniques (e.g., MIMO transmission). In that case, with a complex-field precoding operation at the transmitter side, the equivalent channel in the frequency domain becomes discontinuous when using different precoders on neighboring subcarriers/sym-

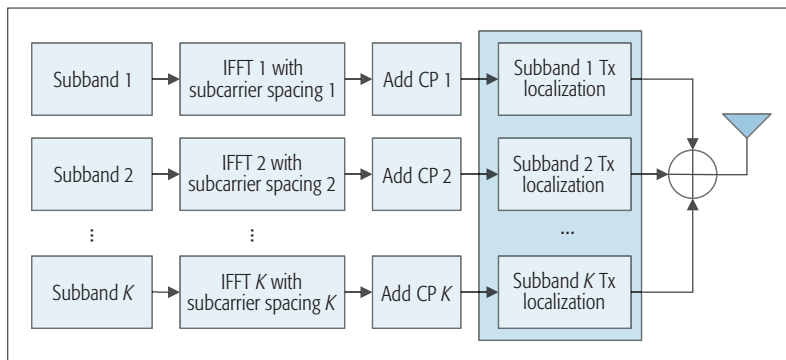


Figure 2. Illustration of subband-wise filtering.

bols. This makes it a challenging task to suppress interference with a simple equalization scheme at the receiver side. Furthermore, with extended pulse shaping, long filter tails appear at both ends of the signal burst, which causes extra overhead and must be taken care of properly.

To avoid the above-mentioned limitations of FBMC-OQAM, efforts have been made to reduce the intrinsic interference in complex field and thus enable QAM transmission. As such, the associated waveform is generally referred to as FBMC-QAM [11]. The basic idea in FBMC-QAM is to apply different filter banks for different subcarriers (e.g., even and odd), offering additional freedom to suppress the interference between adjacent symbols in both the frequency and time domains. In this case, if the signal-to-interference ratio (SIR) can be made sufficiently high, pilot design and MIMO transmission can be as simple as that in CP-OFDM systems. However, as shown in the next section, the SIRs provided by the pulse shaping filters proposed for FBMC-QAM are still not high enough to support 64-QAM transmission without a large gap from CP-OFDM. In addition, the filter tail issue still exists with FBMC-QAM.

Similar to FBMC, GFDM [12] also applies subcarrier filtering for pulse shaping, but in a block-wise and circular manner to avoid inter-burst tails. In addition, block-wise CP is added to eliminate inter-block interference in multi-path channels. In contrast to FBMC-OQAM/QAM, which still pursues orthogonality in the real/complex field, GFDM is generally non-orthogonal. Therefore, the same potential issues of pilot design and MIMO transmission as in FBMC-OQAM/QAM also exist in GFDM. Additionally, the decoding latency becomes a challenge for GFDM, as the detection process can only be started after the entire block is received. The other problem of GFDM lies in its high receiver complexity, as a result of large-sized FFT and the successive interference cancellation (SIC) used.

By changing the prototype filter, both FBMC and GFDM can achieve a lower OOB compared to CP-OFDM. However, as discussed above, side-effects also arose, such as more complicated pilot design and MIMO support. Furthermore, FBMC and GFDM require transceiver structures which are different from that used in CP-OFDM. This fact makes them less compatible with the existing 4G LTE systems, demanding more efforts in standardization.

	CP-OFDM	f-OFDM	UFMC	W-OFDM	FBMC OQAM	FBMC QAM	W-GFDM
FFT size	1024	1024	1024	1024	4096	4096	1024 × 14
Filter type	–	Windowed Sinc (RRC)	Chebyshev	–	PHYDYAS	[11]	RRC with rolloff 0.1
Filter length	–	512	72	–	4096	4096	1024
Filtering granularity	–	25 RB	Per RB	–	Subcarrier	Subcarrier	Subcarrier
Receiver processing	–	Filtering	Double-sized FFT	Windowing	–	LMMSE	SIC
Window length	–	–	–	TX 52 RX 10	–	–	–

Table 1. Evaluation assumptions for waveform comparison.

QUANTITATIVE ANALYSIS

A merely qualitative analysis of the pros and cons of different waveforms is far from enough to decide on the waveform for 5G. In this section, quantitative analysis and comparison among the waveforms under discussion are also provided.

BLER AND SPECTRUM EFFICIENCY IN FDD MODE

Although there are many different ways to measure the performance of waveforms, the block error rate (BLER) and spectrum efficiency in typical deployment scenarios are still the most important metrics. MIMO transmission is not considered since the associated transmission schemes for some candidates are still immature. The data bandwidth is 25 resource blocks (RBs), and ideal channel estimation is assumed to investigate the best performance of each waveform. A brief summary of evaluation assumptions is provided in Table 1, and specific assumptions are indicated along with the results. Note that additional windowing can also be applied to GFDM to soften the transition between adjacent blocks and thus reduce the OOB, and the resulting waveform is denoted as W-GFDM.

Single Band DL Transmission: We start with single band DL transmission. For FDD transmission, filter tails are overlapped with the adjacent symbols. The spectrum efficiency is determined by the BLER and the guard band overhead, which is the additional transition band needed to meet the spectrum mask in LTE. A modified Rapp power amplifier (PA) model with a transmit power of 46 dBm and a power backoff of 11.6 dB is used [13]. The BLER and spectrum efficiency of different waveforms are depicted in Fig. 3.

Two main observations can be made from Fig. 3. First, due to the intrinsic interference, it is rather difficult for the new transceiver required waveforms, including FBMC-QAM and GFDM, to support high-order modulations, even with more complicated receivers. Second, for the legacy OFDM transceiver supportable waveforms, both W-OFDM and UFMC incur significant performance losses compared to CP-OFDM in the high signal-to-noise ratio (SNR) region. The underlying reasons are similar, that is, the reduced CPs and ZPs are insufficient for sup-

pressing the ISI resulting from rich multi-path scattering. On the contrary, with properly designed filters (e.g., with confined time-domain energy spread and flat frequency response over the entire subband), the effectiveness of the CP in f-OFDM is less impacted; also, a sufficiently suppressed OOB and thereby guard band utilization can be achieved. For these reasons, f-OFDM provides the best BLER performance and the highest spectrum efficiency among all the discussed waveform candidates.

Multiple Subband UL Transmission: We then proceed to the case with multiple-subband UL transmission, with different and mixed numerologies. In addition to CP-OFDM in LTE, two more candidates, f-OFDM and W-OFDM, are investigated here, as they are the most discussed waveforms in the Third Generation Partnership Project (3GPP). A recently agreed polynomial power amplifier (PA) model (transmit power 22 dBm, backoff 8 dB, phase distortion 76.3° [13]) is also adopted for the UL BLER evaluation. More details about the scenarios under consideration can be found in [14, references therein].

As indicated in Fig. 4, with a fixed amount of guard band between adjacent subbands, f-OFDM provides more protection against interference from different numerologies, which results in better BLER performance and higher spectrum efficiency compared to W-OFDM. For the sake of brevity, we did not include all the evaluation results in this article. Interested readers are referred to [14, references therein].

OOBE AND OVERHEAD IN TDD MODE

OOBE reduction is a common motivation for all the new waveforms under consideration. As mentioned before, OOB reduction may typically be achieved by two methods: filtering or pulse shaping in either the time or frequency domain (e.g., f-OFDM and FBMC-QAM), and windowing in the time domain (e.g., W-OFDM and W-GFDM). Such filtering and windowing operations will extend the symbol duration and cause additional time-domain overhead, especially for TDD transmission. Theoretically, if the tolerable overhead is sufficiently large, the spectrum mask can always be satisfied. Therefore, the study of OOB should not be separated from the associated time-domain overhead in TDD mode.

As in the FDD single-band case, a modified

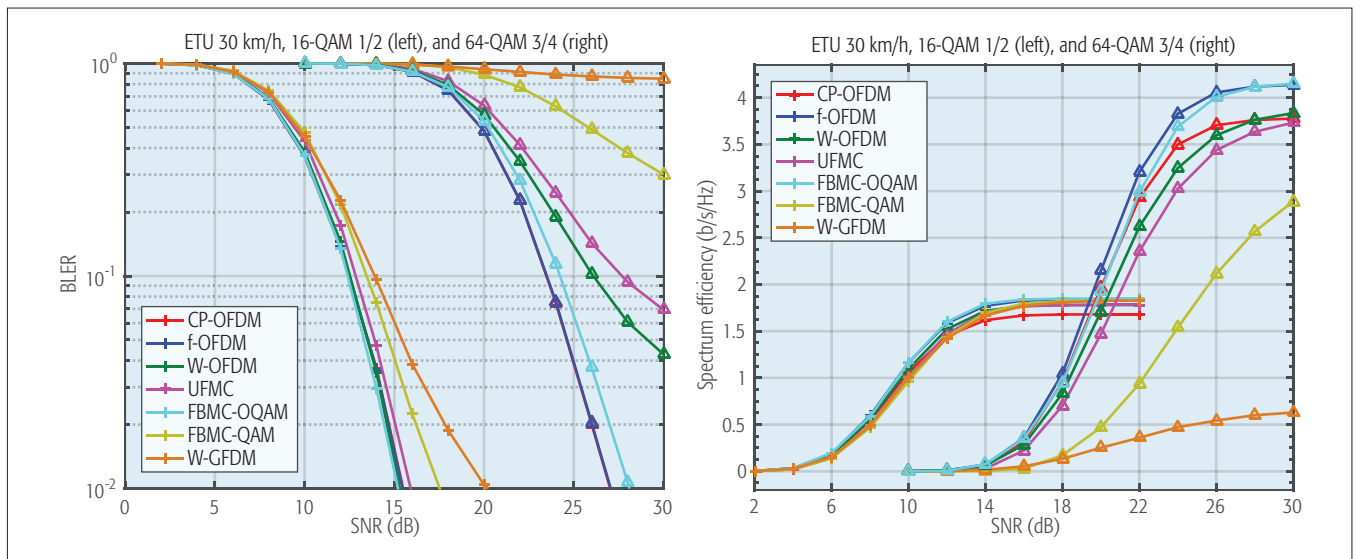


Figure 3. BLER and spectrum efficiency with single-band transmission.

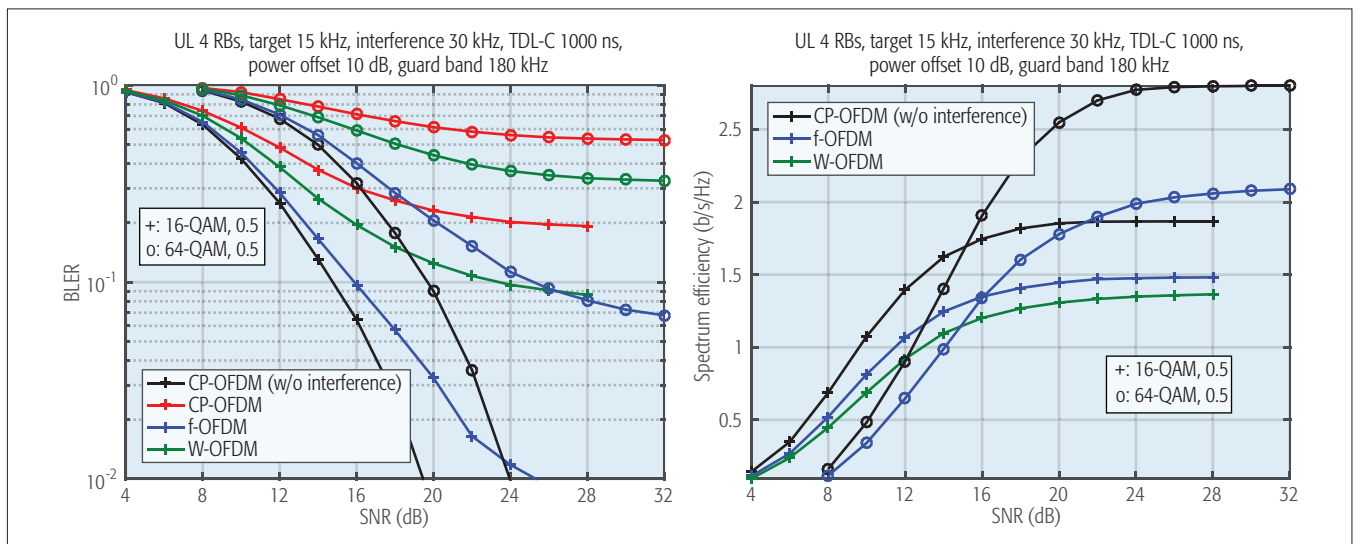


Figure 4. BLER and spectrum efficiency with multi-subband transmission.

Rapp PA model is used for investigating the DL OOB, with a transmit power of 46 dBm and a power backoff of 11.6 dB [13]. The minimum lengths of residual filter tails (after windowing and truncation) or extra guard interval for windowing are derived under the conditions that the LTE spectrum mask is met and the averaged error vector magnitude (EVM) [15] is less than 3 percent, required by LTE for 256-QAM. (Due to intrinsic interference, FBMC-QAM failed to meet this requirement.) The associated time-domain overheads of the waveforms, that is, the ratio of the length of residual tails or additional guard interval to the time duration of each transmission, are calculated as W-OFDM: ~0 percent, UPMC: 0 percent, f-OFDM: ~0 percent, FBMC-QAM: ~8 percent, W-GFDM: ~1 percent.

As can be seen from Fig. 5, the far-end OOB is dominated by the PA. Windowing operation does not help much in sharpening the transition region. In other words, non-negligible guard bands are still needed for W-OFDM

to allow for neighboring non-orthogonal transmissions. Similarly, the spectrum confinement achieved by UPMC is also insufficient. One can see that with properly designed filters, f-OFDM enables a sharp transition region, which could be beneficial for allowing neighboring non-orthogonal transmissions. Furthermore, FBMC-OQAM and W-GFDM provide the first and second best OOB performance, at the cost of additional tail overhead. Note that with negligible duration and energy spread, the tails in W-OFDM and f-OFDM can be absorbed in the guard period (GP) during switching between DL and UL transmissions.

To summarize the quantitative analysis, after reviewing the BLER and spectrum efficiency in FDD mode, and the OOB and tail overhead in TDD mode, one can observe that f-OFDM presents the best trade-off as follows: first, almost the same BLER performance as CP-OFDM; second, sufficiently low OOB with sharp transition regions; and third, negligible time-domain overhead.

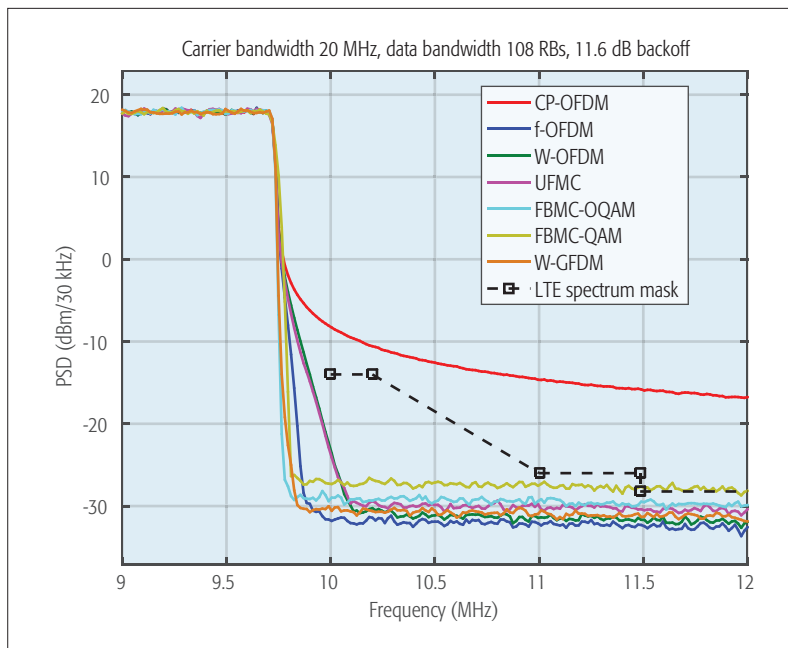


Figure 5. OOB of different waveforms.

CONCLUSION

In this article, an overview and an in-depth analysis of the requirements on the 5G waveform, as well as qualitative and quantitative comparison of the most discussed waveform candidates, including CP-OFDM, W-OFDM, UPMC, f-OFDM, FBMC, and GFDM, are presented. By exposing the underlying methodologies, associated advantages and disadvantages, the BLER performance and spectrum efficiency in FDD mode, and the OOB and tail overhead in TDD mode, the following design guidelines and operational suggestions are obtained for the 5G waveform:

- To achieve high spectrum efficiency in complex multi-path fading channels and support efficient MIMO transmission, the 5G waveform should have a high level of orthogonality in the complex field. In this way, the pilot scheme and precoding schemes developed for CP-OFDM can be maximally reused, and standardization efforts can be saved considerably. In addition, by having a negligible amount of ISI/ICI in the waveform itself, high-order modulations are not restricted, and low-complexity reception is enabled.

- In terms of BLER, spectrum efficiency, OOB, and tail overhead, as one may expect, applying filtering is generally better than windowing, at the cost of extra complexity. Furthermore, filtering is also preferred for achieving a sharp transition region, and thus high spectrum efficiency, for which windowing alone appears insufficient. With proper filtering, windowing could also be additionally incorporated to reduce the tail overhead and filtering complexity.

Taking all these observations into consideration, f-OFDM is recommended as the fundamental waveform for 5G, with which the flexibility required for supporting the diversity in services and deployment scenarios can be provided.

REFERENCES

- [1] J. G. Andrews *et al.*, "What Will 5G be?" *IEEE JSAC*, vol. 32, no. 6, Jun. 2014, pp. 1065–82.
- [2] ITU-R, "IMT Vision Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," tech. rep. M.2083-0, Sept. 2015.
- [3] 3GPP, "Study on Scenarios and Requirements for Next Generation Access Technologies," tech. rep. 38.913, Feb. 2016, v. 0.3.0.
- [4] P. Banelli *et al.*, "Modulation Formats and Waveforms for 5G Networks: Who Will Be the Heir of OFDM? An Overview of Alternative Modulation Schemes for Improved Spectral Efficiency," *IEEE Signal Processing Mag.*, vol. 31, no. 6, Nov. 2014, pp. 80–93.
- [5] Qualcomm Inc., "Waveform Candidates," R1-162199, Busan, Korea, Apr. 11–15, 2016.
- [6] Nokia, Alcatel-Lucent Shanghai Bell, "New Radio Waveforms for the Multi-Service Air Interface below 6 GHz," R1-165012, Nanjing, China, May 23–27, 2016.
- [7] J. Abdoli *et al.*, "Filtered OFDM: A New Waveform for Future Wireless Systems," *Proc. IEEE Int'l. Wksp. Signal Processing. Advances Wireless Commun.*, Stockholm, Sweden, June 2015, pp. 66–70.
- [8] X. Zhang *et al.*, "Filtered-OFDM – Enabler for Flexible Waveform in the 5th Generation Cellular Networks," *Proc. IEEE GLOBECOM*, San Diego, CA, Dec. 2015.
- [9] Huawei, HiSilicon, "f-OFDM Scheme and Filter Design," R1-165425, Nanjing, China, May 23–27, 2016.
- [10] Idaho National Laboratory, "Waveform for the Next Generation Radio Interface," R1-162248, Busan, Korea, Apr. 11–15, 2016.
- [11] C. Kim *et al.*, "QAM-FBMC: A New Multi-Carrier System for Post-OFDM Wireless Communications," *Proc. IEEE GLOBECOM*, San Diego, CA, Dec. 2015.
- [12] N. Michailow *et al.*, "Generalized Frequency Division Multiplexing for 5th Generation Cellular Networks," *IEEE Trans. Commun.*, vol. 62, no. 9, Sept. 2014, pp. 3045–61.
- [13] Nokia, Alcatel-Lucent Shanghai Bell, "[85-18] PA Assumptions for NR; Email Discussion Summary," R1-167297, Gothenburg, Sweden, Aug. 22–26, 2016.
- [14] Huawei, HiSilicon, "Summary on NR Waveform Evaluation Results," R1-166121, Gothenburg, Sweden, Aug. 22–26, 2016.
- [15] 3GPP Tech. Spec. 36.104, June 2016, v. 14.0.0.

BIOGRAPHIES

XI ZHANG [S'11, M'14] received his B.E. degree in communication engineering from the University of Electronic Science and Technology of China in 2010, and his Ph.D. degree in electronic and computer engineering from the Hong Kong University of Science and Technology in 2014. He is now working in the Communication Technology Laboratory, Huawei Technologies Co., Ltd. His current research interests are in the fields of wireless communication with millimeter wave and massive MIMO techniques.

LEI CHEN received his B.Sc. and M.Sc. degrees in communication and information engineering from the University of Electronic Science and Technology of China in 2005 and 2008, respectively. He is currently with Huawei Technologies Co., Ltd. His research interests are in physical layer design, including waveform, MIMO, and so on.

JING QIU received her Ph.D. degree in communication and information systems from the Beijing University of Posts and Telecommunications, China, in 2006, and worked at Chongqing University from 2006 to 2008. Currently she is working on 5G research with Huawei Technologies Co., Ltd. Her research interests are in air interface design and signal processing in wireless systems.

JAVAD ABDOLI [S'09, M'13] received his B.Sc. and M.Sc. degrees in electrical engineering from Sharif University of Technology, Tehran, Iran, in 2003 and 2006, respectively, and his Ph.D. degree in communication systems from the University of Waterloo, Ontario, Canada, in 2012. He is currently with Huawei Technologies Canada Co., Ltd. His research interests are in air interface design, including waveform and multiple access, and signal processing in wireless systems.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

SOFTWARE DEFINED VEHICULAR NETWORKS: ARCHITECTURES, ALGORITHMS, AND APPLICATIONS

BACKGROUND

Noticeable advance of wireless communications and pervasive use of mobile electronics have made vehicular networks no longer a futuristic promise but rather an attainable technology to meet the imminent demands towards reduced accidents, and improved road safety and efficiency. In addition to the initial safety applications, there are increasing demands from traveling users to access Internet for infotainment using IP-enabled smart devices, e.g., video streaming, web browsing, and file downloading, etc.. Such infotainment applications are likely to motivate a huge mass market in connected vehicles, which is expected to reach \$131.9 billion by 2019.

Due to the random vehicle mobility and varying communication environment, an integrated vehicular network comprising of heterogeneous access technologies, e.g., DSRC, WiFi, 4G/LTE, 5G, TV white space, etc., will be indispensable to provide the reliable and ubiquitous mobile coverage. Although the heterogeneous networking has been extensively studied for long in different contexts, the salient features of vehicular communications, e.g., varying road density, fast mobility, observable social pattern, have brought new challenges and led to fundamental and interesting research issues, e.g., how to flexibly configure and efficiently conduct resource allocation under a specific type of wireless network, how to enable the interoperation among multiple co-existing wireless networks, and how to effectively accommodate a large number of travelling users with various kinds of smart devices.

In addition to the advances in the underlying access technologies, cloud computing as a centralized control and management solution has become mature, representing an indispensable component for large scale vehicular networks and Intelligent Transportation Systems. In particular, the Software-Defined Networking (SDN) has been emerging as a promising paradigm to control the network in a systematic way, and gaining attention from both academia and industry. The flexibility and programmability of SDN that are lacking in today's distributed wireless substrate, not only make it attractive to satisfy the Quality of Services (QoS) requirements of vehicular multimedia services, but also greatly simplify the resource management in the heterogeneous vehicular networks with different access technologies.

This Feature Topic (FT) focuses on the crossroads between scientists, industry practitioners, and researchers from different domains in the wireless technologies, mobile computing and smart environments. We envision to provide a platform for researchers to further explore the domain and explore the challenges. In this FT, we invite researchers from academia, industry, and government to discuss challenging ideas, novel research contributions, demonstration results, and standardization efforts on Software Defined Vehicular Networks.

In this FT, we would like to try to answer some (or all) of the following questions:

How to define the vehicular tailor-made SDN architecture to support diverse vehicular applications and ubiquitous vehicular services in a flexible and efficient manner? How to utilize the SDN paradigm to improve the vehicular communication performance and service quality in Vehicle-to-Vehicle (V2V) or Vehicle-to-Infrastructure (V2I) networks? How to enable highly scalable and manageable vehicular networking and computing and how to enable the secure and efficient software defined vehicular communications, among others?

Topics of interest include, but are not limited to:

- Flexible architecture in Software Defined Vehicular Networks
- Vehicular applications and service scenarios in Software Defined Vehicular Networks
- Protocols for Software Defined Vehicular Networks (MAC, routing, mobility management, geo-networking, etc.)
- Testbed, implementation and deployment for Software Defined Vehicular Networks
- Internetworking technologies for heterogeneous Software Defined Vehicular Networks
- Data offloading in Software Defined Vehicular Networks
- Resource management and QoS Provisioning in Software Defined Vehicular Networks
- SDN Enabled Content Distribution in Vehicular Networks
- Software Defined Networking-based Vehicular Networks with Fog Computing
- Security and privacy in Software Defined Vehicular Networks

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 style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words (from introduction through conclusions). Figures and tables should be limited to a combined total of six. The number of references is recommended not to exceed 15. In some rare cases, more mathematical equations, figures, and tables may be allowed if well-justified. In general, however, mathematics should be avoided; instead, references to papers containing the relevant mathematics should be provided. Complete guidelines for preparation of the manuscripts are posted at <http://www.comsoc.org/commag/paper-submission-guidelines>. Please send a PDF (preferred) or MSWORD formatted paper via Manuscript Central (<http://mc.manuscriptcentral.com/commag-ieee>). Register or log in, and go to Author Center. Follow the instructions there. Select "July 2017/Software Defined Vehicular Networks: Architectures, Algorithms and Applications" as the Feature Topic category for your submission.

IMPORTANT DATES

- Manuscript Submission: December 1, 2016
- Decision Notification: March 1, 2017
- Final Manuscripts Due: April 15, 2017
- Publication Date: July 2017

GUEST EDITORS

Guangjie Han
hanguangjie@ieee.org

Yuanguo Bi
biyuanguo@mail.neu.edu.cn

Ammar Rayes
rayes@cisco.com

Mohsen Guizani
mguizani@ieee.org

Kaoru Ota
ota@csse.muroran-it.ac.jp

Wael Guibene
wael.guibene@intel.com

Tom H. Luan
tom.luan@deakin.edu.au

Haibo Zhou
h53zhou@uwaterloo.ca

Interference Management via Sliding-Window Coded Modulation for 5G Cellular Networks

Kwang Taik Kim, Seok-Ki Ahn, Yong-Seok Kim, Jeongho Park, Chiao-Yi Chen, and Young-Han Kim

SWCM aims to mitigate the adverse effects of interference at the physical layer by tracking the optimal maximum likelihood decoding performance with low-complexity decoding and minimal coordination overhead. The authors review the basic structure of the SWCM scheme built on the principles of network information theory, and discuss how it can be extended and implemented in practical wireless communication systems.

ABSTRACT

SWCM aims to mitigate the adverse effects of interference at the physical layer by tracking the optimal maximum likelihood decoding performance with low-complexity decoding and minimal coordination overhead. This article reviews the basic structure of the SWCM scheme built on the principles of network information theory, and discusses how it can be extended and implemented in practical wireless communication systems. Using a representative implementation based on LTE OFDM MIMO systems, extensive link-level and system-level performance simulations are carried out, which demonstrate that SWCM offers significant gains for all users over conventional interference-aware communication schemes. Network operating prerequisites for SWCM are also discussed to facilitate the standardization effort for its adoption in the fifth generation cellular network.

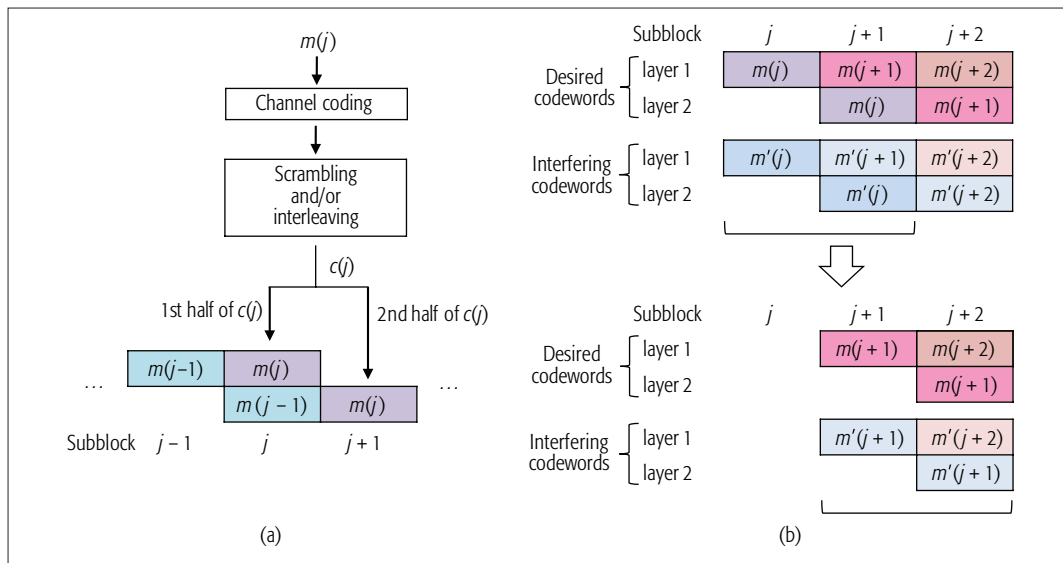
INTRODUCTION

According to the Third Generation Partnership Project (3GPP) study on scenarios and requirements for next-generation access technologies [1], the cell edge performance (typically measured by the spectral efficiency of the 5th-percentile user) of the fifth generation (5G) mobile network must be at least three times higher than that of IMT-Advanced. This cell edge performance, one of the key performance indicators of 5G, is crucially affected by co-channel interference from neighboring cells, which becomes more severe as small cells are deployed more densely in 5G. Consequently, proper mitigation of the performance degradation due to interference becomes one of the key challenges in the 5G radio access network. As a first step in addressing the co-channel interference problem on the receiver side, conventional linear receivers have been improved by incorporating statistical information on interference channels as in linear minimum mean square error interference rejection combining (LMMSE-IRC) receivers [2, 3], which are now widely used in commercial systems. As a next step, network assisted interference cancellation and suppression (NAICS) has been proposed, first as a study/work item for the 3GPP Long Term Evolution-Advanced (LTE-A) in Release 12, in which several types of interference-aware receivers and control signal-

ing for them were discussed. As a result of the standardization of NAICS [4], symbol-level interference-aware detection (IAD) [5] is expected to be implemented in many commercial receivers in the near future.

In network information theory, several decoding techniques have been studied for point-to-point (P2P) random codes over interference channels. Among these, treating interference as Gaussian and discrete noise roughly corresponds to LMMSE-IRC and IAD, respectively. Treating interference as noise, despite its advantage of low-complexity implementation, typically suffers from a significant rate loss when the interference signal strength at the receiver is moderate to high. In comparison, sequence-level interference-aware decoding such as information-theoretic simultaneous nonunique decoding (SND) can asymptotically achieve the rate region of the optimal maximum likelihood decoding (MLD) [6, references therein]. Despite this performance advantage, however, MLD and SND rely on some form of multiuser sequence detection, which cannot be implemented in practice due to its high complexity. The question then naturally arises as to how one can achieve the MLD/SND performance at low complexity, which would be the ultimate goal of physical-layer interference management.

A few approaches have been proposed recently to tackle this problem. First, instead of conventional P2P codes such as low-density parity check (LDPC) and turbo codes, one can utilize novel error correcting codes such as spatially coupled codes and polar codes adapted to simultaneous decoding of desired as well as interfering codewords. The resulting multiuser codes [7, 8], however, are often of very long block lengths, not suitable for typical wireless applications. Second, by tweaking decoding steps for conventional turbo codes, one can iteratively decode for both desired and interfering codewords. This interference-aware successive decoding (IASD) scheme [9] performs well in general and better than noniterative successive interference cancellation decoding in particular. However, the IASD scheme still falls short of achieving the MLD/SND performance, especially under a moderate interference level. Third, by transmitting codewords over multiple layers and multiple subblocks, one can achieve the MLD/SND per-



As its name suggests, a receiver in the SWCM scheme recovers desired messages by sliding the decoding window over multiple subblocks. In each decoding window, the desired codeword and/or the interfering codeword is recovered successively.

Figure 1. Basic SWCM encoder and decoder structures: a) basic encoder structure with each message transmitted over two layers and two subblocks. The (scrambled and/or interleaved) codeword $c(j)$ carries message $m(j)$; b) basic decoder structure with sliding-window decoding and successive cancellation decoding. To recover both the desired codeword $c(j)$ and the interfering codeword $c'(j)$, the receiver applies successive cancellation decoding according to a particular decoding order.

formance for P2P random codes using low-complexity successive cancellation decoding. This sliding-window superposition coding (SWSC) scheme [10, 11] is built on basic components of network information theory, carefully combining the ideas of block Markov coding, sliding-window decoding (both commonly used for multihop relaying and feedback communication), superposition coding without rate splitting (codewords mapped into multiple layers), and successive cancellation decoding (codewords recovered one by one). This conceptual coding scheme has been further developed into an implementable coded modulation scheme, whereby conventional binary codes (LTE turbo codes, to be specific) are mapped to transmitted symbols in staggered subblock layering and recovered successively in sliding decoding windows at the receivers [12, 13]. Even under simple bit-mapping rules and hard decision decoding, this sliding-window coded modulation (SWCM) scheme closely tracks the MLD/SND performance for two-user-pair Gaussian interference channels, with significant performance gain over LMMSE-IRC and IAD.

The goal of this article is twofold. First, we review the basic structure of SWCM and present how it can be enhanced and adapted to practical multiple-input multiple-output (MIMO) transceivers in cellular wireless networks. In particular, we discuss how SWCM can be implemented under typical LTE orthogonal frequency-division multiplexing (OFDM) resource allocation and MIMO transmission/reception scenarios. Second, we provide an extensive study on the performance of SWCM for plausible use cases, and demonstrate the performance gain over the aforementioned LMMSE-IRC, IAD, and IASD schemes. At the link level, we evaluate the performance of SWCM for the 2×2 MIMO Ped-B fading channel [14] under varying signal-to-noise ratios. At the system level, we simulate a macro urban environment network under 3GPP Release

12 NAICS assumptions, and measure average and cell edge throughput performance. In both cases, SWCM achieves significant performance gain over the existing schemes.

The rest of the article is organized as follows. The next section explains the basic structure of the SWCM scheme and its enhancement for adaptive transmission and reception. The following section demonstrates how the scheme can be implemented in practical OFDM MIMO systems. We then compare the performance of SWCM and competing schemes via link-level and system-level simulations, respectively. The next section discusses necessary network-side operations to enable interference-aware transmitters and receivers based on SWCM. The final section concludes the article.

SLIDING-WINDOW CODED MODULATION

BASIC ENCODER AND DECODER STRUCTURES

In the SWCM scheme, a single communication block consists of multiple, say b , subblocks, each consisting of n transmitted symbols. Each transmitted symbol is assumed to be decomposable into multiple layers; for example, a 16-quadrature amplitude modulation (QAM) symbol can be viewed as a combination of two 4-QAM layers, two 4-phase amplitude modulation (PAM) layers, or four binary phase shift keying (BPSK) layers. Each message is transmitted over multiple subblocks and multiple layers. Figure 1a illustrates the encoder structure in which each message is transmitted over two layers and two subblocks. To be concrete, we assume that a 4-PAM signal is transmitted, represented as a weighted superposition of two BPSK layers. To communicate the message $m(j)$ for $j = 1, 2, \dots, b-1$, the encoder uses a binary channel code of length $2n$ bits to form a codeword and its scrambled (and interleaved, if needed) version $c(j)$. For transmission in subblock j , the second half of

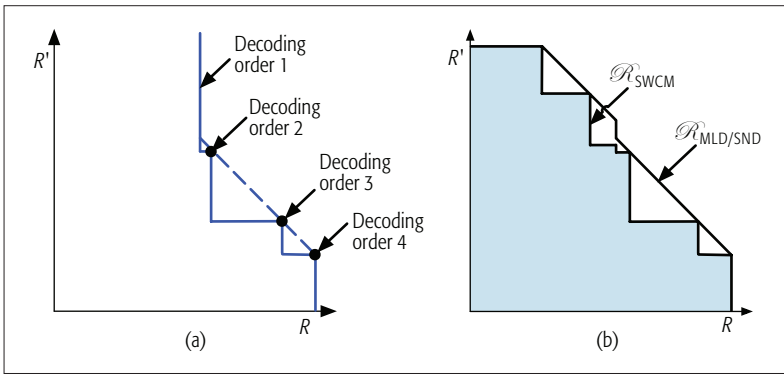


Figure 2. SWCM and MLD/SND achievable rate regions: a) SWCM and MLD/SND achievable rate regions at a receiver. The dotted diagonal line reflects the optimal MLD/SND rate region; b) SWCM and MLD/SND achievable rate regions of a system with two sender-receiver pairs. The SWCM region $\mathcal{R}_{\text{SWCM}}$ is the shaded inner region, while the MLD/SND region $\mathcal{R}_{\text{MLD/SND}}$ is the nonconvex polytope outer region.

the codeword $c(j - 1)$ from the previous message $m(j - 1)$ is mapped to one BPSK layer, while the first half of the codeword $c(j)$ from the current message $m(j)$ is mapped to the other BPSK layer. Thus, the superposition of the two BPSK layers, which contains partial information of two messages, $m(j - 1)$ and $m(j)$, forms a sequence of n 4-PAM symbols to be sent in subblock j . Overall, $(b - 1)$ messages are transmitted over b subblocks with one BPSK layer each missing during the initial and final subblocks, resulting in a slight rate loss from the actual code rate. A similar structure can easily be adapted for other transmission symbols, such as higher-order PAM/QAM and PSK symbols. It is worthwhile to point out the key difference in the encoder structure between SWCM and widely used bit-interleaved coded modulation (BICM). The latter is equivalent to mapping the scrambled codeword to BPSK layers transmitted in the same subblock, while in SWCM the scrambled codeword is mapped to BPSK layers over multiple subblocks in a staggered manner. This structure allows SWCM to mitigate interference within a codeword and from other codewords more efficiently, but it suffers the aforementioned rate loss (the larger b , the better) as well as error propagation over multiple subblocks (the smaller b , the better).

Each message is transmitted over multiple subblocks and multiple layers. As its name suggests, a receiver in the SWCM scheme recovers desired messages by sliding the decoding window over multiple subblocks. In each decoding window, the desired codeword and/or the interfering codeword is recovered, successively. Figure 1b illustrates an example of the decoding operation that corresponds to the encoder structure in Fig. 1a. In the decoding window spanning over subblocks j and $j + 1$, the receiver recovers the desired codeword $c(j)$ (using a conventional P2P decoder) and the interfering codeword $c'(j)$ (again using a conventional P2P decoder) successively. It then slides the decoding window to subblocks $j + 1$ and $j + 2$. Note that successive cancellation decoding is performed within and across decoding windows. If both the desired and interference encoders use the two-layer SWCM

encoder structure illustrated in Fig. 1a, there are four possible decoding orders at a receiver:

1. The desired codeword $c(j)$ only
2. $c(j)$, then the interfering codeword $c'(j)$
3. $c'(j)$, then $c(j)$
4. The next interfering codeword $c'(j + 1)$, then $c(j)$

with different trade-offs between the rate R of the desired codeword and the rate R' of the interfering codeword. Figure 2a illustrates the resulting achievable rate region of (R, R') and compares it with the optimal MLD/SND achievable rate region; see [10, 11] for the detailed information-theoretic analysis. When two sender-receiver pairs are communicating over an interference channel, the intersection of the achievable rate regions at both receivers determine the achievable rate region of the entire system, as illustrated in Fig. 2b.

ADAPTIVE TRANSCIEVER DESIGN TECHNIQUES FOR SWCM

The SWCM scheme discussed in the previous subsection can be enhanced to improve performance and robustness in a broad range of channel conditions to satisfy the desired quality of service (QoS) [13]. The basic SWCM encoding structure can be modified in several ways to increase achievable rates. First, one can adaptively control the number of superposition layers for both transmit points (TPs), which provides multiple code rate options by changing the shape of the corner points of the SWCM achievable rate region in Fig. 2b. Second, one can select various bit mapping rules at each TP, which allows different rate points to be achieved in Fig. 2b for the desired QoS. Third, one can assign different power allocation parameters between superimposed layers, which can potentially enlarge both MLD/SND and SWCM achievable rate regions.

The SWCM decoding structure can be modified in multiple ways to improve robustness under channel state information (CSI) mismatch. First, the decoding order (Fig. 2b) can be switched from one to another as the channel condition changes. Second, “soft” information can be utilized instead of “hard” information in each stage of successive cancellation decoding by storing log-likelihood ratios (LLRs) from the decoder output, which can then be used to calculate LLRs for the decoder input in the next stage. Third, iterative decoding, in addition to successive cancellation decoding, can be used; for instance, iterative decoding can be performed throughout the entire block (subblocks 1 through b), which can better track the optimal MLD/SND performance.

IMPLEMENTATION TECHNIQUES OF SWCM

As a coded modulation scheme that interfaces channel coding (typically binary) and physical-layer modulation, SWCM can be applied to different radio access technologies. In this section, we present a representative implementation for LTE OFDM MIMO systems using turbo codes.

MULTIPLE-INPUT MULTIPLE-OUTPUT

We modify the architecture of the MIMO transmission system in the LTE standard with three key changes to implement an SWCM MIMO

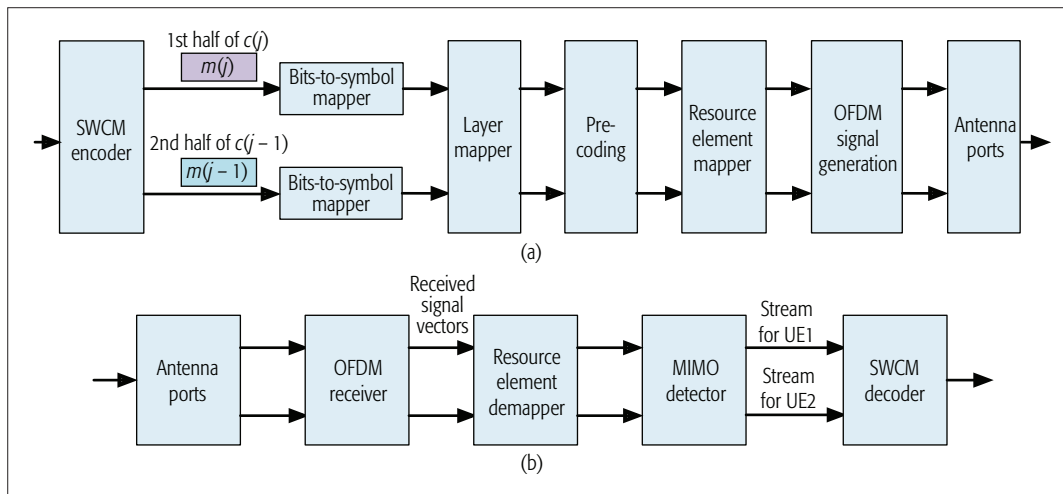


Figure 3. The SWCM MIMO system architecture: a) SWCM encoder, layer mapper, precoder, and resource element mapper for MIMO transmission; b) interference-aware MIMO detector and SWCM decoder.

system architecture. For simplicity, we explain these changes with an SWCM scheme using two layers for a 2Tx–2Rx antenna configuration as shown in Fig. 3, under the assumption that a single stream is communicated. Note that this illustrative design can readily be extended to multiple streams.

First, SWCM replaces BICM over QAM as the coding and modulation mapper. Unlike BICM, any of QAM, PAM, and PSK may be used as an SWCM layer to accommodate flexibility in symbol mapping. To be concrete, assume throughout this subsection that two SWCM layers use two 4-QAM or 4-PAM symbols as a pair. To send the message $m(j)$ for $j = 1, 2, \dots, b-1$, the encoder uses a (binary) LTE turbo code of length $4n$ bits to form a scrambled codeword $c(j)$. This is in comparison to the codeword of length $2n$ bits used for two BPSK layers as in the previous section. For transmission in subblock j , the second half of the previous codeword $c(j-1)$ is mapped to one 4-QAM (or 4-PAM) layer, while the first half of the current codeword $c(j)$ is mapped to the other 4-QAM (or 4-PAM) layer. The two 4-QAM (or 4-PAM) layers are fed into the layer mapper. A similar procedure can readily be adapted to binary or higher-order PSK and higher-order QAM/PAM symbols.

Second, the layer mapper takes a symbol pair from the two layers to form an input to the precoder. There are two operation modes in the layer mapper: the transmit diversity mode and the spatial multiplexing mode. In the transmit diversity mode, the layer mapper combines one symbol each from the two layers to form a scalar QAM symbol input to the precoder. If the two layers use 4-QAM symbols, the 4-QAM symbol from one layer is superimposed on the 4-QAM symbol from the other layer to form a 16-QAM symbol. If the two layers use 4-PAM symbols, the pair of PAM symbols are taken as in-phase and quadrature components of a 16-QAM symbol. In the spatial multiplexing mode, the layer mapper simply takes two 4-QAM symbols from the two layers to form a vector input to the precoder.

Third, the precoder uses beamforming matrices (2×1 in the transmit diversity mode and 2

$\times 2$ in the spatial multiplexing mode) to achieve the best performance under the SWCM encoding/decoding structure instead of interference as noise (IAN) as in LTE. As the output of the precoder, two sequences of n complex symbols are sent in subblock j via two antenna ports for $j = 1, 2, \dots, b$.

FRAME STRUCTURE AND RESOURCE ALLOCATION

As in the LTE standard, the scheduler allocates resource blocks (RBs) or resource elements (REs) in a subframe to one or more UE. To be concrete, assume that $nb = 1200$ physical REs are allocated to a UE for SWCM transmission in a subframe. Physical REs are mapped to virtual REs by pseudorandom interleaving for frequency diversity. The corresponding 1200 virtual REs are divided into $b = 10$ subblocks, each can carry $n = 120$ 4-QAM or 16-QAM symbols. The size of a SWCM subblock (480 bits for the entire codeword when 4-QAM or 4-PAM layers are used) is well within the maximum range of the quadratic permutation polynomial (QPP) interleaver, which is 6144 for the LTE turbo code.

LINK-LEVEL PERFORMANCE

As expected from the corresponding results in network information theory [6, 10], the SWCM scheme can increase spectral efficiency and reduce decoding error rates over conventional schemes, by utilizing the layered signaling and staggered transmission structure. First, layered signaling enables more flexible decoding of some layers of the codeword from another UE, achieving higher rates than nonlayered schemes. Second, staggered transmission allows a longer codeword to be sent over multiple layers, achieving diversity and robustness for decoding, as either the desired or interference signal, at different receivers. This section presents link-level performance simulation results to demonstrate these benefits quantitatively.

We consider the two-user-pair 2×2 symmetric MIMO interference channel model with International Telecommunication Union (ITU) Ped-B fading and additive Gaussian noise under

The SWCM scheme can increase spectral efficiency and reduce decoding error rates over conventional schemes, by utilizing the layered signaling and staggered transmission structure.

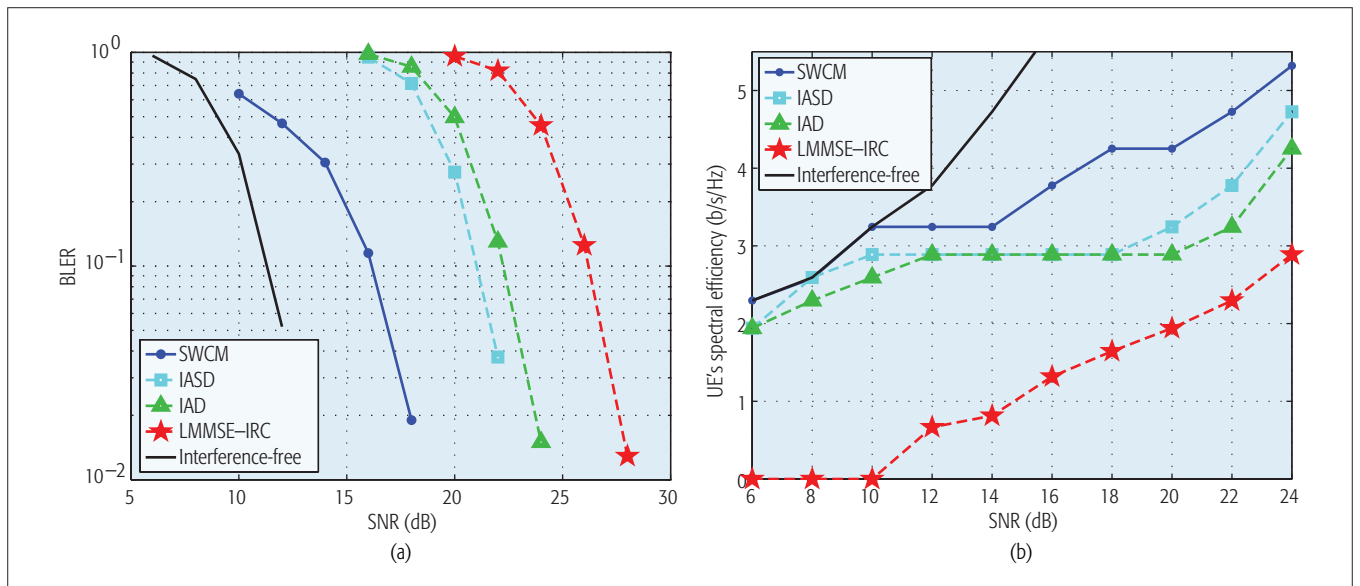


Figure 4. Link-level performance for the 2/2-layer SWCM MIMO scheme with adaptive iterative soft decoding, IASD, IAD, and LMMSE-IRC in the 2×2 MIMO symmetric Ped-B interference channel with average received INR = 15 dB: a) BLER vs. average received SNR at MCS 13 (16-QAM) and RI = 2; and b) achievable UE rates vs. average received SNR at BLER = 0.1 and RI = 2.

the 3GPP LTE standard subframe structure [15] and resource allocation described in the previous section. Ped-B channel gains for an OFDM symbol of the subframe over 2048 subcarriers are obtained by taking a fast Fourier transform of six Rayleigh distributed multipath channel taps [14]. The Jakes model is used for time correlation of the channel gains due to pedestrian mobility (3 km/h in the Ped-B channel model) between OFDM symbols of 32.5 ns duration. The channel gains for all links are generated by applying this process independently. The average signal-to-noise ratio (SNR) (or interference-to-noise ratio [INR]) is defined as the ratio of the average received power of the sum of six multipath signals of the desired codeword (or the average received power of the sum of six multipath signals of the interference codeword) to the average power of the noise at the receiver.

The transmitter uses the single-stream two-layer SWCM scheme with an identity matrix (in the spatial multiplexing mode) as the precoder by mapping the sequence of each SWCM layer directly to each precoder output in order to evaluate its simplest open loop spatial multiplexing performance. The block length of SWCM is matched to that of IASD, IAD, and LMMSE-IRC. The receiver uses the adaptive iterative soft decoding technique with eight iterations for each codeword (the same number of iterations used for other schemes to be compared).

We evaluate the performance of SWCM and existing schemes in the 2×2 MIMO Ped-B interference channel with average INR of 15 dB. As shown in Fig. 4a, simulation results for the SNR gain at block error rate (BLER) of 0.1 demonstrate that the SWCM scheme outperforms LMMSE-IRC [2, 3] (now widely used in commercial systems) by 10.1 dB, IAD [5] (Release 12 NAICS receiver) by 6.1 dB, and IASD [9] (one of the most advanced interference-aware receivers prior to SWCM) by 4.9 dB. As shown

in Fig. 4b, simulation results for the achievable UE rates under symmetric QoS, again at BLER of 0.1, demonstrate that the SWCM scheme outperforms LMMSE-IRC, IAD, and IASD over all SNR regimes, when the best modulation and coding scheme (MCS) is chosen for each scheme at a given SNR. At the same time, for each fixed MCS, SWCM uniformly outperforms the other schemes (data not shown). For example, SWCM outperforms IAD and IASD by 47 percent and LMMSE-IRC by 159 percent at the average SNR of 18 dB.

SYSTEM-LEVEL PERFORMANCE

The improved link-level performance does not necessarily translate to a system-wide throughput gain of the same magnitude, since the scheduler based on proportional fairness (PF) metrics may not always select a UE with a channel that has a moderate to high interference level, and thus the performance would benefit greatly from interference-aware decoding as shown in the previous section. In this section, we present simulation results for system-level performance gains of SWCM over conventional schemes, which partly answer the question of to what extent the promising link-level performance of SWCM can carry over to that of the entire network.

Since there is no system specification for 5G yet, we carry out the system-level performance simulation based on 3GPP Release 12 NAICS evaluation assumptions [4]. In particular, we consider NAICS scenario 1, a wrap-around homogeneous macro network with 19 hexagonal cells (3 sectors per cell) and use the 2D spatial channel model (SCM) with a 2Tx-2Rx cross polarized antenna configuration for urban macro (UMa) scenarios between a TP and a UE moving at 3 km/h. The system bandwidth is 10 MHz, and a file transfer protocol traffic model is used for a partial-load scenario. UEs feed back wideband channel quality indicator

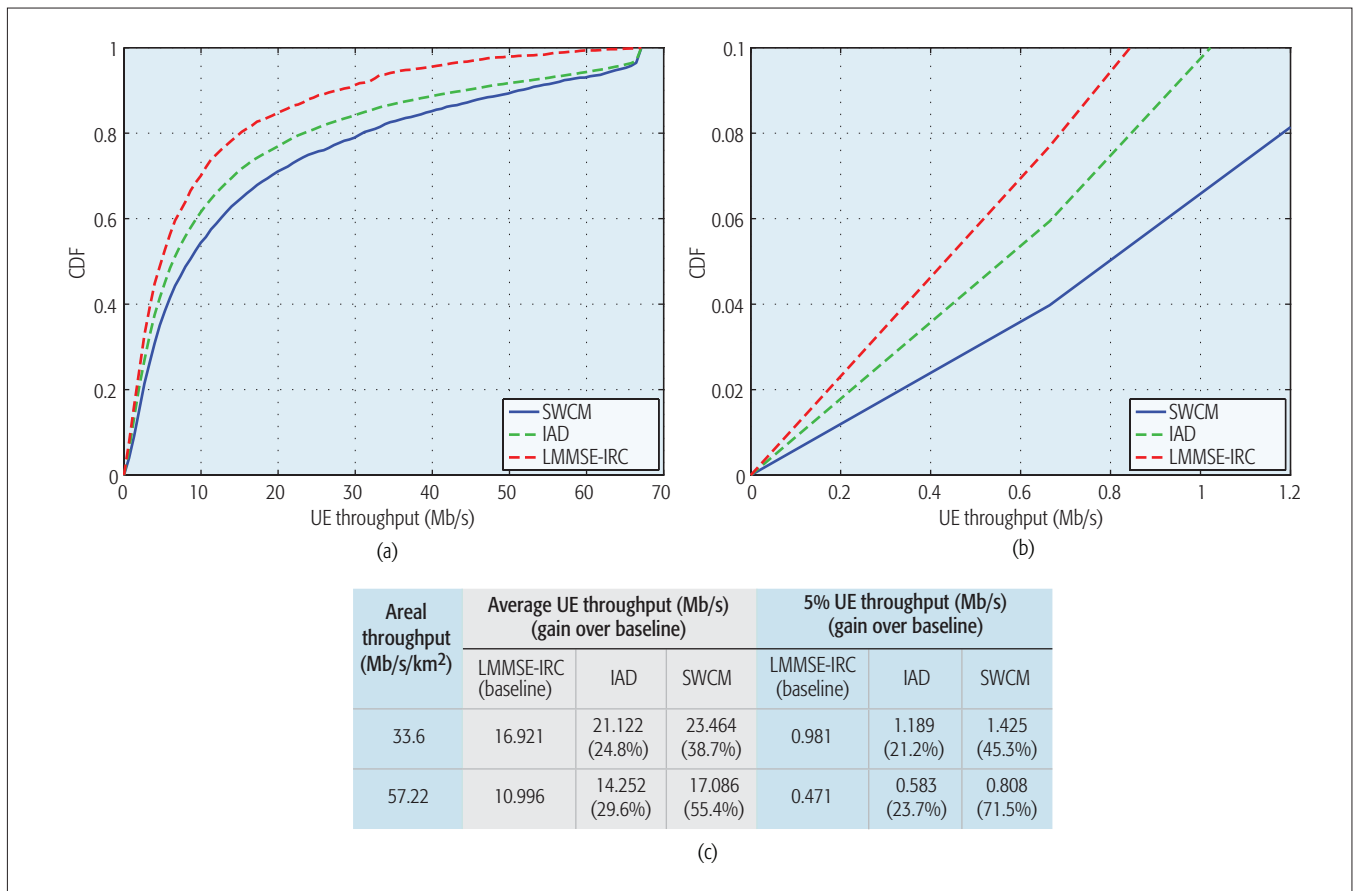


Figure 5. System-level performance in terms of the distribution of UE throughput for the 2/2-layer SWCM MIMO scheme, IAD, and LMMSE-IRC in the NAICS homogeneous scenario: a) UE throughput CDF; b) bottom 10 percent UE throughput CDF; c) summary.

(CQI), precoding matrix indicator (PMI), and rank indicator (RI) (with channel estimation errors reflected) to the TP at every predefined 5 ms interval, and feedback delay is also assumed to be 5 ms. Two LTE schemes (LMMSE-IRC and IAD) and SWCM are compared. UEs with LMMSE-IRC, IAD, or SWCM are configured to adapt PMIs and RIs according to the channel status. Two UEs in two different serving cells are paired for SWCM transmission based on PF scheduling metrics.

Figure 5 shows the cumulative distribution functions (CDFs) of cell-wide and cell-edge throughputs of the three schemes. SWCM achieves significant gains in both average UE throughput (the area above CDF) and cell-edge throughput (5th percentile CDF). SWCM outperforms LMMSE-IRC by 71.5 percent and IAD by 38.6 percent at the cell edge under areal throughput of 57.22 Mb/s/km² (or resource utilization of 54.44–63.94 percent). Note that user perceived throughput (during active time) is defined as the size of a burst divided by the time between its first packet arrival and its last packet reception. As for the average UE throughput, SWCM achieves gains of 55.4 percent and 19.9 percent over LMMSE-IRC and IAD, respectively, under the same areal throughput. As shown in Fig. 5c, the gain is more significant for the higher areal throughput environment, since UEs suffer from more severe co-channel interference.

NETWORK OPERATING PREREQUISITES FOR SWCM

As a sequence-level interference-aware scheme, network-side supports are crucial for SWCM, since it requires information about interfering cells as well as the serving cell. Specifically, the SWCM scheme needs CSI feedback from a UE to a serving TP for multiple cooperating cells, joint scheduling over cooperating TPs for the SWCM transmission, a reference signal design for interfering channel estimation, and downlink control signaling for the SWCM decoding operation at UEs. Table 1 compares the requirements of different interference-aware transceiver techniques and the corresponding 3GPP standardization states.

CHANNEL STATE INFORMATION FEEDBACK FOR MULTIPLE COOPERATING CELLS

For joint scheduling of multiple cells, it is important to infer the SWCM achievable rate region associated with each dominant interfering cell from CSI feedback. For this purpose, the UE reports to its serving TP the required CSI including RIs and PMIs associated with the combination of a serving cell and a particular interfering cell, and CQIs corresponding to corner points of the SWCM achievable rate region in Fig. 2a. There is an obvious trade-off between the amount of CSI reported (uplink control overhead) and the accuracy of the inferred achievable rate regions.

Changes from conventional networks	Required by				Standardized as
	LMMSE-IRC	IAD	IASD	SWCM	
CRS-associated information sharing for interference channel estimation	No	Yes	Yes	Yes	Release 12
PMI, RI, MCS, UE-specific RNTI sharing for interference decoding	No	No	Yes	Yes	N/A
Joint scheduling among TPs	No	No	Yes	Yes	Operator specific issue
A new Tx mode	No	No	No	Yes	N/A

Table 1. Prerequisites for interference-aware transceivers and standardization states.

JOINT SCHEDULING OVER COOPERATING TPs

Joint scheduling is performed over the cooperating TPs based on the CSI feedback by comparing the inferred SWCM achievable rate regions of different pair combinations under fairness criteria. Since data sharing is not required to allocate resources jointly, to determine data rates and transmission schemes, or to perform link adaptation for the SWCM transmission, only CSI reports are shared over the X2 interface among cooperating TPs, requiring very small backhaul bandwidth compared to the conventional backhaul bandwidth necessary for data traffic.

Efficient joint scheduling typically relies on low-latency backhaul and CSI feedback. Thus, in order to maintain the full gain of the SWCM scheme, the network is responsible for meeting the latency requirement, in most cases one to three transmission time intervals (TTIs) due to hybrid automatic repeat request (HARQ) processing for 4G or 5G networks. Note that even under some loss due to feedback and backhaul latencies, the SWCM gain can be maintained to some extent since joint scheduling is performed based on the latest CSI reports from serving UEs, and each TP adjusts scheduling accordingly on the TTI basis. Long-term outer-loop rate control (OLRC) and UE-side techniques such as adaptive decoding in the earlier subsection can also alleviate the performance loss.

DOWNLINK CONTROL AND REFERENCE SIGNALING

The SWCM reception procedure requires additional dynamic signaling besides higher-layer signaling supported in the 3GPP Release 12 NAICS for symbol-level interference management. This signaling requirement includes semistatic parameters (information that does not change often, once they are allocated) such as cell-specific reference signal antenna port (CRS AP), cell ID, data to RS energy per resource element (EPRE) P_A (UE-specific) and P_B (serving and interfering cell-specific) for channel estimation, system bandwidth, multicast-broadcast single-frequency network (MBSFN) configuration, resource allocation, precoding granularity information (1RB–4RB pairs), and transmission mode (TM1–TM9) associated with physical cell ID. Additional dynamic signaling may be carried by the downlink control information (DCI) message and consists of dynamic parameters (information

that may vary per subframe depending on the UE channel status) such as scheduling information jointly configured among the involved cells; MCS, PMI, RI, and radio network temporary identifier (RNTI) of the paired UE belonging to the interfering cell; and an indication of the use of SWCM transmissions. It is also crucial for a UE to estimate channels from interfering TPs as well as a serving TP in order to perform the SWCM reception and report CSI feedback to the serving TP for joint scheduling.

CONCLUDING REMARKS

As one of the most advanced interference-aware communication techniques, the sliding-window coded modulation scheme closely tracks the performance of maximum likelihood sequence decoding at low complexity, effectively mitigating adverse effects of co-channel interference. Confirming previously reported theoretical gains, link-level and system-level performance simulations of our representative implementation under the current LTE OFDM MIMO system architecture demonstrate that SWCM offers significant gains in both cell average and cell edge throughput over conventional schemes at comparable decoding complexity, with a few feasible modifications of the conventional network operations. These results indicate that the SWCM scheme has promising potential to become a basic building block for interference management in 5G and subsequent generation cellular systems.

ACKNOWLEDGMENTS

The authors would like to thank Lele Wang and Hosung Park for their earlier contributions and helpful discussions that led to the current work.

REFERENCES

- [1] 3GPP TR 38.913, "Scenarios and Requirements for Next-Generation Access Technologies," v. 0.3.0, 2016.
- [2] 3GPP TR 25.963, "Feasibility Study on Interference Cancellation for UTRA FDD UE," v. 11.0.0, 2012.
- [3] Y. Ohwatori *et al.*, "Performance of Advanced Receiver Employing Interference Rejection Combining to Suppress Inter-Cell Interference in LTE-Advanced Downlink," *Proc. IEEE VTC-Fall*, San Francisco, CA, Sept. 2011, pp. 1–7.
- [4] 3GPP TSG-RAN WG1 Meeting #78, "R1-143535: LS for Rel-12 NAICS," Rel. 12, 2014.
- [5] J. Lee, D. Toumpakaris, and W. Yu, "Interference Mitigation via Joint Detection," *IEEE JSAC*, vol. 29, no. 6, June 2011, pp. 1172–84.
- [6] B. Bandemer, A. El Gamal, and Y.-H. Kim, "Optimal Achievable Rates for Interference Networks with Random Codes," *IEEE Trans. Info. Theory*, vol. 61, no. 12, Dec. 2015, pp. 6536–49.
- [7] A. Yedla *et al.*, "Universal Codes for the Gaussian MAC via Spatial Coupling," *Proc. 49th Annual Allerton Conf. Commun. Control Comp.*, Monticello, IL, Sept. 2011, pp. 1801–08.
- [8] L. Wang, and E. Sasoglu, "Polar Coding for Interference Networks," 2014, <http://arxiv.org/abs/1401.7293>.
- [9] J. Lee, H. Kwon, and I. Kang, "Interference Mitigation in MIMO Interference Channel via Successive Single-User Soft Decoding," *Proc. UCSD Info. Theory Appl. Workshop*, La Jolla, CA, Feb. 2012, pp. 180–85.
- [10] L. Wang, E. Sasoglu, and Y.-H. Kim, "Sliding-Window Superposition Coding for Interference Networks," *Proc. IEEE Int. Symp. Info. Theory*, Honolulu, HI, July 2014, pp. 2749–53.
- [11] L. Wang, *Channel Coding Techniques for Network Communication*, Ph.D. thesis, UCSD, 2015.
- [12] H. Park, Y.-H. Kim, and L. Wang, "Interference Management via Sliding-Window Superposition Coding," *Proc. Int'l. Wksp. Emerging Technologies for 5G Wireless Cellular Networks, IEEE GLOBECOM*, Austin, TX, Dec. 2014, pp. 1057–61.
- [13] K. T. Kim *et al.*, "Adaptive Sliding-Window Coded Modulation in Cellular Networks," *Proc. IEEE GLOBECOM*, San Diego, CA, Dec. 2015, 7 pp.
- [14] ITU-R Rec. M.1225, "Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000," 1997.
- [15] 3GPP TS 36.211, "Physical Channels and Modulation," Release 12, 2013.

BIOGRAPHIES

KWANG TAIK KIM [M] (kwangtaik.kim@samsung.com) received his B.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology in 2001, and his M.S. and Ph.D. degrees in electrical and computer engineering from Cornell University in 2006 and 2008, respectively. He is currently a principal engineer in the Next Generation Communications Business Team at Samsung Electronics, where he does research on fifth generation cellular systems. He is a recipient of the 2014 Samsung CEO Award of Honor in the Technical Division. His research interests lie in information theory, and the fundamental and theoretical aspects of wireless and wireline communications.

SEOK-KI AHN (seokki.ahn@samsung.com) received his B.S., M.S., and Ph.D. degrees in electronics and electrical engineering from Pohang University of Science and Technology in 2006, 2008, and 2013, respectively. From 2010 to 2013, he was a student on scholarship at the Digital Media and Communications R&D Center of Samsung Electronics. Since 2013, he has been with Samsung Electronics as a senior engineer. His current research interests include channel coding, MIMO transceiver design, and broadband communications.

YONG-SEOK KIM (yongseok45.kim@samsung.com) received his M.S. degree in electrical engineering from Seoul National University, Korea, in 2004. In 2004, he joined Samsung Electronics, where he is currently a senior engineer in the Next Generation Communications Team. He worked on system designs and analysis for pre-4G trial (4x4 MIMO) systems, IEEE 16e/16m, and LTE systems. His current research interests include system designs and analysis for mmWave communications and new radio access technology in 5G.

JEONGHO PARK (jeongho.jh.park@samsung.com) received his B.S., M.S., and Ph.D. degrees in electronic engineering from Yonsei University in 1997, 2000, and 2005, respectively. Since he joined Samsung Electronics in 2005, he has mainly been engaged in research and development of wireless communications. Standardization is also his interest including IEEE 802, 3GPP RAN,

and ITU-R IMT-Systems. Currently, he is a director of the Next Generation Communications Business Team at Samsung Electronics and takes the lead in research and development of breakthrough technologies in 5G area.

CHIAO-YI CHEN (chc111@ucsd.edu) received his B.S. degrees in electrical engineering, and computer science and information engineering from National Taiwan University, Taipei, in 2006. During his undergraduate studies, he spent one year analyzing the impact of synchronization errors in CDMA systems with frequency domain equalization at Tohoku University, Japan. He received his M.S. degree in electrical and computer engineering from the University of California, San Diego (UCSD), and is pursuing his Ph.D. degree at the same university. He worked as an intern at Broadcom Corporation, Qualcomm Inc., and Blue Danube Systems. His research interests include information theory, communication theory, and universal processing. He received the 2002 National Taiwan University Presidential Award, the 2005 Japan Student Service Organization Scholarship, and the 2008 UCSD Electrical and Computer Engineering Departmental Fellowship.

YOUNG-HAN KIM [F] (yhk@ucsd.edu) received his B.S. degree with honors in electrical engineering from Seoul National University, Korea, in 1996, and his M.S. degrees in electrical engineering and statistics, and his Ph.D. degree in electrical engineering from Stanford University in 2001, 2006, and 2006, respectively. In 2006, he joined UCSD, where he is currently an associate professor of electrical and computer engineering. His research interests are in information theory, communication engineering, and data science. He coauthored the book *Network Information Theory* (Cambridge University Press, 2011). He is a recipient of the 2008 NSF Faculty Early Career Development (CAREER) Award, the 2009 U.S.-Israel Binational Science Foundation Bergmann Memorial Award, the 2012 IEEE Information Theory Paper Award, and the 2015 IEEE Information Theory Society James L. Massey Research and Teaching Award for Young Scholars. He served as an Associate Editor of *IEEE Transactions on Information Theory* and a Distinguished Lecturer for the IEEE Information Theory Society.

Waveform and Numerology to Support 5G Services and Requirements

Ali A. Zaidi, Robert Baldemair, Hugo Tullberg, Hakan Björkegren, Lars Sundström, Jonas Medbo, Caner Kilinc, and Icaro Da Silva

The authors propose a flexible physical layer for the NR to meet the 5G requirements. A symmetric physical layer design with OFDM is proposed for all link types, including uplink, downlink, device-to-device, and backhaul. A scalable OFDM waveform is proposed to handle the wide range of carrier frequencies and deployments.

ABSTRACT

The standardization of the next generation 5G radio access technology has just started in 3GPP with the ambition of making it commercially available by 2020. There are a number of features that are unique for 5G radio access compared to the previous generations such as a wide range of carrier frequencies and deployment options, diverse use cases with very different user requirements, small-size base stations, self-backhaul, massive MIMO, and large channel bandwidths. In this article, we propose a flexible physical layer for the NR to meet the 5G requirements. A symmetric physical layer design with OFDM is proposed for all link types, including uplink, downlink, device-to-device, and backhaul. A scalable OFDM waveform is proposed to handle the wide range of carrier frequencies and deployments.

INTRODUCTION

The standardization of the next generation radio technology has started in the Third Generation Partnership Project (3GPP) this year (2016) with the ambition of making fifth generation (5G) wireless systems commercially available around 2020. There are three main challenges that need to be addressed by 5G radio access technology to enable a truly networked society: a massive growth in the number of connected devices, a massive growth in traffic volume, and an increasingly wide range of applications with varying requirements and characteristics. Broadly, we can classify 5G use cases (or services) in the following groups:

- Enhanced mobile broadband (eMBB), requiring very high data rates and large bandwidths
- Ultra-reliable low-latency communications (URLLC), requiring very low latency, and very high reliability and availability
- Massive machine type communications (mMTC), requiring low bandwidth, high connection density, enhanced coverage, and low energy consumption at the user end.

The requirements for the above mentioned 5G services are diverse, and have implications for new spectrum and deployments. New spectrum for 5G is expected to be available by 2020. The actual frequency bands and the amount of spectrum have not been identified yet. All bands,

from below 1 GHz up to 100 GHz, are potential candidates for 5G [1]. 5G services will require a range of different bandwidths. At the low end of the scale, support for massive machine connectivity with relatively low bandwidths is envisioned. In contrast, very wide bandwidths may be needed for high-capacity scenarios, for example, 4K video and future media. Millimeter-wave spectrum bands (i.e., near and above 30 GHz) will play a role in some deployments to reach the envisioned capacity [2].

3GPP aims to develop and standardize components for a new radio access technology (RAT), which is envisioned to operate in frequencies up to 100 GHz to serve the diverse use cases. The new RAT is referred to as NR throughout this article, which is currently the accepted acronym in 3GPP [3]. NR is intended to be optimized for performance without considering backward compatibility in the sense that legacy Long Term Evolution (LTE) user equipments (UEs) do not need to be able to camp on an NR carrier. LTE is also expected to evolve to capture a part of the 5G requirements. The vision of 5G wireless access is shown in Fig. 1, where NR and LTE evolution are integral parts of 5G. LTE evolution is expected to operate below 6 GHz frequencies, and NR is envisioned to operate from sub-1 GHz up to 100 GHz. A tight integration of NR and LTE is envisioned in order to efficiently aggregate NR and LTE traffic.

The first step of NR development will be the physical layer design. This article provides principles for the design of waveform and numerology.¹ The article is organized as follows. In the following section, we highlight key design requirements for NR. Based on the design requirements, we then propose waveform and numerology. Finally, we conclude the article.

PHY DESIGN REQUIREMENTS FOR NR

In the following, we list important features of NR that have implications on new waveform and numerology:

- NR has to support a wide range of frequencies, bandwidths, and deployment options. NR should support diverse use cases such as eMBB, URLLC, and mMTC. These requirements ask for a flexible waveform, numerology, and frame structure.
- NR has to support applications with very low latency, which requires very short subframes.

¹ Numerology refers to waveform parametrization, such as cyclic prefix and subcarrier spacing in OFDM.

- NR should support both access and backhaul links by dynamically sharing the spectrum. NR should also support device-to-device (D2D) communication, including vehicle-to-anything (V2X) communication. This implies that NR waveform and numerology should be designed keeping in view various link types including uplink (UL), downlink (DL), sidelink,² and backhaul.

- NR has to enable the full potential of multi-antenna technology. The number of antenna elements may vary, from a relatively small number of antenna elements in LTE-like deployments to many hundreds in NR, where a large number of active or individually steerable antenna elements are used for beamforming, single-user multiple-input multiple-output (SU-MIMO) and multi-user MIMO (MU-MIMO). NR waveform and numerology must unleash the full potential of massive MIMO.

- NR is envisioned to mainly be based on time-division duplex (TDD) at high frequencies (above 3 GHz) and mainly on frequency-division duplex (FDD) at lower frequencies. The waveform, numerology, and frame structure should be chosen to enable efficient time/frequency utilization for FDD and TDD deployments, respectively.

- At very high frequencies, base stations can be small (low-cost) access nodes, putting similar requirements in DL as in UL (transmit power, hardware impairments, etc.). This suggests a physical layer design that is symmetric in UL and DL.

The key features of NR that have implications on the design of waveform and numerology are also summarized in Fig. 2.

NR WAVEFORM

OFDM is currently used in LTE for DL transmission. In March 2016, 3GPP agreed to study various features of NR assuming orthogonal frequency-division multiplexing (OFDM) unless significant gains can be demonstrated by any other waveform [3]. This section assesses OFDM for a number of key performance indicators (KPIs) for different link types (UL, DL, sidelink, backhaul), and concludes that OFDM is indeed an excellent choice for NR. A few other relevant multi-carrier and single-carrier waveforms are also discussed briefly.

ASSESSMENT OF OFDM

OFDM has been widely studied in the literature [4]. In the following, we assess the performance of OFDM for a number of KPIs. Different link types impose different levels of requirements on the waveform performance indicators at different frequencies. An assessment of OFDM is therefore made for all link types.

The key performance indicators for NR waveform are as follows.

Spectral Efficiency: OFDM is well known to be highly spectral-efficient. Spectral efficiency is vital to meet extreme data rate requirements. In general, spectral efficiency is more crucial at lower carrier frequencies than at higher frequencies, since the spectrum is not as precious at higher frequencies due to the availability of potentially much larger channel bandwidths. Spectral efficiency is very important for UL and

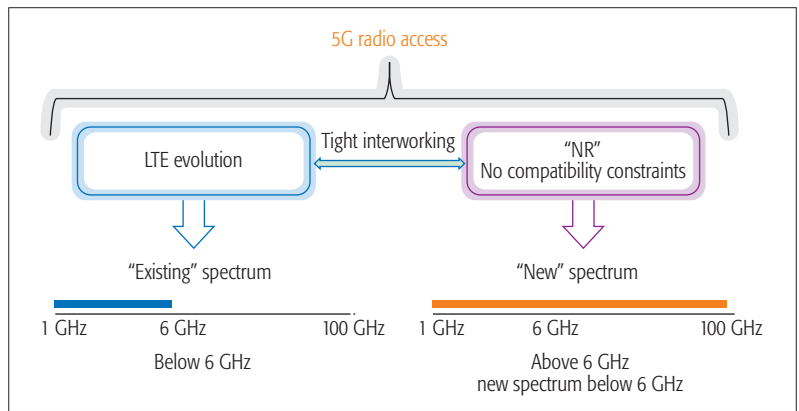


Figure 1. Radio access vision for 2020 and beyond: 5G radio access comprises LTE evolution and a new RAT (NR) that is not backward-compatible with LTE and is operable from sub-1 GHz to 100 GHz.

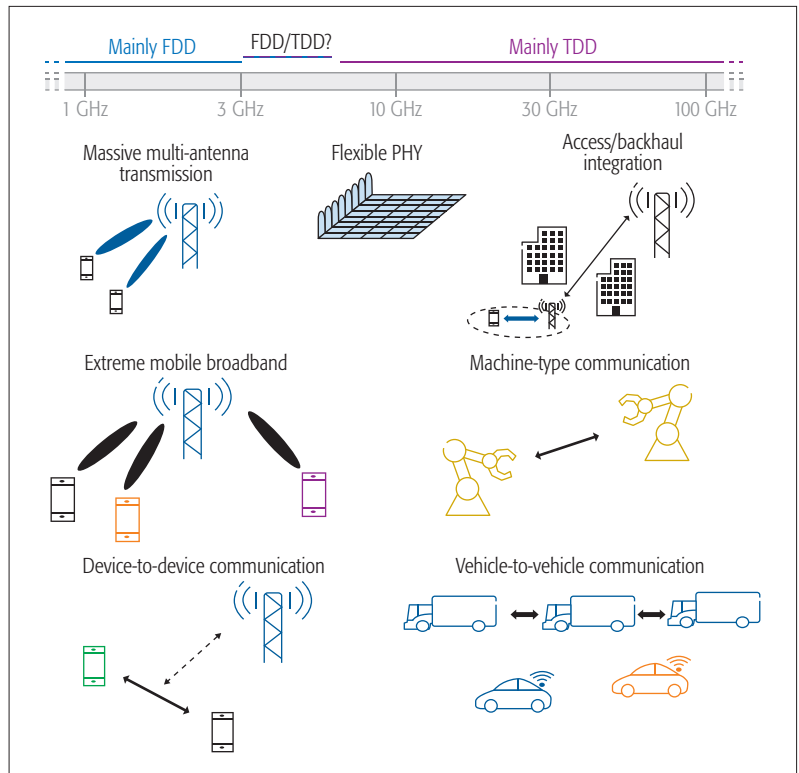


Figure 2. Massive MIMO, flexible physical layer, (mobile) self-backhaul operation in sub-1 GHz to 100 GHz with mainly TDD above 6 GHz to support diverse uses cases are the key features of the 5G radio access.

DL; however, the requirements are even more stringent for backhaul (due to the large amount of data). Vehicular communication also requires very high spectral efficiency in dense urban scenarios when the system is capacity limited and the large number of vehicles are periodically broadcasting signals in an asynchronous fashion.

MIMO Compatibility: OFDM enables a straightforward use of MIMO technology. With the increase in carrier frequency, the number of antenna elements will increase in the access nodes (base stations) as well as in the devices. The use of various MIMO schemes will be essential in providing high spectral efficiency (by enabling SU-MIMO/MU-MIMO) and greater coverage (via beamforming). Beamforming will

² D2D link is referred to as sidelink in 3GPP.

Channel frequency-selectivity is always relevant to the transmission of large bandwidth signals over wireless channels. Channel frequency selectivity depends on various factors such as type of deployment, beamforming technique, and signal bandwidth. OFDM is robust to frequency selective channels.

Performance indicators	OFDM assessment	DL req.	UL req.	Sidelink req.	V2X req.	Backhaul req.
Spectral efficiency	High	Very high	Very high	High	Very high	Very high
MIMO compatibility	High	Very high	Very high	High	Very high	Very high
Time localization	High	High	High	High	Very high	Very high
Transceiver baseband complexity	Low	Very high	High	Very high	High	High
Flexibility/scalability	High	High	High	High	High	High
Robust. to frequency selective channel	High	High	High	High	High	High
Robust. to time selective channel	Medium	High	High	High	Very high	Low
Robust. to phase noise	Medium	High	High	High	High	High
Robust. to sync. errors	High	Medium	Medium	High	High	Medium
PAPR	High (can be reduced)	Low	High	High	Medium	Low
Frequency localization	Low (can be improved)	Medium	Medium	Medium	Medium	Low

Table 1. Assessment of OFDM.

be instrumental in overcoming high propagation losses at very high frequencies (coverage limited scenarios);

Peak-to-Average-Power-Ratio: OFDM has high PAPR (like other multi-carrier waveforms). A low PAPR is essential for power-efficient transmissions from devices (e.g., UL, sidelink). Low PAPR becomes even more important at very high frequencies. It is noteworthy that small low-cost base stations are envisioned at high frequencies; therefore, low PAPR is also important for DL. High PAPR in OFDM can also be substantially reduced via various well-known PAPR reduction techniques with only minor compromise in performance [5]. For NR, OFDM with PAPR reduction (without DFT precoding³) is an attractive option for UL and sidelink. The use of one waveform for all link types will also make transceiver designs and implementations symmetric for all transmissions. Moreover, it is important to note that the requirements on PAPR for UL and DL will become more similar in the future due to low-cost small base stations.

Robustness to Channel Time Selectivity: This is vital in high-speed scenarios, which are relevant in large cell deployments. The large cell deployments are not expected at very high frequencies due to harsh propagation conditions (coverage limitation). At very high frequencies, the deployments are expected in the form of small cells where mobility is not a major concern. However, V2X services may be enabled at very high frequencies, making robustness to channel time selectivity a very important performance indicator at very high frequencies. Traditionally, a backhaul link is fixed, and mobility is not a concern; however, for the envisioned mobile backhaul (e.g., access nodes on vehicles), robustness to channel time selectivity will become relevant. OFDM can be made robust to channel

time selectivity by a proper choice of subcarrier spacing.

Robustness to Channel Frequency Selectivity: Channel frequency selectivity is always relevant to the transmission of large bandwidth signals over wireless channels. Channel frequency selectivity depends on various factors such as type of deployment, beamforming technique, and signal bandwidth. OFDM is robust to frequency selective channels.

Robustness against Phase Noise: An OFDM system can be made robust to phase noise by a proper choice of subcarrier spacing. Phase noise robustness is crucial for all link types where a device (transmitter/receiver) is involved. In particular, low-phase noise oscillators may be too expensive and power consuming for devices. Phase noise robustness is also important for future low-cost base stations. Basically, any link that involves a device and/or low-cost base station puts a high requirement on phase noise robustness of waveform, especially if the communication takes place at high frequencies since phase noise increases with carrier frequency.

Transceiver Baseband Complexity: The baseband complexity of an OFDM receiver is lowest among all candidate waveforms that have been studied in the past for 5G RAT [7]. Baseband complexity is always very important for the devices, especially from the receiver perspective. For NR, complexity is even a major consideration for base stations, since a base station can be a small access node (especially at high frequencies) with limited processing capability. At very high frequencies and large bandwidths, the receiver may also have to cope with severe RF impairments.

Time Localization: OFDM is very well localized in the time domain, which is important to efficiently enable (dynamic) TDD and support latency-critical applications such as URLLC. Dynamic TDD is envisioned at high frequencies,

³ LTE uses DFT-spread OFDM (DFTS-OFDM) for both UL and sidelink due to its lower PAPR than OFDM. However, DFTS-OFDM has certain drawbacks compared to OFDM such as less flexibility for scheduling (in the case of SC-FDMA) and a more complex MIMO receiver with degraded link-level and system-level performance [6]. Since MIMO will also be a key component for UL and sidelink in NR, DFTS-OFDM is not a preferred option.

and provision of low latency is essential for all link types; in particular, backhaul and V2X links may impose very high requirements.

Frequency Localization: OFDM is less localized in the frequency domain. Frequency localization can be relevant to support coexistence of different services potentially enabled by mixing different waveform numerologies in the frequency domain on the same carrier. Frequency localization is also relevant if asynchronous access is allowed in UL and sidelink. In general, frequency localization of a waveform may not be important at high frequencies where a large amount of channel bandwidth is available.

Robustness to Synchronization Errors: The provision of cyclic-prefix in OFDM makes it robust to timing synchronization errors. Robustness to synchronization errors is relevant when synchronization is hard to achieve such as sidelink. It can also be relevant if asynchronous transmissions are allowed in the UL.⁴

Flexibility and Scalability: OFDM is a flexible waveform that can support diverse services in a wide range of frequencies by proper choice of subcarrier spacing and cyclic prefix (CP). Further discussion on OFDM numerology design that fulfills a wide range of requirements is given in the next section.

In Table 1, we provide a summary of OFDM assessment. An OFDM assessment of “High” in second column means that OFDM has good performance in general for the given KPI, whereas a link requirement “High” for a KPI tells that the given waveform KPI is important for the given link type in general. We assess D2D and V2X cases separately due to different levels of requirements. For example, V2X communication has higher requirements on mobility and system capacity, but lower requirements on power efficiency when compared to UE-to-UE communication. Based on the assessment in Table 1, we conclude that OFDM is an excellent choice for the NR air interface.

OTHER MULTI-CARRIER WAVEFORMS

In recent years, a number of multi-carrier and single-carrier waveforms have been investigated and proposed for 5G RATs. An assessment of these multi-carrier and single-carrier waveforms can be found in [7, 8] for all KPIs given above. Besides OFDM, the other major multi-carrier waveforms (filter bank multicarrier orthogonal quadrature amplitude modulation, FBMC-OQAM and FBMC-QAM) are based on filter bank implementations where each subcarrier is filtered. OFDM is well localized in time and less localized in frequency, whereas FBMC is less localized in time but well localized in frequency. The good time localization of OFDM along with its lower implementation complexity than FBMC makes OFDM the preferred choice for NR that has to support TDD, delay-critical use cases, and efficient processing of large bandwidth signals. If necessary, the frequency localization of OFDM can be improved via low complexity windowing [9, 10] or subband filtering. The windowing or filtering can be employed at either the transmitter or the receiver, or both transmitter and receiver. An example of transmitter and receiving windowing in OFDM is provided below.

SINGLE-CARRIER WAVEFORMS

Single-carrier waveforms can be useful at very high frequencies, where power-efficient transmission is desired. Among single-carrier waveforms, there are two main categories: discrete Fourier transform single-carrier OFDM (DFTS-OFDM), and pure single-carrier. Pure single-carrier waveforms can have very low PAPR and are inherently robust to phase noise and Doppler. However, they do not allow efficient and flexible spectrum resource utilization; require more complex receiver design due to lack of frequency domain equalization (if CP is not enabled); and have lower compatibility with MIMO and are less spectrally efficient in general. On the other hand, DFTS-OFDM offers better scheduling flexibility, allows low-complexity frequency domain equalization, and has higher compatibility with MIMO than pure single-carrier waveforms. DFTS-OFDM has lower PAPR than OFDM [11], but not as low as pure single-carrier waveforms. These properties make DFTS-OFDM an attractive option for UL and DL at very high frequencies, where low PAPR is desired.

OFDM NUMEROLOGIES FOR NR

NR is envisioned to operate from sub-1 GHz to 100 GHz for a wide range of deployment options and to support a variety of services. It is not possible for a single waveform numerology to fulfill all these requirements. Therefore, we propose to adopt a family of OFDM numerologies for NR air interface.

NUMEROLOGY DESIGN PRINCIPLES

A given carrier frequency, phase noise, and Doppler set requirements on the minimum subcarrier spacing. Use of smaller subcarrier spacings would result in either high error vector magnitude (EVM) due to phase noise or undesirable high requirements on the local oscillator. Too narrow subcarrier spacings also lead to performance degradations in high Doppler scenarios. Required CP overhead (and thus anticipated delay spread) sets an upper limit for the subcarrier spacing; selecting too large subcarriers would result in undesirable high CP overhead. The maximum fast Fourier transform (FFT) size of the OFDM modulator together with subcarrier spacing determines the channel bandwidth. Based on these relations the subcarrier spacing should be as small as possible, while still being robust against phase noise and Doppler and providing the desired channel bandwidth. Below we provide further discussion on the choice of sub-carrier spacing and CP taking into account phase noise effect and realistic delay spread at different carrier frequencies.

As discussed earlier, a set of OFDM numerologies has to be defined for NR to handle a wide range of frequencies and deployment options. These OFDM numerologies could be unrelated to each other, that is, OFDM numerology for a given frequency and deployment is only based on this frequency and deployment, not considering numerologies for other frequencies and deployments at all. Another possibility is to define a family of OFDM numerologies that are related to each other via scaling, that is,

NR is envisioned to operate from sub-1 GHz to 100 GHz for a wide range of deployment options and to support a variety of services. It is not possible for a single waveform numerology to fulfill all these requirements. Therefore, we propose to adopt a family of OFDM numerologies for the NR air interface.

⁴ We note that LTE only supports synchronous uplink transmission (except for PRACH), which is realized via timing advance at the UEs.

$$\Delta f_i = n_i \Delta f_{i-1}, \quad T_{cp,(i)} = \frac{T_{cp,(i-1)}}{n_i}, \quad (1)$$

where Δf_i and $T_{cp,(i)}$ denote subcarrier spacing

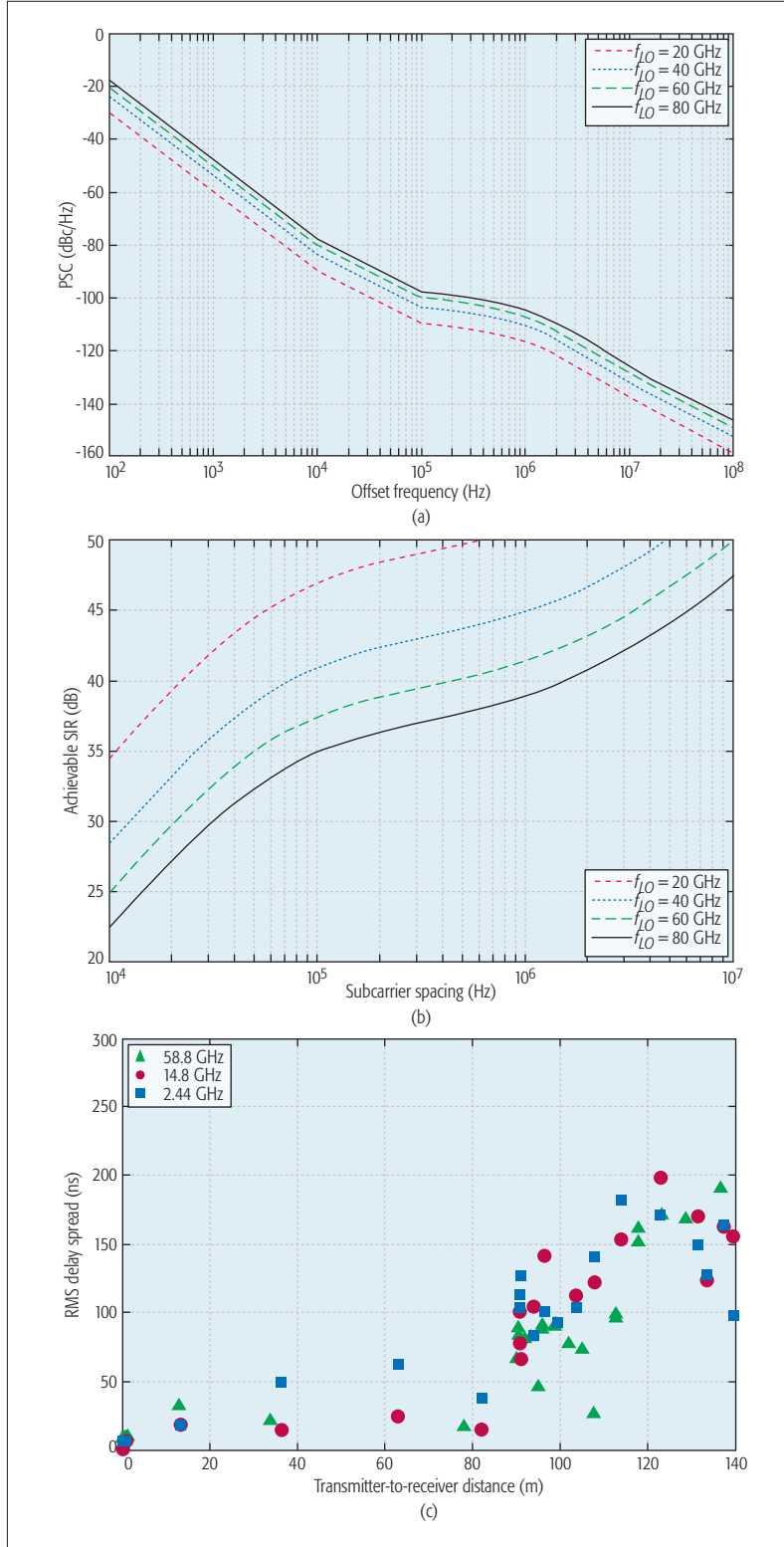


Figure 3. Phase noise power spectral densities at different oscillator frequencies: a) phase noise power spectral densities at different oscillator frequencies; b) achievable SIR subject to phase noise (due to inter-carrier interference) at different oscillator frequencies; c) channel delay spread has a weak dependence on carrier frequency.

and cyclic-prefix duration of the i th numerology and $n_i \in \mathbb{N}$ is a scaling factor. The duration of an OFDM symbol is inverse to subcarrier spacing. With this scaling approach, sampling clock rates of different OFDM numerologies relate to each other via the scaling factors $\{n_i\}$, which simplifies the implementation. We therefore propose to adopt this scaling approach, that is, OFDM numerologies are derived from a base OFDM numerology via the scaling. In principle, the scaling factors $\{n_i\}$ can be selected independent of each other; however, it is desirable that the scaling factors follow a certain relationship (given in Eq. 2), which will be discussed in the following.

We propose that the number of OFDM symbols per subframe should be equal for all numerologies, meaning that the subframe duration would shrink with the increase in subcarrier spacing. Maintaining an equal number of OFDM symbols per subframe for all numerologies simplifies scheduling and reference signal design. Furthermore, this would enable shorter latencies for wider subcarrier numerologies (to be used in high-frequency small cell deployments where some of the URLLC applications are envisioned). If an equal number of OFDM symbols are assumed for all numerologies, the following relationship holds for subframe durations between different numerologies:

$$T_{sf,(i)} = \frac{T_{sf,(i-1)}}{n_i} = \frac{T_{sf,(i-2)}}{n_i n_{i-1}} = \dots = \frac{T_{sf,(1)}}{\prod_{k=1}^i n_k}.$$

For adjacent TDD networks using different OFDM numerologies, it is desirable that an integer number of subframes from one OFDM numerology fits into one subframe of the other OFDM numerology to enable time aligned DL and UL periods. If the sub-frame durations of different numerologies do not fulfill the above condition, two neighboring TDD networks would require guard time in the frame structure to enable synchronous operation, which is not efficient resource utilization. Therefore, we propose that the scaling factors are chosen such that a subcarrier spacing is an integer divisible by all smaller subcarrier spacing, that is,

$$\Delta f_i = 2^{L(i)} \Delta f_1, \quad \forall i \in \{1, 2, \dots, M\}, \quad (2)$$

where $L(i) \in \mathbb{Z}$, M is the number of OFDM numerologies, and Δf_1 is sub-carrier spacing of the base numerology. This implies that the scaling factor in Eq. 1 should be chosen as $n_i = 2^L$, where L is an integer.

IMPACT OF PHASE NOISE AND CHANNEL DELAY SPREAD

Phase noise in an OFDM system causes two main effects: common phase error (CPE), and inter-carrier interference (ICI) [12]. CPE refers to phase rotation of all subcarriers by an equal amount and can be corrected easily through the use of pilot subcarriers. ICI is an additive noise (not always Gaussian) and usually hard to compensate for depending on how fast the phase variations are. In the following, we evaluate the effect of ICI in OFDM as a function of subcarrier spacing at different oscillator frequencies. First, we briefly describe the phase noise model

used in the evaluations and then present the evaluation results.

The local oscillator (LO) consists of a crystal oscillator (XO) and a voltage-controlled oscillator (VCO) connected in a phase locked loop (PLL). At low offset frequencies, the LO phase noise is dominated by the XO phase noise, shifted up by $20\log(f_{LO}/f_{XO})$. At high offset frequencies, the LO phase noise is dominated by the -20 dB/dec of the VCO. In the following evaluations, the considered LO design is based on XO running at 490 MHz and VCO with figure-of-merit⁵ (FOM) = -190 dB and a power consumption of 30 mW. With this design, the power spectral density (PSD) of the phase noise is given in Fig. 3a. The signal-to-interference ratio (SIR) due to ICI for a subcarrier can be computed according to the expression in [13, Sec. 5.2]. In Fig. 3b, we have evaluated the SIR of the middle subcarrier (suffering from highest ICI) as a function of subcarrier spacing for four different oscillator frequencies. According to Fig. 3b, 40 dB signal-to-noise ratio (SNR) can be achieved with $\Delta f = 30$ kHz at 20 GHz oscillator frequency, $\Delta f = 60$ kHz at 40 GHz oscillator frequency, and $\Delta f = 500$ kHz at 60 GHz oscillator frequency.

For a fixed CP overhead in an OFDM symbol, larger subcarrier spacing implies smaller CP. The CP has to be greater than the delay spread of the channel. Therefore, channel delay spread sets an upper limit on the subcarrier spacing. Some recent channel measurements at different carrier frequencies (2.44 GHz, 14.8 GHz, and 58.8 GHz) in a street microcell scenario have shown that delay spread is similar at different frequencies assuming omnidirectional antennas (Fig. 3c) [14]. Similar conclusions are made in a recent white paper [15], which shows that delay spread has a weak dependence on frequency. Furthermore, it has been observed that delay spread is much lower in line of sight (LOS) conditions compared to non-LOS (NLOS) conditions. According to Fig. 3c, the maximum value of root mean square (RMS) delay spread is $0.2 \mu\text{s}$, which is important to keep in mind while setting the upper limit on subcarrier spacing. It is also important to note that the observed delay spread of the channel depends on a few other factors such as deployment scenario and beamforming. Delay spread is usually smaller in indoor environments, and use of narrow beams may reduce delay spread as well.

PROPOSED NUMEROLOGIES

We now propose a set of OFDM numerologies following the design principles discussed previously and the important observations made earlier related to impact of phase noise and channel delay spread at different carrier frequencies. We choose LTE numerology as the base numerology, that is, $\Delta f_1 = 15$ kHz, $T_{ofdm,(1)} = 66.67 \mu\text{s}$, and $T_{cp,(1)} = 4.69 \mu\text{s}$. The other numerologies are derived from the base numerology according to Eqs. 1 and 2. The derived numerologies are given in Table 2. We note that in LTE, CP duration of the first OFDM symbol in a slot is $5.2 \mu\text{s}$. We propose the same for NR base numerology (which is LTE numerology), although this is not explicitly mentioned in Table 2. Moreover, LTE provides an option for extended CP, which should also exist in NR. As proposed in Table

OFDM parameters	Up to 6 GHz	Up to 20 GHz	Up to 40 GHz	Above 40 GHz
Subcarrier spacing	15 kHz	30 kHz	60 kHz	$2^L \times 60$ kHz
Clock frequency	61.44 MHz	122.88 MHz	245.76 MHz	$2^L \times 245.76$ MHz
Samples per OFDM symbol	4096	4096	4096	4096
OFDM symbol duration	66.77 μs	33.33 μs	16.67 μs	$16.67/2^L \mu\text{s}$
CP samples	288	288	288	288
CP duration	4.69 μs	2.35 μs	1.17 μs	$1.17/2^L \mu\text{s}$

Table 2. Proposed OFDM Numerologies.

2, different numerologies are suitable for different frequency ranges considering the achievable SNR subject to phase noise and channel delay spread discussed previously. According to Fig. 3c, CP duration must be greater than $0.2 \mu\text{s}$, which implies $L = 3$ in the last column of Table 2, meaning that the largest subcarrier spacing should be 480 kHz if the numerologies are derived according to Eq. 2.⁶ We recall that in the presence of ICI (due to phase noise), 480 kHz subcarrier spacing can achieve approximately 40 dB SNR at 60 GHz oscillator frequency and approximately 35 dB SNR at 80 GHz oscillator frequency (Fig. 3b).

There are a few important reasons for proposing LTE numerology as the base numerology.

- 3GPP has specified LTE numerology for narrowband Internet-of-Things (NB-IOT). NB-IOT devices are designed to operate for 10 years or more on a single battery charge. Once such an NB-IOT device is deployed, it is likely that within the device lifetime the embedding carrier gets refarmed to NR.

- NR deployments can happen in the same band as LTE. With adjacent carrier LTE TDD, NR must adopt the same UL/DL switching pattern as LTE TDD does. Every numerology where (an integer multiple of) a subframe is 1 ms can be aligned with regular subframes in LTE. In LTE, duplex switching happens in a special subframe. To match the transmission direction in special subframes, the same numerology as in LTE is needed.

- LTE Release 8 was standardized after a thorough numerology study; therefore, it is reasonable to aim for similar numerology at LTE-like frequencies and deployments.

FRAME STRUCTURE

In LTE, one radio frame comprises 10 subframes, and each subframe consists of two slots with seven OFDM symbols per slot. The notion of a slot may not be necessary; therefore, we only define the subframe for NR. The proposed subframe consists of N_{symb} OFDM symbols, but not all symbols in a subframe are always used for active transmission. We define two basic subframe types, one for UL and one for DL. Transmission in a DL subframe always starts at the beginning of the subframe and can extend from 0 up to at most N_{max} OFDM symbols. Transmission in a UL subframe always stops at the end

⁵ FOM has been defined according to (1) in [12].

⁶ In practice, maximum delay spread is typically four to five times greater than RMS delay spread and CP should be chosen accordingly.

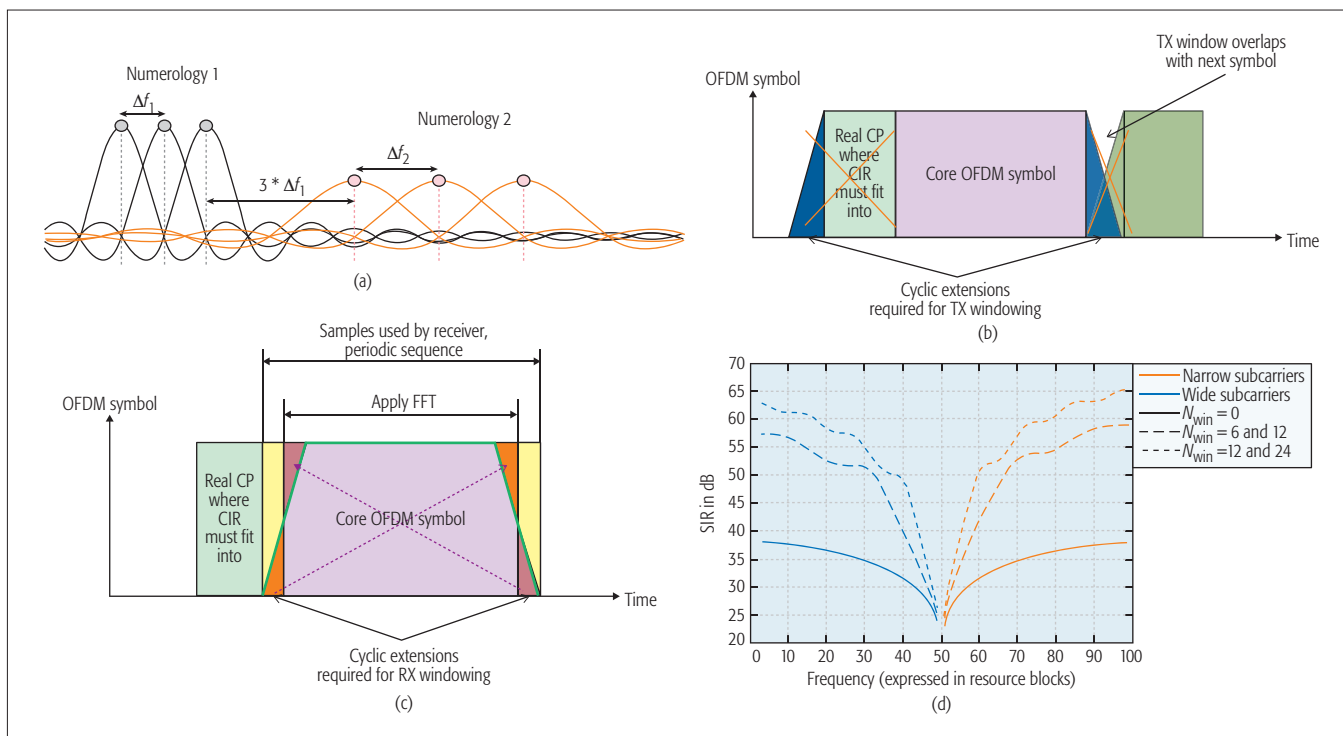


Figure 4. Transmitter and/or receiver windowing is an attractive option if different OFDM numerologies are mixed on the same carrier: a) inter-numerology interference; b) an illustration of transmitter windowing; c) an illustration of receiver windowing; d) windowing reduces inter-numerology interference.

of the subframe and can extend from 0 up to at most N_{max} OFDM symbols. The gaps between DL and UL transmission, if present, are used as guard in TDD for transmission in the reverse direction within a subframe.

The duration of a single subframe has to be very short. Depending on the numerology, a subframe duration can be tens of microseconds to a few hundred microseconds. For the OFDM numerologies given in Table 2, we propose seven OFDM symbols per subframe, that is, $N_{max} = 7$. This implies subframe duration of 500 μ s for 15 kHz numerology, 250 μ s for 30 kHz numerology, 125 μ s for 60 kHz numerology, and reaching down to 15.62 μ s for 480 kHz subcarrier spacing. Very short subframes are important for URLLC applications requiring low latency, and such devices will typically check for control signaling transmitted at the beginning of every DL subframe. Given the latency-critical nature, the transmission itself can also be very short (e.g., a single subframe). For eMBB devices, extremely short subframes are typically not needed. Therefore, one can aggregate multiple subframes and schedule the subframe aggregate using a single control channel.

MIXING NUMEROLOGIES

For some use cases, mixing different numerologies on the same carrier frequency may be beneficial, for example, to support different services with very different latency requirements. In an OFDM system with different numerologies (subcarrier bandwidth and/or CP length) multiplexed in the frequency domain, only subcarriers within a numerology are orthogonal to each other. Subcarriers of one numerology interfere with subcarriers of another numerology, since energy leaks

outside the subcarrier bandwidth and is picked up by subcarriers of the other numerology. The inter-numerology interference is illustrated in Fig. 4a, where a numerology based on subcarrier spacing Δf_1 interferes with another numerology based on subcarrier spacing Δf_2 , even though there is a small guard band between the two transmissions.

The inter-numerology interference can be reduced by either applying time-domain filtering per numerology (sub-band) or time-domain windowing. In the following, we consider the windowing approach due to its low-complexity implementation and superior performance [10].

TRANSMITTER WINDOWING

The main reason for the slow decay of OFDM spectrum is signal discontinuities at OFDM symbol boundaries. With transmitter windowing, the boundaries of each OFDM symbol are multiplied with a smooth slope in the time domain, increasing smoothly from 0 to 1 (increasing slope) or 1 to 0 (decreasing slope) (Fig. 4b). The increasing slope is applied at the beginning of the CP, while the decreasing slope is applied after the end of the core OFDM symbol within an extra added cyclic suffix. Figure 4b also shows that the increasing slope of the next OFDM symbol overlaps with the decreasing slope of the previous OFDM symbol. Since the receiver only keeps the samples of the core OFDM symbol, transmitter windowing is transparent to the receiver.

RECEIVER WINDOWING

A standard OFDM receiver cuts out the desired OFDM symbol period by applying a rectangular window in the time domain to the received sig-

nal and subsequently applies an FFT. Application of a rectangular window in the time domain corresponds to convolution in the frequency domain with a sinc function. The sinc-like function leads to high interference pickup from adjacent non-orthogonal signals such as OFDM signals with other numerologies. To reduce interference pickup, the rectangular window must be replaced by a smooth window function. To this end, a smooth increasing window slope is applied at the boundary between the CP and the core OFDM symbol (half within each); a decreasing smooth window slope is applied at the boundary between the core OFDM symbol and the added cyclic suffix (Fig. 4c). If the applied window slopes fulfil the Nyquist criteria (i.e. they are center asymmetric), the signal part cut away by the decreasing windowing slope (indicated by the upper right orange triangle in Fig. 4c) is the same as the remaining signal part after application of the increasing window slope within the CP (indicated by the lower left orange triangle in Fig. 4c) since the CP is a copy of the last part of OFDM symbol. If the windowed CP part (lower left orange triangle in Fig. 4c) is added to the last part of the core OFDM symbol, the core OFDM symbol is restored at its second boundary. The core OFDM symbol can also be restored at the first symbol boundary by applying the same trick. Now the complete OFDM is restored, and subcarriers are orthogonal again. The FFT is applied to the restored core OFDM symbol as indicated in Fig. 4c. Interference pickup remains reduced as long as the interference does not have a periodicity equal to the OFDM symbol duration.

In Fig. 4d, we show the effect of transmitter and receiver windowing on inter-numerology interference assuming 15 kHz and 30 kHz numerologies (Table 2) multiplexed in the frequency domain. It can be observed that windowing substantially increases the achievable SIR. (SIR is averaged across subcarriers within one resource block, which is assumed to have 12 subcarriers.) Windowing has extremely low complexity. Only the windowed samples are scaled, and overlap-and-add over the windowed periods is performed.

CONCLUSIONS

We propose a symmetric physical layer for all link types (e.g., UL, DL, sidelink, backhaul link) based on OFDM with scalable numerology. OFDM is assessed for a number of performance indicators, link types, and frequency ranges. We observe that OFDM is an excellent choice for all link types in NR, due to its high time localization, low-complexity transceiver design, high spectral efficiency, and easy integration with MIMO technologies. The main drawback of OFDM (as with all multi-carrier waveforms) is its high PAPR, which can be a limitation at very high frequencies. There are well-known methods to reduce PAPR of OFDM with minor degradation in performance. OFDM with PAPR reduction can be particularly useful for UL and sidelink. For very high frequencies, DFTS-OFDM may also be an interesting waveform due to its low PAPR and frequency domain equalization. However, further

investigations are necessary to discover if DFT precoding is necessary at very high frequencies.

We propose a family of OFDM numerologies considering implementation complexity, phase noise robustness, and realistic channel delay spreads at different carrier frequencies. The proposed family of numerologies consists of a base numerology, and the remaining numerologies in the family are derived by scaling up the subcarrier spacing and scaling down the cyclic prefix of the base numerology by the same factor. The scaling approach is simple to implement. Enabling different numerologies merely requires scaling of the sampling clock frequency without changing any other waveform (OFDM) parameter. Furthermore, the preferred option for the numerology scaling factor is 2^L times the base numerology, where L is an integer. Such scaling is important to allow two neighboring TDD networks to enable two different numerologies without any resource waste (i.e., without using guard time). The preferred choice for the base numerology is LTE numerology for various reasons. The most important reason is coexistence with NB-IOT, for which 3GPP has already specified LTE numerology. Finally, we show that if different numerologies are multiplexed on the same carrier, the low-complexity (transmitter/receiver) windowing of OFDM can be useful to significantly suppress inter-numerology interference.

ACKNOWLEDGMENT

The research leading to these results received funding from the European Commission H2020 Programme under grant agreements n671650 (5G PPP mmMAGIC project) and n671680 (5G PPP METIS-II project).

REFERENCES

- [1] S. Methley *et al.*, "5G Candidate Band Study: Study on the Suitability of Potential Candidate Frequency Bands above 6 GHz for Future 5g Mobile Broadband Systems," Quotient Associates Ltd, tech. rep., Mar. 2015.
- [2] J. Luo, *et al.*, "Millimeter-Wave Air-Interface for 5G: Challenges and Design Principles," *Proc. ETSI Wksp. Future Radio Technologies – Air Interfaces*, 2016, pp. 1–6.
- [3] RP-160671, "Study on NR New Radio Access Technology," 3GPP TSG RAN Meeting 71, Mar. 2016.
- [4] T. Hwang *et al.*, "OFDM and its Wireless Applications: A Survey," *IEEE Trans. Vehic. Tech.*, vol. 58, no. 4, May 2009, pp. 1673–94.
- [5] D.-W. Lim, S.-J. Heo, and J.-S. No, "An Overview of Peak-to-Average Power Ratio Reduction," *J. Commun. and Networks*, vol. 11, no. 3, June 2009, pp. 229–39.
- [6] J. Zhang *et al.*, "Comparison of the Link Level Performance Between OFDMA and SC-FDMA," *Proc. Communications and Networking in China (ChinaCom)*, Oct. 2006, pp. 1–6.
- [7] A. A. Zaidi *et al.*, "A Preliminary Study on Waveform Candidates for 5G Mobile Radio Communications above 6 GHz," *Proc. IEEE VTC-Spring*, May 2016.
- [8] A. A. Zaidi *et al.*, "Evaluation of Waveforms for Mobile Radio Communications Above 6 GHz," *Proc. IEEE GLOBECOM 2016*.
- [9] S. H. Muller-Weinfurter, "Optimum Nyquist Windowing in OFDM Receivers," *IEEE Trans. Commun.*, vol. 49, no. 3, Mar 2001, pp. 417–20.
- [10] E. Bala, J. Li, and R. Yang, "Shaping Spectral Leakage: A Novel Low-complexity Transceiver Architecture for Cognitive Radio," *IEEE Vehic. Tech. Mag.*, vol. 8, no. 3, Sept. 2013, pp. 38–46.
- [11] G. Huang, A. Nix, and S. Armour, "Impact of Radio Resource Allocation and Pulse Shaping on PAPR of SC-FDMA Signals," *Proc. IEEE 18th Int'l. Symp. Personal, Indoor and Mobile Radio Commun.*, Sept 2007, pp. 1–5.
- [12] J. Stott, "The Effects of Phase Noise in COFDM," *EBU Tech. Rev.*, 1998.
- [13] L. Fanori and P. Andreani, "Highly Efficient Class-C CMOS VCOs, Including a Comparison with Class-B VCOs," *IEEE J. Solid-State Circuits*, vol. 48, no. 7, July 2013, pp. 1730–40.
- [14] R1-160846, "Street Microcell Channel Measurements at 2.44, 14.8 & 58.68 GHz," 3GPP TSG-RAN WG1 No. 84, Feb. 2016.
- [15] 5GCM White Paper, "5G Channel Model for Bands Up to 100 GHz," <http://www.5gworkshops.com>, Dec. 2015.

The proposed family of numerologies consists of a base numerology and the remaining numerologies in the family are derived by scaling up the subcarrier spacing and scaling down the cyclic-prefix of the base numerology by the same factor. The scaling approach is simple to implement.

BIOGRAPHIES

ALI A. ZAIDI received M.Sc. and Ph.D. degrees in telecommunications from KTH Royal Institute of Technology, Sweden, in 2008 and 2013, respectively. He is a senior researcher at Ericsson, driving research and standardization of future radio access technologies at Ericsson Research as well as international fora such as 3GPP and 5GPPP. He currently holds a task leadership role in the mmMAGIC project, co-funded by the European Commission's 5GPPP program. His research contributions include 40+ peer reviewed research papers, 10+ filed patents, and a book chapter on information and control in networks.

ROBERT BALDEMAIR received his Dipl.Ing. and Dr. degree from the Vienna University of Technology in 1996 and 2001, respectively. In 2000 he joined Ericsson, where he was initially engaged in research and standardization of digital subscriber line technologies ADSL and VDSL. Since 2004 he has been working on research and development of radio access technologies for LTE, and since 2011 with wireless access for 5G. Currently he holds a master researcher position at Ericsson. In 2014 he and colleagues at Ericsson were nominated for the European Inventor Award, the most prestigious inventor award in Europe, for their contribution to LTE.

HUGO TULLBERG received his M.Sc. degree in electrical engineering from Lund University, and his Ph.D. degree in electrical engineering, communication theory, and systems from the University of California at San Diego (UCSD). He was the Technical Manager of the EU FP7 project METIS. He is currently a master researcher at Ericsson Research, where he works with 5G communication systems. His research interests include communication and information theory, machine learning and inference systems, cognitive radio, and ad hoc networking.

HAKAN BJÖRKEGREN received his M.S. degree in computer science from Luleå University of Technology, Sweden, in 1989 and his Ph.D. degree in signal processing from the same university in 1995. Since then he has been working on radio, physical layer protocols, algorithms and base station architectures in various positions at Ericsson AB. His research interests include signal processing algorithms, protocols, physical layer design and optimization, and link evaluation.

LARS SUNDSTRÖM received his Ph.D. in applied electronics from Lund University, Sweden, in 1995. From 1995 to 2000 he was an associate professor in the Competence Center for Circuit Design at the same university, where his research focused on linear radio transmitters and RF ASIC design. In 2000, he joined Ericsson Research, where he presently holds the position of senior specialist with interests ranging from RF, analog, and mixed signal IC design to radio architectures for cellular transceivers.

JONAS MEDBO is currently a senior specialist in applied propagation at Ericsson Research, Sweden. He received his Ph.D. degree in particle physics from Uppsala University, Sweden, in 1997. Since 1997 he has been with Ericsson Research focusing on propagation research. He has contributed to widely used channel models like Hiperlan/2 and 3GPP SCM, and is currently focusing on 5G channel measurements in the range 0.5 to 100 GHz and modeling for 3GPP and ITU.

CANER KILINC has been a member of the Wireless Access Network research group at Ericsson Research Sweden and currently serves as an experienced researcher on 5G design & development and standardizations project. He has numerous patents, and journal and conference papers. He is also representing Ericsson Research in the Metis II European Project and contributing on 5G protocols and architecture. He carried out his Master's studies in computer science and engineering with a specialization in mobile systems at Luleå University of Technology, Sweden, and he obtained his B.Sc. degree in mathematics from Eskisehir Osmangazi University, Turkey, in 2007.

ICARO DA SILVA is a senior researcher working at Ericsson, Stockholm, since 2010. He is currently working as a 3GPP delegate in RAN2 Working Group driving 5G New Radio (NR) topics such as mobility and RRC states. In recent years he has been driving both internal and external 5G architecture initiatives in R&D. He was work package leader in the 5G-PPP project METIS-II responsible for the Overall Control Plane (CP) Design, and he has been an active contributor to the 5G Architecture Working Group in 5GPPP. He is also a co-author of the chapter about 5G architecture in the book *5G Mobile and Wireless Communications Technology* and an author of many conference/journal papers on 5G. He is also an active inventor in the area of mobile networks, architecture, and network management with more than 100 patents. He received his Master of Science from the Universidade Federal of Ceará, Brazil, in 2007 and since then has been involved in R&D projects with Ericsson.

Generalized DFT-Spread-OFDM as 5G Waveform

Gilberto Berardinelli, Klaus I. Pedersen, Troels B. Sørensen, and Preben Mogensen

ABSTRACT

This article introduces G-DFT-s-OFDM as a potential 5G waveform candidate. G-DFT-s-OFDM replaces the CP with a sequence having a tunable length; this sequence is part of the IFFT output rather than being appended to it. Benefits in terms of flexibility, spectral containment, low latency, and robustness to Doppler spread and phase noise are discussed in the article.

INTRODUCTION

Fifth generation (5G) radio access technology is expected to take a huge leap compared to the previous radio generations by supporting machine type communication (MTC), the so-called Internet of Things (IoT), besides traditional mobile broadband (MBB) access [1]. In particular, massive machine type communication (MMC) scenarios may feature a large amount of low-cost devices, for example, sensors, which transmit sporadically and are only coarsely synchronized to the network [2]. Conversely, mission-critical communication (MCC) services, such as car-to-car communication or closed loop control in factory automation, demand ultra-high reliability and low latency (e.g., below 1 ms) [3]. The necessity of supporting the new services targeted by 5G has brought researchers to question the suitability of the orthogonal frequency-division multiplexing (OFDM) waveform as adopted in Long Term Evolution (LTE) [4]. In particular, the demerits of OFDM are recognized to be its large out-of-band emissions, which affect the coexistence of asynchronous services or devices, as well as its sensitivity to phase noise and hardware impairments and the large peak-to-average power ratio (PAPR). The discrete Fourier transform spread-OFDM (DFT-s-OFDM) waveform, standardized in LTE uplink, emulates a single-carrier transmission and has the advantage of a lower PAPR, but does not overcome the other OFDM demerits. A further inefficiency of the LTE waveforms is the insertion of a cyclic prefix (CP) to counteract the multipath effect and enable one-tap frequency domain equalization; the CP introduces additional overhead and negatively impacts the flexibility of the system design.

Significant research effort is currently spent on the design and analysis of alternative solutions aimed at overcoming the inefficiencies of the LTE waveforms. Candidate multicarrier solutions rely on different degrees of filtering for obtaining lower

side lobes than OFDM, thus improving robustness to asynchronous transmission. For example, in filter bank multicarrier (FBMC) approaches, filtering is applied at each subcarrier [5], while in universal filter multicarrier (UFMC), blocks of subcarriers are filtered [6]. Generalized frequency-division multiplexing (GFDM) applies, at each subcarrier, a filter that is circularly shifted in time and frequency; this breaks subcarrier orthogonality, calling for efficient intercarrier interference suppression at the receiver [7]. Even though the mentioned solutions offer undeniable advantages in terms of spectrum containment (as well as further benefits such as reduced overhead), their suitability as 5G waveforms when considering complexity constraints as well as realistic impairments is still disputable. In particular, the extension to multiple-input multiple-output (MIMO) antenna transmission with multiple streams seems rather cumbersome.

This article presents a different perspective on the waveform design for 5G. While recognizing the disadvantages of OFDM/DFT-s-OFDM in supporting diverse services as envisioned by 5G, our hypothesis is that by applying simple enhancements on the DFT-s-OFDM waveform, it becomes possible to deal with the 5G challenges while maintaining similar complexity as in LTE. We refer to this waveform as generalized DFT-s-OFDM (G-DFT-s-OFDM). It can also be seen as a generalization of the zero-tail (ZT) DFT-s-OFDM concept presented in previous contributions [8]. We show how a system employing the G-DFT-s-OFDM waveform can benefit from the time domain granularity of a single-carrier transmission while maintaining the computational advantages of fast Fourier transform (FFT)-based processing; in that respect, G-DFT-s-OFDM can be considered a suitable candidate for a 5G waveform. Throughout this article, we refer to OFDM/DFT-s-OFDM as reference waveforms, highlighting how the proposed concept allows overcoming their demerits. A qualitative comparison with other waveform candidates is also provided.

G-DFT-s-OFDM DESIGN

SIGNAL GENERATION AND DETECTION

The generation of a G-DFT-s-OFDM symbol is depicted in Fig. 1a. The transceiver structure is identical to the traditional DFT-s-OFDM one, excluding the CP insertion, which is absent. The input vector at the DFT is composed of two known sequences, s_h and s_t , located at the head

The authors introduce G-DFT-s-OFDM as a potential 5G waveform candidate. G-DFT-s-OFDM replaces the CP with a sequence having a tunable length; this sequence is part of the IFFT output rather than being appended to it. They discuss the benefits in terms of flexibility, spectral containment, low latency, and robustness to Doppler spread and phase noise.

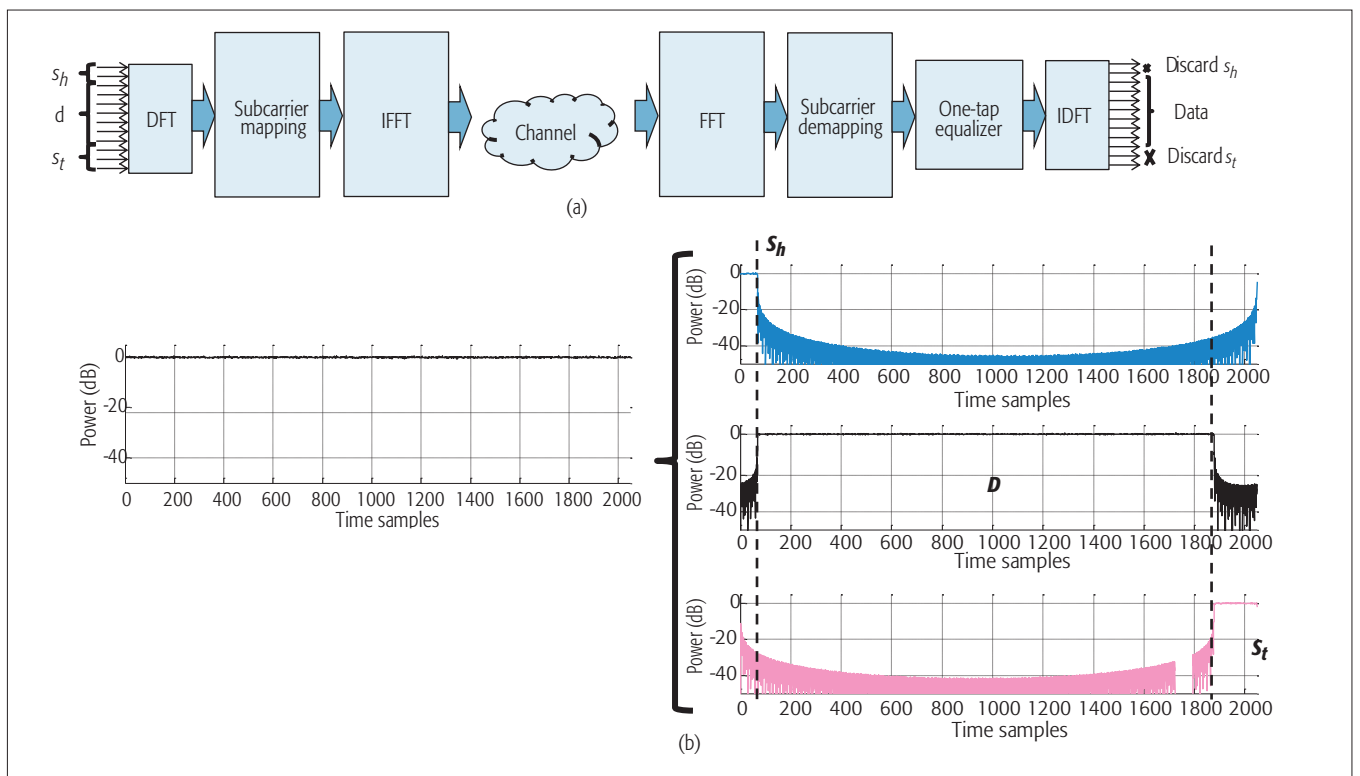


Figure 1. a) G-DFT-s-OFDM transceiver; b) composition of the transmit signal.

and tail, and a data part d in the middle. Different from the data part, the sequences s_h and s_t are set to be identical at consecutive symbols. A possible design for s_h and s_t is discussed later. The ZT DFT-s-OFDM waveform is obtained for the special case where the s_h and s_t are zero vectors. The input vector is converted to frequency domain by a DFT, and the samples are mapped over the used subcarriers. The time domain symbol is generated by an inverse FFT (IFFT).

Figure 1b shows the final time domain symbol and also highlights the contribution of the three portions (S_h , S_t , D) the linear combination of which leads to the time domain symbol.¹ Given the *quasi*-single-carrier nature of DFT-s-OFDM, the most significant energy of the three input sequences is indeed mapped over their respective parts of the symbol. Note that S_h , S_t , and D only overlap in their low energy parts. In the presence of multipath propagation, the sequence S_t is intended to generate to the next symbol the same energy leakage component that the symbol itself experiences at its beginning (provided S_t is set to be longer than the delay spread of the channel). This is meant to emulate the cyclic property of the received signal as in traditional CP-based transmission: a necessary condition for enabling one-tap frequency domain equalization.

The role of the S_h sequence is to avoid a power regrowth of the data part D in the last portion of the symbol due to the cyclicity of the IFFT operation; in that respect, it can be set to be very short (e.g., 6 samples out of 2048 as simulated in [8]). Note that, different from in a CP-based solution, the signal is here only *quasi*-cyclical: the residual energy of the data part D in the last samples introduces some non-cyclical leakage to the next symbol. Nonetheless, such

residual intersymbol interference is generated by samples with magnitude significantly lower than that of the average signal power (25 dB lower as in Fig. 1b), and is therefore expected to have limited impact on the link performance.

Recently, an approach based on the inclusion of a redundant vector has been proposed with the aim of further reducing the power of the last samples of D (e.g., up to 40 dB lower than the average signal power) [9]. The redundant vector is multiplexed with the data samples at the DFT input, and is claimed to consume only 1 percent of the overall transmit energy. The price to pay is a significant computational complexity increase since the redundant vector is derived from the data vector at every symbol transmission.

At the receiver, the reverse operations are applied. The received signal is converted to the frequency domain by an FFT and demapped from the used subcarriers. One-tap channel equalization is applied, and the transmit symbol vector is retrieved after IDFT and removal of the samples corresponding to s_t and s_h . G-DFT-s-OFDM maintains the orthogonality of the frequency subcarriers and allows one-tap frequency domain processing for MIMO as well, thus avoiding the increase of computational burden of waveforms such as FBMC when applied in the spatial domain. In general, a G-DFT-s-OFDM transceiver has the same complexity as an LTE uplink transceiver as currently used in commercial user equipments (UEs) and operator base stations.

DEALING WITH CHANNEL DELAY SPREAD

The G-DFT-s-OFDM waveform copes with the time dispersion characteristics of the radio channel without a CP, thereby avoiding its inefficiencies. In current radio technologies the CP

¹ Note that upper case (S_h, S_t, D) refers to the time domain signals generated by s_h, s_t, d .

is hardcoded in the system numerology, leading to potential spectral efficiency losses in case of propagation over channels with shorter delay spread, or to a block error rate (BLER) increase if its length is insufficient. Further, the insertion of CP leads to poor coexistence among systems operating with different settings. This is shown in the upper part of Fig. 2. In OFDM/DFT-s-OFDM, a larger time dispersion requires a longer CP, but the duration of the IFFT output needs to remain the same to maintain the system numerology. This leads to different symbol durations as well as to a different number of symbols per frame. Cells adopting different CP durations then generate mutual asynchronous interference even when aligned at frame level. Computationally feasible interference-aware receivers such as interference rejection combining (IRC) and successive interference cancellation (SIC) are able to efficiently suppress synchronous interfering streams, but fail at their task in case of asynchronous interference. The G-DFT-s-OFDM case is shown in the lower part of Fig. 2; the length of the S_t sequence, which is meant to cope with the time dispersion of the channel, can be set according to the experienced delay spread and is “absorbed” within the symbol. In this way, G-DFT-s-OFDM allows decoupling the radio numerology from the channel characteristics, simplifying the system design and enabling the possibility of synchronizing radio links regardless of the cell size and experienced propagation conditions. In the case of large cells, the sequence S_t can also mitigate the different propagation delays of the served users, thus avoiding the usage of a timing advance (TA) procedure [10]. We also foresee that the usage of the same known sequence S_t at each symbol can be exploited to track the time variation of the channel by high-speed users.

G-DFT-s-OFDM subsumes a system design where the delay spread is estimated first, for example, based on initial sounding signals, and the corresponding length of the required S_t signaled to the device. In practice, a limited set of options for the S_t length may be adopted due to constraints in the size of the feedback message. In the case of time-division duplex (TDD) mode, the same S_t length can be set in both uplink and downlink, while in frequency-division duplex (FDD) mode, such length can be different, and an estimate of the downlink delay spread at the user is required. In a simpler implementation, the length of S_t can be set on a cell basis given an estimate of the root mean square delay spread in the cell. This may penalize users at the cell edge in the case of large cells and reduce the spectral efficiency of cell center users, although still offering larger granularity than existing CP-based standards (only two CP options have been defined in LTE [4]).

REFERENCE SEQUENCE DESIGN

It is known that channel equalization at the receiver is made possible with the insertion in the transmit signal of known reference sequences for estimating the channel response. A similar approach as in LTE uplink can be used in G-DFT-s-OFDM, where the reference sequence is mapped over an entire time symbol. Well-known sequences with attractive auto-correlation and cross-correlation properties (e.g., Zadoff-Chu

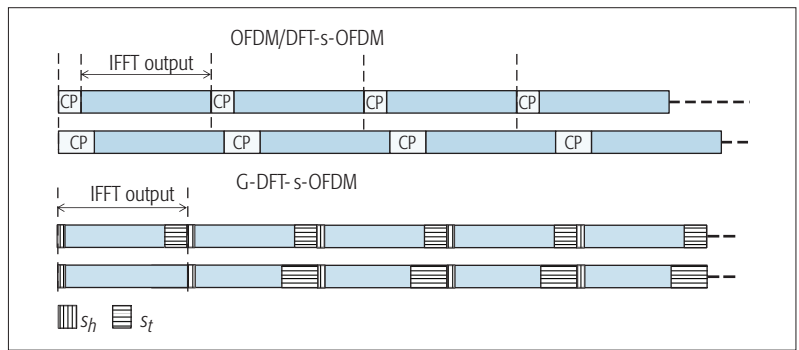


Figure 2. OFDM/DFT-s-OFDM transmission with different CP lengths vs. G-DFT-s-OFDM transmission with different S_t length.

[4]) can be mapped over the reference sequence symbol. In order to preserve the signal cyclic property, the sequences s_t and s_h in each data vector should be obtained as a copy of the first and last samples of the reference sequence vector. This approach exploits the degree of freedom offered by the presence of the s_t and s_h vectors to harmlessly accommodate the reference sequence in the radio frame. However, its suitability for the case of redundant vector insertion is still disputable.

It is possible to reduce overhead by transmitting the reference sequence over a shorter symbol at the expense of a reduction in frequency resolution. If its duration corresponds to a fraction $1/X$ of the original symbol duration, the subcarrier spacing is X times larger, and the resultant channel frequency response is then sampled with a lower resolution. Similar to the case of OFDM with frequency interleaved pilots, the channel response over all the subcarriers of the data symbol can be obtained by linear interpolation.

COMMON CLOCK WITH LTE

It has been widely argued that the minimum transmission time interval (TTI) in 5G is intended to be significantly lower than the LTE one (1 ms) due to the necessity of supporting latency-critical MCC applications. A minimum TTI duration of 0.2 ms has been justified for 5G in recent contributions (e.g., [11]). Future devices are likely to support different radio technologies, and chip manufacturers would benefit from a system that maintains a common clock with already supported standards. In that respect, it would be beneficial to design a 5G numerology based on the same reference clock as LTE.

However, when using CP-based transmission, the 5G TTI duration cannot be obtained by simply parsing the LTE TTI due to the usage of 14 symbols in 1 ms, which do not downscale linearly to the 0.2 ms. This requires the usage of a different subcarrier spacing and therefore a different sampling rate with respect to the LTE one. G-DFT-s-OFDM allows instead using the 0.2 ms TTI while maintaining identical clock rate as LTE. 6 symbols having duration $66.67 \mu\text{s}$ ($=1/15 \text{ kHz}$, where 15 kHz is the LTE subcarrier spacing) indeed cope perfectly with the 0.2 ms duration. Moreover, the same TTI duration also copes with an $N \times 15 \text{ kHz}$ subcarrier spacing, with $N \geq 1$, by increasing the number of symbols by a factor N . This eases the support of different radio standards in the same chip.

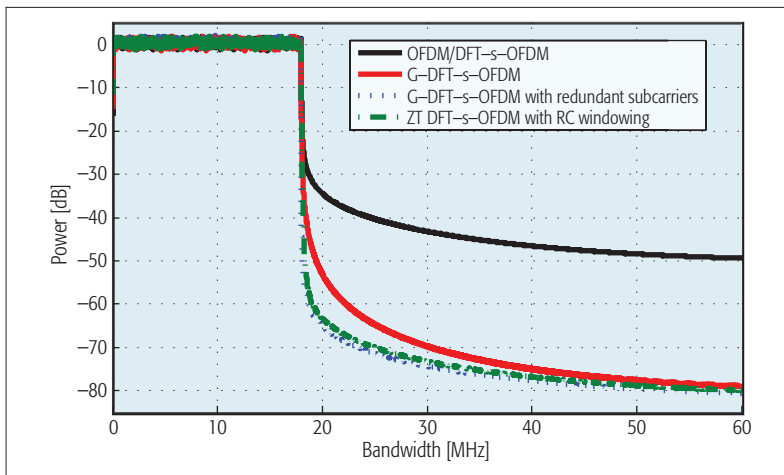


Figure 3. Out-of-band emissions of G-DFT-s-OFDM.

G-DFT-s-OFDM PROPERTIES

PAPR

Low PAPR is the main selling point of the DFT-s-OFDM waveform, and is preserved in G-DFT-s-OFDM. This makes the waveform particularly suited for transmission at mmWave frequencies. In such a spectrum region, the OFDM benefits in terms of granular frequency domain user multiplexing are not particularly relevant due to the availability of large bands, and it is therefore not worth accepting its large PAPR drawback. The aforementioned ZT DFT-s-OFDM, which can be seen as a special case of G-DFT-s-OFDM, suffers from a limited PAPR increase due to the presence of low power head and tail in the signal, while maintaining a significant margin over OFDM (e.g., around ~ 2.5 dB with respect to the ~ 3 dB of DFT-s-OFDM) [8].

FREQUENCY LOCALIZATION

It is known that asynchronous users operating over adjacent bands are likely to experience significant energy leakage from the neighbor bands in case the reference waveforms are used. The poor spectral containment of OFDM/DFT-s-OFDM is due to the rectangular time domain window, which introduces an abrupt transition between adjacent time symbols, leading to a sinc-shape of the frequency subcarriers. Raised cosine (RC) windowing is widely applied for smoothening such abrupt transitions. However, the RC windowing distorts the original signals and leads to intercarrier interference and therefore to a BLER penalty. In G-DFT-s-OFDM, the usage of the same sequences S_h and S_t at the beginning and end of each symbol smoothen the transition between adjacent symbols. This improves the spectral containment of the signal without creating any distortion to the natural output of the IFFT. The specific ZT DFT-s-OFDM case is particularly suited for the application of RC windowing to further smoothen the symbol transitions and benefit in terms of spectral containment. Since the windowing is applied over samples already having significant lower magnitude than the data samples, the distortion on the data part is expected to be minimal.

The out-of-band emission performance of G-DFT-s-OFDM and reference waveforms is reported in Fig. 3, assuming LTE configuration

parameters for a 20 MHz bandwidth. G-DFT-s-OFDM leads to up to 30 dB energy leakage reduction with respect to OFDM/DFT-s-OFDM. As expected, a further improvement is obtained with ZT DFT-s-OFDM with RC windowing. The case of G-DFT-s-OFDM with the redundant vector approach mentioned above is also included, and shown to provide similar benefits as the previous case. Although the results shown are still far from the performance of several FBMC schemes presented in the literature (e.g., [5]), they do not come at the expense of a significant increase of the computational complexity due to the filtering operation.

TIME LOCALIZATION

Restricting the allocation of a radio waveform within a limited time interval enables latency reduction mechanisms and also simplifies the synchronization procedure. OFDM/DFT-s-OFDM are well localized within the FFT window plus CP; the latter also allows correlation-based solutions for tracking and fine tuning of the received signal. Poor time localization is instead a recognized drawback of waveforms such as FBMC offset quadrature amplitude modulation (OQAM) where symbols are overlapping in time. Time localization of UFMC symbols can be slightly compromised by the inter-symbol interference in time dispersive channels [6]. G-DFT-s-OFDM localizes every symbol within an FFT window, while maintaining cyclicity at the receiver. Similar to CP-based solutions, the acquisition of the symbol timing can be obtained by correlating against a copy of the S_h and S_t sequences.

LINK PERFORMANCE

Since G-DFT-s-OFDM is a modified version of traditional DFT-s-OFDM, their link performance is similar. DFT-s-OFDM link performance is extensively addressed in the literature (e.g., [12]) and therefore not reported here. It is known that DFT-s-OFDM transmission over a frequency selective channel suffers from noise enhancement due to the IDFT operation at the receiver, which spreads the noise over the faded subcarrier across the entire data set. This leads to a BLER penalty with respect to OFDM. However, this performance gap tends to vanish with sufficient antenna diversity branches or when a turbo equalizer receiver is used. Another approach for improving the G-DFT-s-OFDM performance over frequency selective channels is to apply the DFT on a physical resource block (PRB) basis rather than across the entire bandwidth, as envisioned in a previous contribution [13]. If the PRB size is smaller than the coherence bandwidth of the channel, flat fading is experienced at each PRB, and link performance is not harmed. Applying the DFT on a PRB basis also allows frequency selective link adaptation. However, this comes at the expense of an increase of the PAPR.

An additional penalty of G-DFT-s-OFDM becomes visible at high signal-to-noise ratios (SNRs), that is, above 25 dB, where the impact of the energy leakage of the last samples of the data part D over the next symbol affects the cyclic property of the received signal [8]. In this case, high order modulation and coding schemes (MCSs) can be penalized. The usage of the aforementioned redundant vector approach significantly reduces the energy of the last sam-

ples of D , presumably improving the link performance at high SNR.

ADDRESSING THE 5G CHALLENGES

SUPPORT OF DIVERSE SERVICES AND HIGH SPEED

5G aims to support, in a flexible air interface, a set of diverse services that can be mapped over adjacent frequency bands. Since different services can significantly benefit from specific numerologies, recent research contributions have addressed the possibility of treating the subcarrier spacing as an additional degree of freedom in the operational mode [14]. For example, MBB users with high-quality transceivers can adopt a short subcarrier spacing, since the usage of longer time symbols reduces the relative overhead of CP (or S_T tail). Conversely, MCC users can benefit from a larger subcarrier spacing (and therefore shorter symbols) for the sake of latency reduction. Further, 5G will be operating over significantly higher carrier frequencies than LTE (ranging from fragmented bands below 6 GHz to large spectrum allocations at mmWave frequencies), and is expected to support user speed up to 500 km/h [15]. In that respect, a larger subcarrier spacing may also be beneficial for phase noise limited devices (e.g., for MMC) and high-speed users.

In traditional CP-based transmission, users adopting different subcarrier spacing will necessarily be asynchronous, even when aligned at a frame level: if they are operating over adjacent bands, they will be penalized by the poor spectral containment of the OFDM/DFT-s-OFDM waveform. The usage of a different subcarrier spacing also affects the coexistence between neighbor cells operating over the same frequency resources, since the resulting asynchronous interference cannot be canceled efficiently by interference suppression receivers. For instance, a low-speed broadband device operating with a short subcarrier spacing will not be able to efficiently suppress the interference coming from a neighbor cell serving a high-speed user operating with a large subcarrier spacing in the same frequency chunk.

G-DFT-s-OFDM allows supporting users with different subcarrier spacings while drastically reducing the OFDM/DFT-s-OFDM limitations. First, its improved spectral containment diminishes the residual interference to asynchronous devices transmitting in the neighbor bands [10]. Further, the absence of CP allows proportional scaling of the number of symbols per frame with the same increase/decrease factor for the subcarrier spacing. This scalability, combined with the quasi-single carrier nature of G-DFT-s-OFDM, improves the coexistence of time-aligned neighbor cells operating with different settings.

The principle is shown in Fig. 4. Base station B is serving a phase noise limited user (or a high-speed user) operating over the same frequency chunk where base station A is serving a static broadband user equipped with a high-quality transceiver. If the two base stations are time aligned, UE A1 will receive in its FFT window a combination of its desired symbol and an integer number of symbols transmitted by base station B: the multiple interfering symbols can be processed by UE A1 as a unique longer symbol, with an error vector magnitude (EVM) below ~ 1 percent. This is possible thanks to not only the absence of CP, but also the

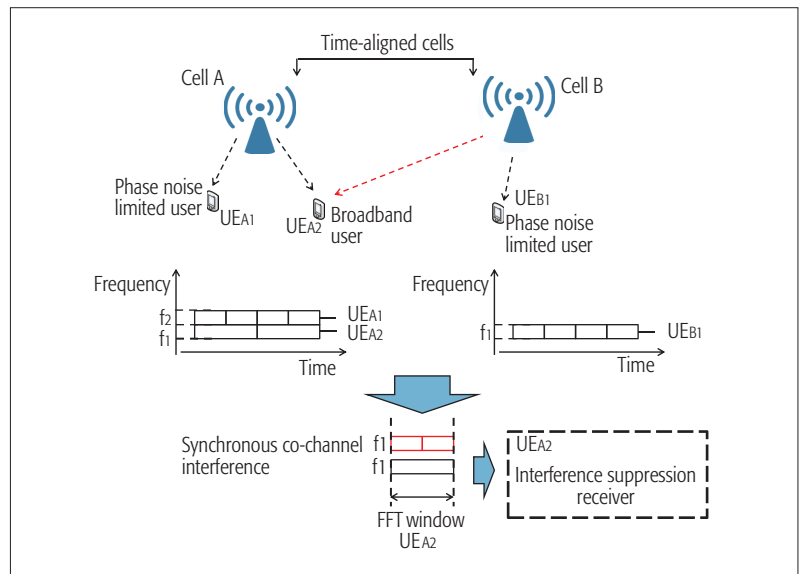


Figure 4. Coexistence of time aligned cells operating with different subcarrier spacing.

quasi-single carrier nature of the G-DFT-s-OFDM waveform, where data symbols are transmitted serially in time rather than in parallel as in traditional multicarrier techniques. UE A1 will then perceive a synchronous interference contribution by base station B. This can be efficiently suppressed by IRC or SIC receivers, provided specific arrangements on the reference sequences are taken such that both desired and interfering channel responses can be estimated. The underlying assumption is that signals from base station A and base station B are received with nearly the same delay; the proposed approach is then intended to work in cells with limited range and propagation delay not exceeding the S_T sequence duration. It is worth observing that the same does not strictly hold for UE B1. Its shorter FFT window only collects a fraction of the cell A long symbol; the received interfering signal is therefore not cyclic due to the presence of intersymbol interference. This compromises the possibility of properly suppressing it.

SUPPORT OF ULTRA-LOW LATENCY AND LOW POWER CONSUMPTION

Envisioned MCC applications such as car-to-car communication and closed loop control in industrial automation may require physical layer latencies at a sub-millisecond level [3]. Since the processing time is recognized as the bottleneck of the round-trip time budget, striving for extremely short transmissions is of fundamental importance for making the challenging 5G latency target a reality. In traditional CP-based systems, the minimum transmission sample corresponds to an OFDM/DFT-s-OFDM symbol, that is, around $\sim 71 \mu\text{s}$ duration in LTE. The multiplexing of users within the same OFDM/DFT-s-OFDM symbol is obtained with frequency domain scheduling, while time domain scheduling can only be applied by considering the entire OFDM/DFT-s-OFDM symbol as a minimum unit. Data processing can start only upon reception of an entire symbol, thus impacting the latency budget. Devices need to transmit at least for an entire OFDM/DFT-s-OFDM symbol even in the case of a minimum

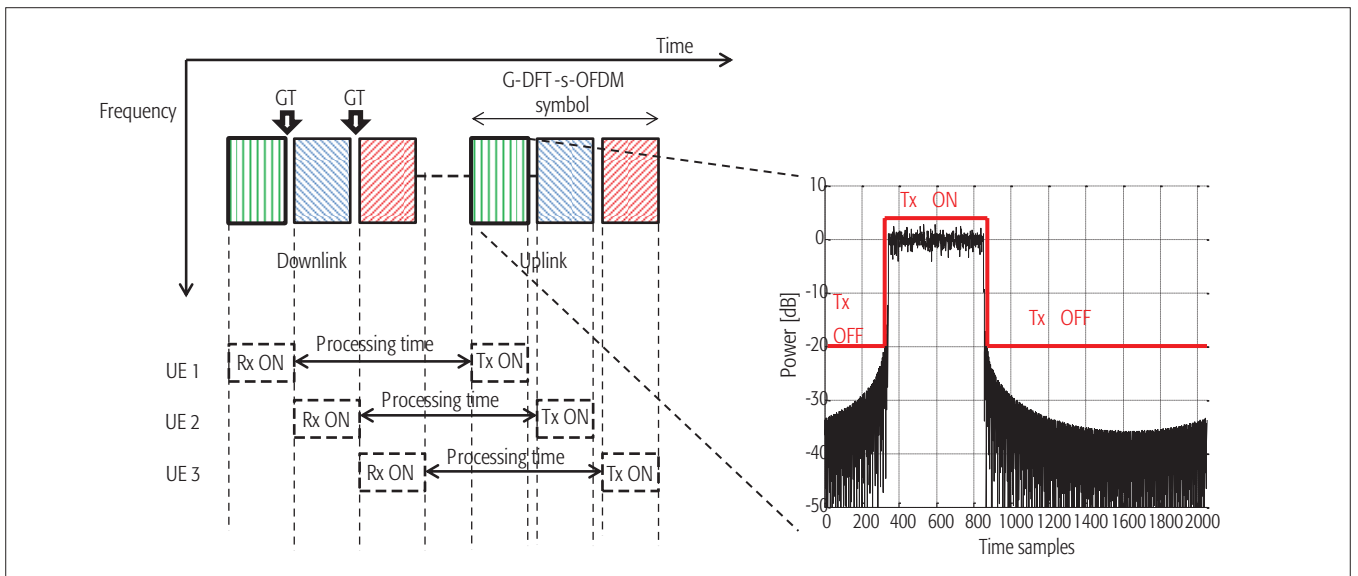


Figure 5. Reducing round-trip time and power consumption with G-DFT-s-OFDM.

amount of data (e.g., acknowledgment reports) or if they are transmitting the last bytes in their data queue. Obviously, the transmission time can be reduced by shortening the symbol duration. However, this translates to an increase of the subcarrier spacing. As mentioned in the previous section, a larger subcarrier spacing improves robustness to phase noise and Doppler spread, but may also affect the capability of the receiver to properly track the frequency selectivity of the channel for a correct equalization.

In G-DFT-s-OFDM data symbols are transmitted serially, thus enabling the possibility of transmitting only a portion of the IFFT output samples without compromising the subcarrier spacing. The principle is shown in Fig. 5. Three users are time domain scheduled on a sub-symbol basis, provided a sufficient guard time (GT) is inserted between them for accommodating the delay spread of the channel. At each receiving UE, the meaningful samples are zero-padded in the portions corresponding to the data of the other UEs, and detected as a G-DFT-s-OFDM symbol. The UEs are then able to process the data before receiving the entire set of samples. By anticipating the receiver processing, it is possible to reduce the time required for generating the acknowledgment message, thus shrinking the overall round-trip time. In the uplink, the s_i and s_j vectors at each user are set to be zero vectors, and their length is tuned such that the data part is allocated over a specific portion of the output time symbol. Only this portion is transmitted over the air. A further benefit is energy saving, since the devices can power on their transmitter/receiver chain only for a limited fraction of the symbol. Note that the same principle can also be used to embed in the symbol a dedicated guard period in TDD mode, thus avoiding its insertion as a distinct overhead term in the frame [10]. In general, G-DFT-s-OFDM offers the possibility of achieving high time granularity while maintaining the advantages of FFT processing. The price to pay is an increase of the EVM of the received data due to replacing the low power head/tail with zeros.

QUALITATIVE COMPARISON WITH OTHER WAVEFORMS

Table 1 summarizes the properties of the discussed waveforms. For the 5G candidates, we restrict our focus to FBMC-OQAM and UFMC besides the proposed G-DFT-s-OFDM. This table is intended as a high-level qualitative overview since the properties of FBMC and UFMC solutions are strongly dependent on specific implementation and filter characteristics. In general, G-DFT-s-OFDM combines the simplicity of OFDM processing with the benefits of single-carrier transmission and the multi-service support offered by UFMC and FBMC. A quantitative comparison with the other solutions is left for future work. We expect potential synergies of G-DFT-s-OFDM with the UFMC waveform to be thoroughly explored with the objective of a combined waveform design. The aim is to preserve the attractive PAPR and time localization properties of G-DFT-s-OFDM while further improving its spectral containment.

CONCLUSIONS

In this article we have introduced the generalized DFT-spread-OFDM (G-DFT-s-OFDM) waveform without CP and qualitatively discussed its suitability in addressing the 5G challenges. In general, G-DFT-s-OFDM combines the time granularity of the single carrier transmission with the computational benefits of FFT-based processing. It avoids the inefficiencies due to the CP insertion, and supports asynchronous access and multi-service integration due to its improved spectral containment. Further, it removes the time constraints of OFDM/DFT-s-OFDM, reducing latency and power consumption. The scalability to larger subcarrier spacing boosts the robustness to hardware impairments and high-speed transmission without compromising the possibility of efficiently exploiting interference suppression receivers. Its compatibility with the LTE clocking eases the integration with existing standards in commercial chips.

Future work will address a detailed comparison with other waveforms, with the aim of concretely assessing its potential for 5G. Further, the

	OFDM	DFT-s-OFDM	FBMC-QAM	UFMC	G-DFT-s-OFDM
Complexity	It requires IFFT/FFT at transmitter/receiver. One-tap equalization.	It requires (DFT +IFFT)/(FFT + IDFT) at transmitter/receiver. One-tap equalization.	It requires filtering per subcarrier. Number of operations varies depending on specific implementation and filter length. One-tap equalization might not suffice.	It requires filtering per block of subcarriers. Number of operations varies depending on implementation. One-tap equalization might suffice.	It requires (DFT +IFFT)/(FFT+ IDFT) transmitter/ receiver. One-tap equalization.
Overhead	Given by hard-coded CP	Given by hard-coded CP	No CP needed	Given by the tail of the filter	Tunable, according to the estimated delay spread of the channel
PAPR	High, given the multicarrier nature.	Low, given the quasi-single-carrier nature.	High, given the multicarrier nature.	High, given the multicarrier nature.	Low, given the quasi-single carrier nature.
Time localization	Excellent, symbol confined within the FFT window plus CP.	Excellent, symbol confined within the FFT window plus CP.	Poor, set of adjacent symbols overlapping in time	Good, slightly compromised by intersymbol interference.	Excellent, symbol confined within the FFT window
Frequency localization	Poor, it can be improved with windowing	Poor, it can be improved with windowing	Excellent with sufficiently long prototype filters [5]	Excellent with sufficiently long Dolph-Chebyshev filters [6]	Good, it can be further improved with the approach in [9]
Ultra-low latency support	Enabled by using shorter symbols.	Enabled by using shorter symbols.	Enabled by using shorter symbols, but compromised by the necessity of accommodating filter tails.	Enabled by using shorter symbols.	Enabled by using shorter symbols and/or transmitting only a portion of the IFFT output
Robustness to Doppler spread or phase noise	Obtained by enlarging the subcarrier spacing, but this may affect coexistence with users in neighbor bands	Obtained by enlarging the subcarrier spacing, but this may affect coexistence with users in neighbor bands	Obtained by enlarging the subcarrier spacing. This has no impact on users in neighbor bands.	Obtained by enlarging the subcarrier spacing. This has limited or no impact on users in neighbor bands.	Obtained by enlarging the subcarrier spacing. This has limited impact on users in neighbor bands.
Extension to MIMO	Straightforward, thanks to subcarrier-wise processing	Straightforward, thanks to subcarrier-wise processing	Not straightforward	Straightforward, thanks to subcarrier-wise processing	Straightforward, thanks to subcarrier-wise processing

Table 1. Waveform properties.

potential of a combined waveform design with UFMC will be investigated.

ACKNOWLEDGMENTS

Part of this work has been performed in the framework of the Horizon 2020 project FANTASTIC-5G (ICT-671660) receiving funds from the European Union. The authors would like to acknowledge the contributions of their colleagues in the project, although the views expressed in this contribution are those of the authors and do not necessarily represent the project.

REFERENCES

- [1] E. Dahlman *et al.*, "5G Wireless Access: Requirements and Realization," *IEEE Commun. Mag. Communications Standards Supplement*, Dec. 2014, pp. 42–47.
- [2] G. Wunder *et al.*, "5GNOW: Non-Orthogonal, Asynchronous Waveforms for Future Mobile Applications," *IEEE Commun. Mag.*, Feb. 2014, pp. 97–105.
- [3] A. Osseiran *et al.*, "Scenarios for 5G Mobile and Wireless Communications: The Vision of the METIS project," *IEEE Commun. Mag.*, May 2014, pp. 26–35.
- [4] H. Holma and A. Toskala, *LTE for UMTS: OFDMA and SC-FDMA Based Radio Access*, Wiley, 2009.
- [5] B. Farhang-Boroujeny, "OFDM versus Filter Bank Multicarrier," *IEEE Signal Processing Mag.*, vol. 28, no. 3, May 2011, pp. 92–112.
- [6] V. Vakilian *et al.*, "Universal-Filtered Multi-Carrier Technique for Wireless Systems beyond LTE," *Proc. IEEE GLOBECOM Wksp.*, Dec. 2013, pp. 223–28.
- [7] N. Michailow *et al.*, "Generalized Frequency Division Multiplexing for 5th Generation Cellular Networks," *IEEE Trans. Commun.*, vol. 62, no. 9, Sept. 2014, pp. 3045–61.
- [8] G. Berardinelli *et al.*, "Zero-Tail DFT-Spread-OFDM Signals," *Proc. IEEE GLOBECOM Wksp.*, Dec. 2013, pp. 229–34.
- [9] A. Sahin *et al.*, "An Improved Unique Word DFT-Spread-OFDM Scheme for 5G Systems," *Proc. IEEE GLOBECOM Wksp.*, Dec. 2015, pp. 1–6.
- [10] G. Berardinelli *et al.*, "On the Potential of Zero-Tail DFT-Spread-OFDM as 5G Waveform," *Proc. IEEE VTC-Fall*, Sept. 2014, pp. 1–5.
- [11] K. I. Pedersen *et al.*, "A Flexible 5G Frame Structure Design for Frequency-Division Duplex Cases," *IEEE Commun. Mag.*, Mar. 2016, pp. 53–59.

- [12] B. E. Priyanto *et al.*, "Initial Performance Evaluation of DFT-Spread OFDM Based SC-FDMA for UTRA LTE Uplink," *Proc. IEEE VTC-Spring*, Apr. 2007, pp. 3175–79.
- [13] G. Berardinelli *et al.*, "On the Potential of OFDM Enhancements as 5G Waveforms," *Proc. IEEE VTC-Spring*, May 2014, pp. 1–5.
- [14] F. Schaich and T. Wild, "Subcarrier Spacing- A Neglected Degree of Freedom?" *Proc. IEEE 16th Int'l. Wksp. Signal Processing Advances in Wireless Commun.*, June 2015, pp. 56–60.
- [15] IMT Vision, "Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," ITU doc., Radiocommunication Study Groups, Feb. 2015.

BIOGRAPHIES

GILBERTO BERARDINELLI (gb@es.aau.dk) received his first and second level degrees in telecommunication engineering, cum laude, from the University of L'Aquila, Italy, in 2003 and 2005, respectively, and his Ph.D. degree from Aalborg University, Denmark, in 2010. He is currently an associate professor at the Wireless Communication Networks (WCN) section of Aalborg University, also working in tight cooperation with Nokia Bell Labs. His research interests are mostly focused on physical layer and radio resource management design for 5G systems.

KLAUS I. PEDERSEN (klaus.pedersen@nokia.com) received his M.Sc. degree in electrical engineering and Ph.D. degree from Aalborg University in 1996 and 2000, respectively. He is currently leading a Nokia Bell Labs research team in Aalborg, and is a part-time professor at Aalborg University in the WCN section. His current research work is related to 5G design, including radio resource management aspects and end-to-end performance. He is also contributing to the EU funded research project FANTASTIC-5G.

TROELS B. SØRENSEN (tbs@es.aau.dk) received his Ph.D. degree in wireless communications from Aalborg University in 2002. Upon completing his M.Sc. E.E. degree in 1990, he worked with a Danish telecom operator developing type approval test methods. Since 1997 he has been at Aalborg University, where he is now an associate professor in the WCN section. His current research and teaching activities include cellular network performance and evolution, radio resource management, and related experimental activities.

PREBEN MOGENSEN (pm@es.aau.dk) received his M.Sc. E.E. and Ph.D. degrees in 1988 and 1996 from Aalborg University. He is currently a professor at Aalborg University, leading the WCN section. He is also associated on a part-time basis with Nokia Bell Labs as a principal engineer. His current research interests include 5G and MTC/IoT.

Flexible DFT-S-OFDM: Solutions and Challenges

Alphan Şahin, Rui Yang, Erdem Bala, Mihaela C. Beluri, and Robert L. Olesen

The authors present further modifications of DFT-S-OFDM that offer improvements in flexibility, and they discuss the latest enabling techniques. Considering the flexibility introduced by DFT-s-OFDM and its variations, the DFT-S-OFDM family also offers a set of promising waveforms for 5G networks, which will require a flexible physical layer.

ABSTRACT

Discrete Fourier transform spread orthogonal DFT-S-OFDM, adopted in 3GPP LTE uplink, enables the synthesis of block-based single carrier waveforms with various bandwidths by changing the size of the DFT-spread block. Conceptually, it also allows a transition between block-based multicarrier and single-carrier schemes when multiple DFT-spread blocks are employed in the structure. Recently, it has been shown that DFT-S-OFDM can also accommodate an internal guard period that offers flexibility on the duration of the guard periods without affecting the symbol duration. In this article, we present further modifications of DFT-S-OFDM that offer improvements in flexibility and discuss the latest enabling techniques on this topic. Considering the flexibility introduced by DFT-s-OFDM and its variations, the DFT-S-OFDM family also offers a set of promising waveforms for 5G networks, which will require a flexible physical layer.

INTRODUCTION

The design of the next generation wireless systems is currently underway in academia, industry, and regulatory and standardization bodies. The IMT-2020 Vision [1] sets the framework and overall objectives for the development of the next generation wireless systems. To address the anticipated increase in wireless data traffic, the demand for higher data rates, low latency, and massive connectivity, the IMT-2020 Vision defines the main use cases that drive the fifth generation (5G) requirements: enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine type communications (mMTC). These use cases have widely different targets on peak data rates, latency, spectrum efficiency, and mobility. The IMT-2020 Vision document confirms that the required capabilities depend heavily on the use case. It is therefore important to build flexibility in the 5G designs to efficiently meet the expected use-case-specific requirements. The air interface, specifically the waveform, is one of the key components of the new 5G technology.

Orthogonal frequency-division multiplexing (OFDM) is used for Third Generation Partnership Project (3GPP) Long Term Evolution (LTE)

and IEEE 802.11 systems in part because it allows simple receiver architectures and permits flexible access to frequency domain resources to achieve high spectral efficiency. On the other hand, since OFDM signals have high peak-to-average-power ratio (PAPR), the uplink (UL) of 3GPP LTE systems uses discrete Fourier transform spread OFDM (DFT-S-OFDM) waveform to reduce UE power consumption and power amplifier (PA) cost, and increase coverage range. Although DFT-S-OFDM exhibits high out-of-band (OOB) leakage like OFDM, the low implementation complexity makes OFDM and DFT-S-OFDM still the most likely waveforms for the 5G physical layer (PHY). Therefore, the development of the new 5G PHY will begin with low-complexity OFDM-like waveforms and attempt to improve their flexibility, PAPR, and OOB emissions.

Recently, it has been demonstrated that the structure of DFT-S-OFDM admits low-complexity and flexible mechanisms to realize improvements in OOB emission and PAPR. For instance, zero-tail (ZT) DFT-S-OFDM [2, 3] offers a method that employs a flexible internal guard period in place of the data-dependent CP without changing the original structure of DFT-S-OFDM and reduces the OOB leakage substantially. Similarly, generalized frequency-division multiplexing (GFDM) [4-6] introduces a block transmission scheme with a low-complexity frequency domain windowing approach to change the prototype filter. This approach is also beneficial to modify the original kernel of DFT-S-OFDM (i.e., the Dirichlet sinc function) and yields low PAPR. This article builds on these recent advances of DFT-S-OFDM, and provides an in-depth explanation of the flexible DFT-S-OFDM schemes. It shows how DFT-S-OFDM can be extended with mechanisms such as internal guard period, unique word (UW), and pulse shaping, while providing a survey on a set of promising waveforms for 5G networks.

The article is structured as follows. The next section describes the features of DFT-S-OFDM. We then explain the fundamental building blocks that enable enhancements to DFT-S-OFDM. After that we present the flexible DFT-S-OFDM schemes. Lastly, we present some simulation results for the surveyed waveforms, and the final section lists areas of future development and concludes the discussion.

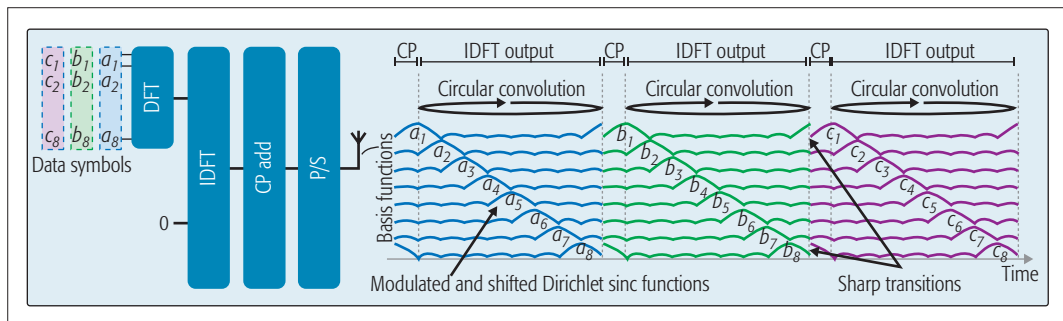


Figure 1. Features of CP DFT-S-OFDM. In this illustration, the DFT size and IDFT size are set to 8 and 64, respectively.

DFT-S-OFDM: FEATURES AND ISSUES

In conventional DFT-S-OFDM, data symbols are first spread with a DFT block, then mapped to the input of the IDFT block. The CP is prepended to the beginning of the symbol in order to avoid inter-symbol interference (ISI) due to multipath of the channel and allow one-tap frequency domain equalization (FDE) at the receiver. The structure of DFT-S-OFDM may be interpreted in different ways. One way is that DFT-S-OFDM is a precoded OFDM scheme, where the precoding with DFT aims to reduce the PAPR [7]. This interpretation has its own merit as it may yield different precoding strategies. Another way is that DFT-S-OFDM can be considered as a scheme that upsamples the data symbols by a factor equal to the ratio of the IDFT and DFT block sizes, and applies a *circular* pulse shaping with a Dirichlet sinc function before the CP extension [4–8], as illustrated in Fig. 1. From this point of view, DFT-S-OFDM is tightly related to SC waveforms that allow using single-tap FDE (commonly referred to as SC-FDE [9]) at the receiver. In SC-FDE, the CP or a fixed sequence (also known as UW sequence) is first attached to the beginning and/or end of each data symbol block. Afterward, the symbols are linearly convolved with a predetermined pulse shaping function. Hence, SC-FDE differs from DFT-S-OFDM in that it uses a linear convolution, and generates the CP (or UW) before the pulse shaping.

Due to the differences in the convolution methods, SC-FDE maintains the signal continuity between adjacent symbols and provides less OOB leakage, while DFT-S-OFDM does not provide smooth transition between the consecutive symbols. For instance, let the DFT size and IDFT size be 8 and 64, respectively. Then the upsampling ratio is $64/8 = 8$, and the number of taps for the Dirichlet sinc function is equal to 64 (i.e., the IDFT size). As illustrated in Fig. 1, the data symbols, for example, $\{a_1, a_2, \dots, a_8\}$, modulate the shifted Dirichlet sinc functions in time, which is similar to single-carrier schemes. However, as the shifts are circular in DFT-S-OFDM, the pulse shapes associated with the data symbols may lose their contiguity in time. For example, the main lobe associated with the first input of the DFT appears at the head and tail parts of the DFT-S-OFDM symbol, thus creating the sharp transitions shown in Fig. 1. In addition, DFT-S-OFDM includes a CP, which introduces

another source of discontinuity between adjacent symbols. In spite of the differences between single-carrier waveforms and DFT-S-OFDM, their joint interpretations pave the way for a flexible framework for non-traditional schemes, which are investigated in this study.

TOWARD FLEXIBLE DFT-S-OFDM

In this section, we provide a primer on the fundamental concepts that are utilized to enhance the flexibility of DFT-S-OFDM. We first discuss the idea of internal guard period and compare it with external guard interval (i.e., the CP). We then discuss GFDM under the family of DFT-S-OFDM and emphasize its frequency domain windowing operation to modify the Dirichlet sinc kernel of the conventional DFT-S-OFDM.

FLEXIBLE INTERNAL GUARD PERIOD

Cyclic prefix is a method to add an external guard period to deal with multipath channels that introduce ISI and mitigate timing synchronization errors. In general, its duration is fixed, and is designed based on the worst-case scenario (i.e., the longest possible delay spread). Therefore, a fixed CP size penalizes the users that experience shorter delay spreads.

The functionality of the CP can also be provided by an internal guard period. In this approach, the total duration of the guard period and data period is fixed. However, the ratio between the data period and the guard period is flexible, as shown in Fig. 2a. Therefore, even if the duration of the guard period changes, the total symbol duration remains constant. In a flexible internal guard period, the benefits of using the CP approach are retained by using a fixed signal, which is also known as a UW signal [10]. Since the UW signal is predefined, the fixed part of the current symbol functions as a CP for the next symbol. Therefore, the channel can still be represented as a circular convolution, which allows single-tap FDE to be used at the receiver. It is worth noting that the use of a fixed sequence at the end of each symbol block has been thoroughly investigated for SC-FDE systems (see, e.g., [9, 11, references listed therein]) and has been adopted in IEEE 802.11ad [12], where the fixed sequence is a Golay sequence. In addition, the idea of UW has been considered for OFDM-based schemes, known as UW-OFDM [10].

Figure 2b shows an uplink scenario that highlights the potential benefits of the flexible internal guard interval. Assuming that the uplink

Cyclic prefix is a method to add an external guard period to deal with multipath channels that introduce ISI and mitigate timing synchronization errors. In general, its duration is fixed, and is designed based on the worst-case scenario, that is, the longest possible delay spread. Therefore, a fixed CP size penalizes the users that experience shorter delay spreads.

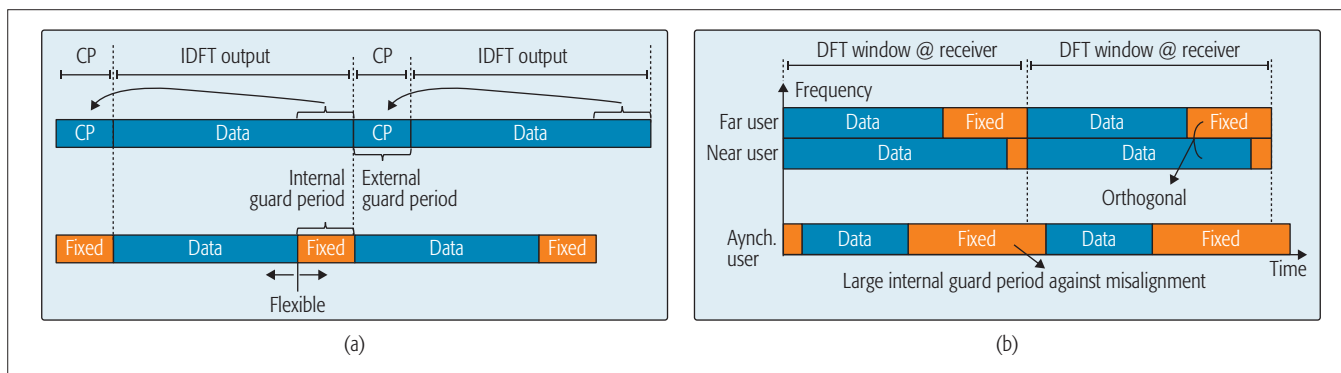


Figure 2. The flexible internal guard period and its features: a) flexible internal guard period; b) exemplary applications of flexible internal guard period in the uplink.

timing alignment is maintained loosely, the internal guard periods of devices may be extended dynamically in order to provide immunity against timing misalignment issues. In another scenario, the network may configure different guard intervals for far and near users without affecting the frame duration and penalizing other users. If the duration of the internal guard interval is long enough, there will be no ISI. Hence, the orthogonality between the signals on different subbands may still be maintained, even though the internal guard period durations on those signals are different.

The internal guard period is also useful for addressing several PHY issues. For example, the existence of a fixed sequence can be exploited in channel and noise variance estimation [13], and phase tracking [14]. Note that the fixed sequence is repeatedly transmitted as part of each symbol. Therefore, the symbols with internal guard periods inherently provide finer resolution for reference signals, which allows a receiver to compensate the phase drifts due to the phase noise or mobility within the time transmission interval (TTI). In addition, as discussed in the next section, the internal guard period provides an effective mechanism to handle the discontinuity problem of conventional DFT-S-OFDM. Hence, waveforms with internal guard periods may cause significantly less adjacent channel interference (ACI) when the orthogonality between the signals transmitted on adjacent subbands cannot be maintained due to using different numerologies or asynchronous access. Finally, the schemes that support internal guard periods offer better symbol energy use since all transmitted samples can be used at the receiver effectively. Note that the CP contains information about the data symbols. When it is discarded at the receiver, a portion of the symbol energy is not utilized during the demodulation.

FLEXIBLE PULSE SHAPING

GFDM is one of the candidate waveforms proposed for 5G that offers flexible pulse shaping [4-6]. The original motivation of GFDM is to generate a generic block-based filtered multicarrier scheme in which the energy of the data symbols is spread in time with a circularly shifted prototype filter. Compared to an OFDM symbol, a GFDM symbol spreads over multiple subcarriers and OFDM symbol durations, which is similar to FBMC. In contrast to the linear convolution used in most FBMC schemes or basic

single-carrier, GFDM brings a circular convolution into the picture. Similar to CP-OFDM and DFT-S-OFDM, GFDM also inserts the CP at the beginning of the GFDM symbol in order to enable single-tap FDE at the receiver.

GFDM can be viewed as an extension of DFT-S-OFDM. GFDM is based on the circular convolution of the up-sampled data symbols with a configurable prototype filter, while DFT-S-OFDM does the same operation with the Dirichlet sinc function. In [4, 5], it is shown that GFDM symbols can be synthesized with a frequency domain windowing approach where the output of the DFT spread block is repeated and windowed with the frequency response of the prototype filter as illustrated in Fig. 3. As a special case, if the frequency response of the filter is a rectangular function and the number of non-zero coefficients is equal to DFT-spread size, GFDM and DFT-S-OFDM become equivalent. From this point of view, GFDM introduces flexibility in the modification of the Dirichlet sinc kernel of the DFT-S-OFDM, which enables different energy distribution of the data symbols in the time-frequency plane and may result in better PAPR characteristics as compared to those of conventional DFT-S-OFDM [5].

FLEXIBLE DFT-S-OFDM: SCHEMES

In this section, we discuss the DFT-S-OFDM-based schemes that exploit the concepts discussed previously. We first investigate ZT DFT-S-OFDM and UW DFT-S-OFDM. We show that DFT-S-OFDM offers a very flexible mechanism for generating an internal guard period, which also leads to joint OOB leakage and PAPR reduction. Afterward, we investigate tail suppression mechanisms to control the leakage from data symbols to the internal guard period, which may cause ISI in a multipath channel. We first examine the tail suppression with a perturbation signal without distorting the data symbols. Then we show that modifying the kernel of the original DFT-S-OFDM with frequency-domain windowing is also beneficial to mitigate the leakage to the internal guard period.

ZERO-TAIL DFT-S-OFDM

In ZT DFT-S-OFDM [2-3], an approximately zero guard interval is generated by zeroing the data symbols that modulate the Dirichlet sinc functions, contributing to the tail of the DFT-

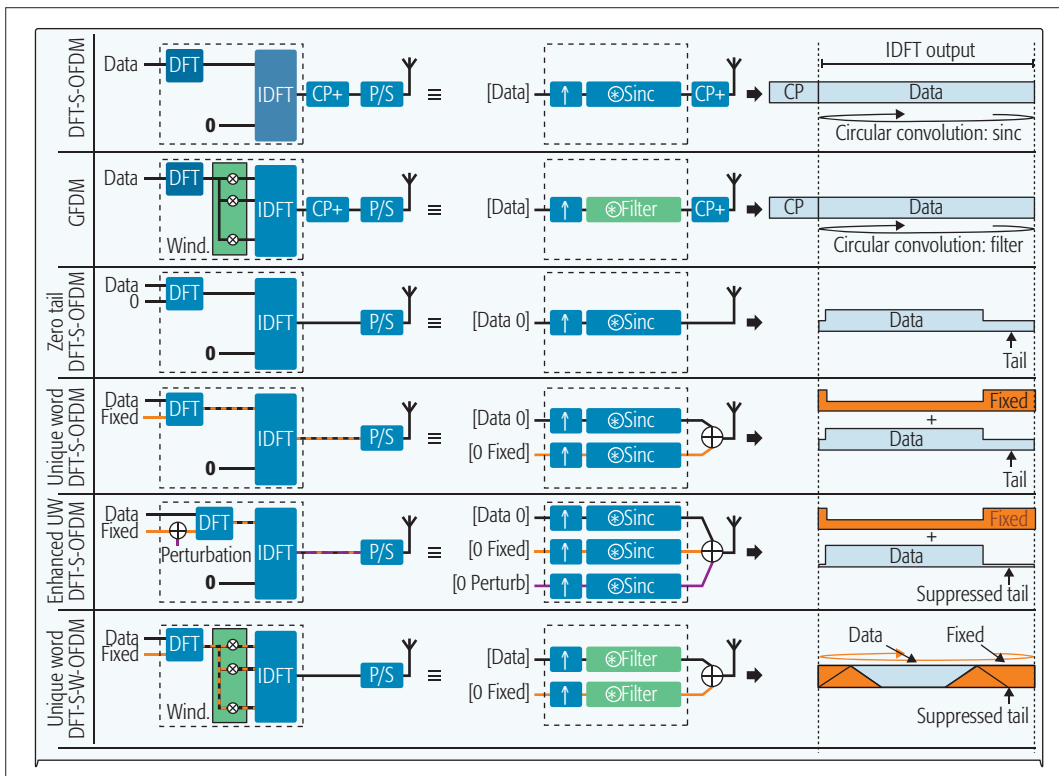


Figure 3. A comparison of DFT-S-OFDM-based waveforms and illustration of the resulting signals in time domain.

GFDM can be viewed as an extension of DFT-S-OFDM. GFDM is based on the circular convolution of the up-sampled data symbols with a configurable prototype filter, while DFT-S-OFDM does the same operation with the Dirichlet sinc function.

S-OFDM symbol. To achieve this, ZT DFT-S-OFDM also needs to have a zero sample at least for the first input of DFT-S-OFDM, which also leads to a short zero-head. In this scheme, the CP is omitted; however, as the tail of the previous symbol is approximately the same as what the CP would have been, the scheme still supports single-tap FDE. In addition, as the samples that cause significant discontinuity between the adjacent symbols are removed, this scheme generates signals that are very similar to basic single-carrier waveform with guard intervals. Therefore, it achieves substantial OOB leakage reduction. The structure of ZT DFT-S-OFDM is the same as DFT-S-OFDM with zero symbols padded to the lower end and upper end of the DFT-spread block and omission of CP insertion, as illustrated in Fig 3. Since the lengths of the tail and head are determined by the number of zeros at the input of the DFT-spread block, this scheme offers significant flexibility in setting the length of the internal guard period.

UNIQUE WORD DFT-S-OFDM

In contrast to zeroing the tail and head samples of each DFT-S-OFDM symbol, this scheme aims to generate a non-zero fixed tail and head by adding a fixed signal (i.e., UW signal) to the ZT DFT-S-OFDM symbols. The UW signal can be inserted in the time domain, or in the frequency domain as the DFT-S-OFDM involves linear operations. There are several considerations on the transmitter and receiver design with the UW signal. First, the UW signal itself should be within the transmission bandwidth. This is particularly important if the UW signal is added after the DFT-S-OFDM signal is generated. Second, the

interference that may occur due to a UW signal needs to be handled at the receiver. This consideration leads to two approaches:

- If the UW is known at the receiver a priori, the receiver needs to remove the UW signal during the equalization.
- If the UW signal is chosen such that it is orthogonal to the DFT-S-OFDM signal, the receiver does not need to know the UW a priori in order to demodulate the symbols.

One natural way of satisfying the orthogonality is to exploit the zero inputs of the ZT DFT-S-OFDM to generate the UW signal. In other words, a fixed sequence (instead of zeroes) will drive the lower end and upper end of the DFT-spread block (Fig. 3). Therefore, similar to the operations for the data symbols, the UW sequence is first up-sampled by the ratio between IDFT and DFT sizes, and then convolved with the Dirichlet sinc function. Adding a UW sequence before DFT-spread block has three main benefits:

- The orthogonality between the UW signal and the data signal is maintained as the signals use different inputs of the DFT block. Therefore, the receiver does not need to know the UW for the demodulation of the data symbols.
- Since the UW sequence is circularly convolved with the sinc function, the contiguity between the UW signals is still maintained. As a result, the existence of UW does not change the OOB leakage characteristics of ZT DFT-S-OFDM.
- Since the UW sequence is fixed and separated from the data symbols, it can be further utilized in PHY and MAC processing jointly.

Unique word DFT-spread windowed OFDM (UW DFT-S-W-OFDM) waveform suppresses the leakage to the tail part of the DFT-S-OFDM symbols by modifying the kernel of conventional DFT-S-OFDM. This is achieved by using a frequency domain windowing approach as in GFDM.

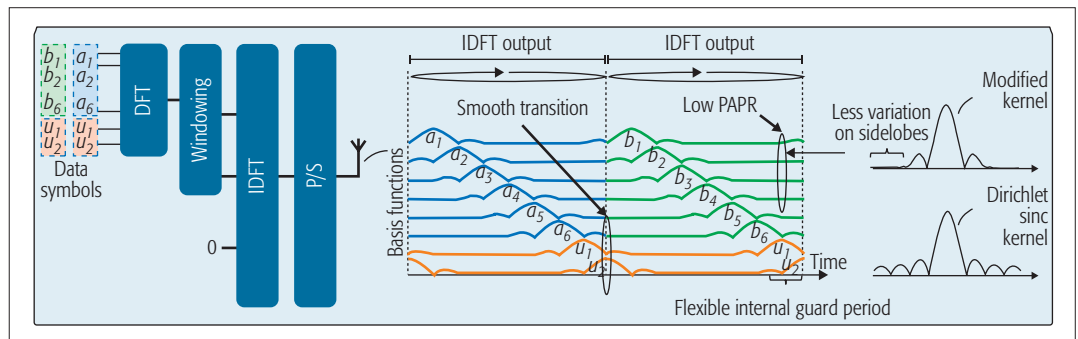


Figure 4. Features of UW DFT-S-W-OFDM. In this illustration, the DFT size and IDFT size are set to 8 and 64, respectively.

ENHANCED UNIQUE WORD DFT-S-OFDM

The main drawback of ZT DFT-S-OFDM is that the scheme may suffer from ISI in large delay spread multipath channels as the tail is not precisely zero and may be different from one symbol to the next. The enhanced UW DFT-S-OFDM addresses this challenge by introducing a perturbation signal such that when added to the original ZT DFT-S-OFDM symbol, the energy that leaks from the data symbols to the tail is suppressed. There are two main considerations for the design of the perturbation signal that may affect the receiver complexity: the energy of the perturbation signal, and the orthogonality between the perturbation signal and data signal. Reference [15] shows that if the perturbation signal is generated by using the same inputs of the DFT-spread block for UW, the energy of the perturbation signal will be small, and the data symbols can be recovered by using the conventional DFT-S-OFDM receiver. The structure of enhanced UW DFT-S-OFDM is illustrated in Fig. 3. The details on how to generate the perturbation signal via a linear precoder for tail suppression are provided in [15].

UNIQUE WORD DFT-S-W-OFDM

Unique word DFT-spread windowed OFDM (UW DFT-S-W-OFDM) waveform suppresses the leakage to the tail part of the DFT-S-OFDM symbols by modifying the kernel of conventional DFT-S-OFDM. This is achieved by using a frequency domain windowing approach as in GFDM. Additionally, a fixed sequence is applied to the lower end of the DFT-spread block to generate a UW signal at the tail of the DFT-S-OFDM symbol, and CP is omitted, as shown in Fig. 3. This scheme has three benefits compared to other schemes:

- If the length of the modified filter is shorter than the number of samples in the tail duration, the scheme leads to a linear pulse shaping for the data symbols and circular pulse shaping for the UW sequence. In this case, the proposed scheme can exactly synthesize a traditional non-block-based single-carrier waveform; for example, the SC-FDE waveform adopted in the IEEE 802.11ad millimeter-wave standard, with a block-based operation as the continuity between consecutive DFT-S-OFDM symbols is perfectly maintained.
- The scheme replaces the Dirichlet sinc function of DFT-S-OFDM with a filter that has lower sidelobes. Therefore, it mitigates the ISI from

the tail part of time domain samples of the previous symbol and decreases PAPR.

- It allows a framework to apply a cyclic shift to the symbol in the time domain in order to adjust the position of the main lobes of the data symbols in time. In other words, if the windowing operation modulates the output of the DFT-spread block, the position of the internal guard period can be shifted circularly with fine granularity, which may reduce the number of head symbols. The features of this waveform are illustrated in detail in Fig. 4.

PERFORMANCE ANALYSIS

In this section, we compare the surveyed schemes through simulations and emphasize the trade-offs introduced by the schemes. For all schemes, we consider an IDFT block of 512 subcarriers, a DFT-spread block of 128 bins, and a guard period length of 64 samples. For ZT DFT-S-OFDM, enhanced UW DFT-S-OFDM, and UW DFT-S-W-OFDM, the last 16 inputs of a DFT-spread block are reserved for generating the internal guard period. In addition, the first input of the DFT-spread block is reserved for ZT DFT-S-OFDM and enhanced UW DFT-S-OFDM in order to maintain the continuity between adjacent DFT-s-OFDM symbols. The modulation order is set to 4-quadrature amplitude modulation (QAM) unless otherwise stated. The UW is chosen as a random fixed sequence in the simulations. As a final point, the windowing function for UW DFT-S-W-OFDM is generated as follows: A root raised cosine filter, where the oversampling rate, the roll-off factor, and the filter length are $512/128 = 4$, 0.1, 33 samples, respectively, is first generated. After the filter is padded with zeroes, its frequency response is calculated. Except for the main lobe of 144 samples, all other coefficients in frequency are then set to zeroes.

In Fig 5a, the time and spectrum characteristics of the waveforms are provided. In order to show the contribution of the data symbols to the tail part of the symbol in time, the fixed signals (i.e., UW signal) are not included. As shown in Fig. 5a, the energy at the tail part of ZT DFT-S-OFDM is 15–20 dB lower than that at the non-tail part. Therefore, the leakage at the tail causes ISI in multipath channels, which can be a limiting factor for this scheme with high-order modulations in a rich scattering environment. In enhanced UW DFT-S-OFDM, the tail is suppressed 5–10 dB more compared to the ZT DFT-S-OFDM, at the expense of calculation of

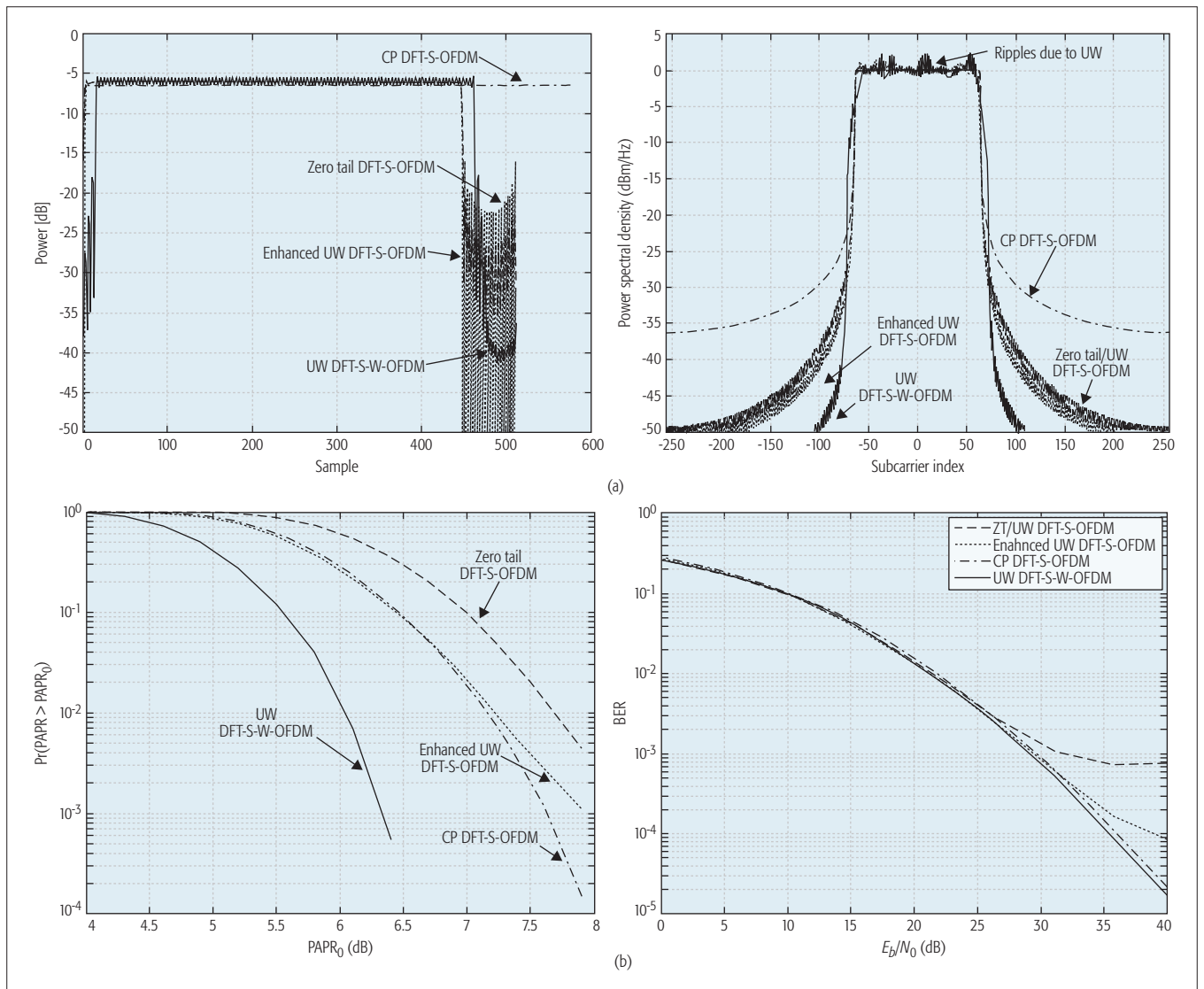


Figure 5. Performance of the investigated schemes: a) time and frequency characteristics with null UW and random UW, respectively; b) PAPR and BER characteristics.

the perturbation symbols. The UW DFT-S-W-OFDM exhibits less contribution from the data symbols to the tail compared to that of other schemes. This is due to the fact that the kernel function used in UW DFT-S-W-OFDM decays faster than that of DFT-S-OFDM. By modulating the output of the DFT-spread block, the position of the internal guard period is also adjusted for UW DFT-S-W-OFDM. Therefore, it uses 16 symbols for the UW sequence, as opposed to 17 symbols for ZT and UW DFT-S-OFDM waveforms. Note that UW DFT-S-W-OFDM spreads the data symbols over 144 subcarriers in frequency, and introduces a trade-off between the tail suppression and utilization of a higher number of subcarriers in frequency compared to other schemes. The use of an internal guard period provides significant improvement on the OOB leakage performance. Since the OOB leakage is a function of continuity of the symbols, the schemes that provide better tail characteristics yield less OOB leakage. As the DFT-S-W-OFDM suppresses the contributions from the data symbols more than other schemes, it pro-

vides superior OOB performance. An artifact of UW is observed as the ripples in the plateau since the UW is a fixed periodic sequence.

In Fig 5b, the PAPR and BER performance of the schemes are provided. Since ZT DFT-S-OFDM symbols have approximately zero tail, the average power for ZT DFT-S-OFDM decreases. Hence, the PAPR of ZT DFT-S-OFDM is higher than that of the other schemes. However, the ZT DFT-S-OFDM transmitter does not need to back off the transmitter power more than the conventional DFT-S-OFDM as the power of the peak samples remains the same. Enhanced UW DFT-S-OFDM and CP DFT-S-OFDM provide similar PAPR performance, while the impact of UW in UW DFT-S-OFDM appears in the tail of the empirical complementary cumulative distribution function (CCDF). Since UW DFT-S-W-OFDM decreases the variation of the prototype filter's sidelobes by replacing the original kernel of DFT-S-OFDM, this scheme offers approximately 1 dB improvement in the PAPR compared to CP DFT-S-OFDM. For the bit error rate (BER) analysis, we consider a multipath

DFT-S-OFDM is a promising candidate waveform for 5G due to its low PAPR and flexibility in accommodating an internal guard period that has several advantages including robustness against timing misalignment and adaptability of the guard period.

channel where the power delay profile is exponential. The unnormalized power of the t th tap is modelled as $\exp(-\tau t)$ where τ is the decaying factor and set to 0.5. In the simulations, the gain on each tap is generated independently and drawn from complex Gaussian distribution. We set the modulation as 64-QAM and use a single-tap MMSE-FDE receiver that allows the receiver to exploit the path diversity for SC systems [11]. No channel coding is considered in the simulations. As shown in Fig. 5b, ZT and UW DFT-S-OFDM exhibit interference-limited behavior. On the other hand, the enhanced UW DFT-S-OFDM and DFT-S-OFDM address the limitation as these schemes suppress the tail of the symbol.

CONCLUSION

DFT-S-OFDM is a promising candidate waveform for 5G due to its low PAPR and flexibility in accommodating an internal guard period that has several advantages including robustness against timing misalignment and adaptability of the guard period. This article has investigated different flavors of DFT-s-OFDM waveforms and analyzed their relationships. Specifically, it was shown that GFDM can be viewed as an extension of DFT-s-OFDM with a flexible choice of pulse shaping kernel. ZT DFT-S-OFDM provides a simple and flexible way to generate an almost zero internal guard period. Two new UW-based waveforms, one using a perturbation signal and another pulse shaping with frequency domain windowing, have been presented. These waveforms offer low OOB and PAPR, while achieving similar or better BER performance in comparison with that of CP DFT-S-OFDM. Furthermore, a deterministic UW signal, which is orthogonal to the data signal, may be added and utilized by the receiver for channel estimation and phase tracking.

REFERENCES

- [1] ITU-R Rec. M.2083-0, "IMT Vision – Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," Sept. 2015.
- [2] G. Berardinelli *et al.*, "On the Potential of OFDM Enhancements as 5G Waveforms," *Proc. IEEE VTC-Spring*, May 2014, pp. 1–5.
- [3] G. Berardinelli *et al.*, "On the Potential of Zero-Tail DFT-Spread-OFDM in 5G Networks," *Proc. IEEE VTC-Fall*, Sept. 2014, pp. 1–6.
- [4] N. Michailow *et al.*, "Generalized Frequency Division Multiplexing: Analysis of an Alternative Multi-Carrier Technique for Next Generation Cellular Systems," *Proc. Int'l. Symp. Wireless Commun. Sys.*, Aug. 2012, pp. 171–75.
- [5] N. Michailow and G. Fettweis, "Low Peak-to-Average Power Ratio for Next Generation Cellular Systems with Generalized Frequency Division Multiplexing," *Proc. IEEE Intell. Signal Processomg and Commun. Sys.*, Nov. 2013, pp. 651–55.
- [6] N. Michailow *et al.*, "Generalized Frequency Division Multiplexing for 5th Generation Cellular Networks," *IEEE Trans. Commun.*, vol. 62, no. 9, Sept. 2014, pp. 3045–61.
- [7] A. Ghosh *et al.*, "LTE-Advanced: Next-Generation Wireless Broadband Technology [Invited Paper]" *IEEE Wireless Commun.*, vol. 17, no. 3, June 2010, pp. 10–22.

- [8] H. G. Myung *et al.*, "Peak Power Characteristics of Single Carrier FDMA MIMO Precoding System," *Proc. IEEE VTC-Fall*, Sept. 2007, pp. 477–81.
- [9] F. Pancaldi *et al.*, "Single-Carrier Frequency Domain Equalization," *IEEE Signal Processing Mag.*, vol. 25, no. 5, Sept. 2008, pp. 37–56.
- [10] M. Huemer *et al.*, "Design and Analysis of UW OFDM Signals," *AEU – Int'l. J. Electronics Commun.*, vol. 68, no. 10, 2014, pp. 958–68.
- [11] D. Falconer *et al.*, "Frequency Domain Equalization for Single-Carrier Broadband Wireless Systems," *IEEE Commun. Mag.*, vol. 40, no. 4, Apr. 2002, pp. 58–66.
- [12] IEEE Std 802.11ad-2012, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band," Dec. 2012, pp. 16–28.
- [13] J. Coon *et al.*, "Channel and Noise Variance Estimation and Tracking Algorithms for Unique-Word Based Single-Carrier Systems," *IEEE Trans. Wireless Commun.*, vol. 5, no. 6, June 2006.
- [14] M. Huemer, H. Witschnig, and J. Hausner, "Unique Word Based Phase Tracking Algorithms for SC/FDE-Systems," *Proc. IEEE GLOBECOM*, vol. 1, Dec. 2003, pp. 70–74.
- [15] A. Sahin *et al.*, "An Improved Unique Word DFT-Spread OFDM Waveform for 5G Mobile Broadband," *Proc. IEEE GLOBECOM Wksp. 5G & Beyond – Enabling Technologies and Applications*, San Diego, CA, Dec. pp. 1–5.

BIOGRAPHIES

ALPHAN ŞAHİN [S'03, M'15] received his B.S. degrees in electrical engineering and telecommunication engineering, and his M.S. degree in electrical engineering from Istanbul Technical University, Turkey, and his Ph.D. degree in electrical engineering from the University of South Florida, Tampa, in 2005, 2006, 2008, and 2013, respectively. From 2006 to 2010, he was with the Scientific and Technological Research Council of Turkey and worked on information security as a researcher. From 2014 to 2015, he visited Texas A&M University, College Station, and Florida International University, Miami, and worked as a postdoctoral research associate. He is currently with InterDigital Communications. His research interests include signal processing techniques emphasizing on communication systems, heterogeneous wireless networks, and machine learning.

RUI YANG [M'87] received his M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1987 and 1992, respectively. He has 17 years of experience in the research and development of wireless communication systems. Since he joined InterDigital in 2000, he has led several product development and research projects. He is currently a principal engineer at InterDigital Labs. His interests include digital signal processing and air interface design for cellular and WLAN networks. He has received more than 15 patents in those areas.

ERDEM BALA received his B.S. and M.S. degrees from Bogazici University, Istanbul, Turkey, and his Ph.D. degree from the University of Delaware, all in electrical engineering. He is currently a research engineer at InterDigital Labs conducting research on future radio access networks. Previously, he was an intern at Mitsubishi Research Labs and a software engineer at Nortel Networks. He has numerous issued and pending patents, and has received the InterDigital Innovation Award and the Nortel Networks Pride Award.

MIHAELA C. BELURI [M'90] received her M.S. degree in electrical engineering from the Polytechnic University of Bucharest, Romania. She is currently a principal engineer with InterDigital Communications, Melville, New York, working on millimeter-wave and 5G technologies. She has authored or coauthored IEEE conference publications and holds several patents. Her research interests include dynamic spectrum management and shared spectrum technologies, algorithm design, modeling, and simulations for wideband code-division multiple access, high-speed packet access, and LTE systems.

ROBERT OLESEN received his M.S. degree in electrical engineering from New York University in 1988, and an E.M.B.A. from Hofstra University in 2006. He has 36 years of experience in research and development of microwave and wireless communication systems. Since he joined InterDigital in 1999 he has been a project lead and program manager for wireless standards related projects including 802.11 WLAN, 3GPP LTE, 5G wireless, and 3GPP New Radio. He is currently a senior director at InterDigital Labs. His interests include wireless MIMO research, microwave research, and next generation air interface design. He is the holder of over 70 patents and has received the Chairman's award for most important technology.

CALL FOR PAPERS

IEEE COMMUNICATIONS STANDARDS MAGAZINE

REAL TIME COMMUNICATIONS IN THE WEB: CURRENT ACHIEVEMENTS AND FUTURE PERSPECTIVES

BACKGROUND

Web Real-time Communication is a joint standardization effort between the Internet Engineering Task Force (IETF) and the World Wide Web Consortium (W3C). Since 2011 the "Real-Time Communication in WEB-browsers" (RtcWeb) Working Group has been working on key aspects like the overall communication infrastructure, the protocols and API (Application Programming Interface) requirements, the security model, the media formats (and related media codecs), as well as advanced functionality like congestion/flow control and interworking with legacy VoIP equipment. While the W3C WebRTC wg has conducted a parallel activity on the definition of a whole set of APIs exposing functions like exploration and access to device capabilities, capture of media from local devices, encoding/processing of media streams, establishment of peer-to-peer connections between web-enabled devices, decoding/processing of incoming media streams and delivery of such streams to the end-user in an HTML5-compliant fashion.

To date, the two mentioned working groups have done a tremendous amount of work, which has brought us close to what can undoubtedly be considered an unprecedented milestone in the field of real-time multimedia communications: the so-called "WebRTC 1.0" standards suite. The idea behind WebRTC 1.0 is to allow all of the involved stakeholders (browser vendors, telecommunication providers, application providers, web developers, etc.) to converge onto a well-defined set of protocols and APIs to be leveraged in order to allow wide-spread deployment on the market of interoperable products offering end-users a media-rich, web-enabled real-time experience.

WebRTC has also had to confront itself with alternative views. In fact, since the beginning of 2014, a brand new initiative has seen the light in the W3C, the ORTC (Object Real-time Communications) Community Group, which has initially been identified as a clear opponent to WebRTC. Fortunately, after the unavoidable initial friction, the international standardization community has decided not to disperse precious energies and has eventually come to a sort of compromise. The idea has been to rapidly converge to a unique agreed-upon solution by allowing the ORTC community to contribute to the finalization of the first version of the standard. At the same time, a common decision has been taken as to welcome most of the key concepts proposed with ORTC's low-level object API to be adopted for the so-called "Next Version" of the standard, which nonetheless has backward compatibility with the 1.0 release among its foundational requirements. This is exactly where we stand now. We are a step away from completing WebRTC 1.0, with our minds already looking at the rising WebRTC-NV initiative.

In light of these trends, this FT (Feature Topic) will both focus on the current state of the art with respect to WebRTC 1.0 completion and introduce the envisioned work program for the second generation of the standard. In doing so, we invite contributions dealing with the hot topics and illustrating how the community has successfully coped with most of them. We also await papers making some projections as to what the standardization community imagines will represent the upcoming milestones and open issues. Tutorial-oriented contributions shedding some light on the current status of standardization, with special focus on the upcoming final release of the so-called WebRTC 1.0 standard ecosystem are most welcome. This FT will take the stock of the situation with respect to topics like, e.g., codecs, session description, data-channel and stream multiplexing. It will also illustrate how standard bodies are dealing with seamless integration of WebRTC with the initially competing ORTC effort. Finally, it will take a look at the future by welcoming contributions about the forthcoming initiative informally known as WebRTC Next Version (WebRTC-NV).

Stated in one sentence, this Feature Topic aims at presenting the consolidated results achieved so far in the area of standardization of real-time communications in the Web, by presenting a comprehensive view of the numerous challenges researchers have had to face before arriving at the first release of an agreed-upon WebRTC standard suite, while also providing useful hints on the upcoming standardization efforts.

Original contributions previously unpublished and not currently under review, are solicited in relevant areas including (but not limited to) the following:

- WebRTC 1.0 standard protocols and APIs;
- WebRTC 1.0 JavaScript programming patterns and APIs;
- Practical experiences with WebRTC 1.0: testbeds and business cases;
- WebRTC Security Architecture: signaling, consent, privacy, communications security, peer authentication, trust relationships;
- ORTC low-level object oriented approach to real-time web communications;
- WebRTC support in the browsers;
- WebRTC business readiness and industry adoption;
- WebRTC gateways for interoperability with legacy VoIP architectures;
- Performance monitoring/evaluation of WebRTC architectures;
- WebRTC Next Version: milestones and work plan.

SUBMISSIONS

Articles should be tutorial in nature and written in a style comprehensible and accessible to readers outside the specialty of the article. Authors must follow the *IEEE Communications Standards Magazine's* guidelines for preparation of the manuscript. Complete guidelines for prospective authors can be found at <http://www.comsoc.org/comstandardsmag/author-guidelines>.

It is important to note that the *IEEE Communications Standards Magazine* strongly limits mathematical content, and the number of figures and tables. Paper length should not exceed 4500 words. All articles to be considered for publication must be submitted through the IEEE Manuscript Central site (<http://mc.manuscriptcentral.com/commag-ieee>) by the deadline. Select "Standards Supplement" from the drop down menu of submission options.

IMPORTANT DATES

- Manuscript Submission Date: January 15, 2017
- Decision Notification: March 15, 2017
- Final Manuscript Due Date: April 15, 2017
- Publication Date: June 2017

GUEST EDITORS

Dr. Salvatore Loreto
Ericsson AB, Stockholm, Sweden

Prof. Simon Pietro Romano
University of Napoli Federico II, Italy

Prof. Carol Davids
Illinois Institute of Technology, USA

COMMUNICATIONS EDUCATION AND TRAINING: EDUCATIONAL SERVICES BOARD



Rulei Ting

David G. Michelson

Michele Zorzi

Education and professional development has been described as the third pillar of the IEEE Communications Society after publications and conferences. Certainly no other Society activity has the potential to touch and impact as broad a group of the membership. In 2008, the IEEE Communications Society renewed its commitment to education with the revitalization of the ComSoc Education Board under the leadership of Prof. Stefano Bregni (2008–2011). Since then, the range of activity sponsored by the Board has expanded tremendously under the leadership of Prof. David Michelson (2012–2013), Prof. Michele Zorzi (2014–2015), and Dr. Rulei Ting (2016–2017). The Board has undergone several name changes in the process. It was known as the Education Board from 2008 to 2013, the Education and Training Board from 2013 to 2015, and the Educational Services Board since 2016.

In “Education and Training in ComSoc: Recent Achievements,” Michele Zorzi shares the highlights of education and training activity within ComSoc while he served as Director of Education and Training, including:

1. Recognition of Telecommunications Engineering as a distinct engineering discipline by ABET
2. Publication of the first four editions of the IEEE Communications Magazine Feature Topic on Education and Training
3. Continued presentation of technical sessions on Education and Training at ICC and GLOBECOM
4. Development of the ComSoc Hands-on Lab Exchange
5. Growth and expansion of ComSoc Training
6. Establishment of the ComSoc summer school.

Moreover, Michele concludes, the way forward is very bright.

By providing direct exposure to the equipment, processes, principles, and outcomes associated with a particular discipline, laboratory assignments and projects provide engineering students with opportunities to develop intuition and appreciation for the field that cannot be matched by lectures or simulation. However, the effort required to develop suitable experiments and associated documents for communications education can be considerable. Historically, there

have been few formal mechanisms to facilitate sharing of laboratory assignments so that others may build on and benefit from previous efforts. In “The ComSoc Hands On Lab Exchange,” Rhys Bowley, Erik Luther, and David G. Michelson describe a new web-based platform developed by the Educational Services Board that will allow a more flexible method for enabling peer review and sharing of laboratory assignments and projects in a timely manner for the benefit of the teaching community. The ultimate intent is to build a community of contributors and users from both industry and academia that will promote and develop this important activity.

Summer schools have a long history of contributing to the training of doctoral students. They allow candidates to expand their horizons and build a sense of community by participating in an intensive schedule that includes attending presentations by experts, participating in technical discussions with peers, sharing their accomplishments with both, and socializing with both. Two years ago, senior leaders within ComSoc recognized the need to overcome the lack of summer school activity in communications and authorized formation of a ComSoc summer school with Prof. Fabrizio Granelli as director. In “Training and Networking for Young Society Members: The ComSoc Summer School Program,” Fabrizio shares the experience and success achieved by the first two editions of the ComSoc Summer School on Communications that were held in Trento, Italy, in 2015 and 2016. The tremendously positive response to this initiative bodes well for future offerings of this type, including activities oriented toward other cohorts within our community.

The Scholarship of Teaching and Learning (SoTL) movement promotes the notion that the same scholarly approach that we apply to research, including scholarly inquiry, peer review, and publication of results, can also be applied to teaching and learning. Although SoTL has gathered a large following within the academic community in recent years, adoption by communications and other engineering disciplines is still in its early stages. In “Integrating the Scholarship of Teaching, Learning, and Research

Supervision into Communications Education,” David G Michelson introduces the precepts of SoTL, methods by which educators can begin contributing to SoTL, and several initiatives that the IEEE Communications Society is undertaking in an effort to promote SoTL in communications education. For those with an interest in SoTL, this article provides an outstanding entry point into a field that will undoubtedly become increasingly important in coming years. It also serves as a natural lead-in (and will encourage submissions) to a Feature Topic on SoTL in Communications Education that will appear in the November 2017 issue of this magazine.

The next Feature Topic on Education and Training will focus on Industry Standards in Education and Training, and will appear in May 2017. See the Call for Papers on the *IEEE Communications Magazine* website.

BIOGRAPHIES

RULEI TING (rt@ieee.org) is a senior technical director at AT&T, currently leading AT&T Collaborate program management initiatives. In addition to his Ph.D. in electrical engineering, he earned an Executive Master's in technology management at the Wharton School and Penn Engineering, University of Pennsylvania. He serves as IEEE Communications Society's Director of Educational Services, 2016–2017, and Committee Chair of IEEE Wireless Communications Engineering Technologies Certification.

DAVID G. MICHELSON (davem@ece.ubc.ca) leads the Radio Science Lab at the University of British Columbia where his research focuses on wireless communications. He completed the UBC Faculty Certificate Program on Teaching and Learning in Higher Education in 2011, served as ComSoc's Director of Education and Training from 2012 to 2013, and has been a member of the ComSoc Educational Services Board from 2012 to present. He is also an elected Member-at-Large of the ComSoc Board of Governors (2013–2015, 2017–2019).

MICHELE ZORZI [F] (zorzi@ing.unife.it) is with the Information Engineering Department of the University of Padova. His present research interests focus on various aspects of wireless communications. He was Editor-in-Chief of *IEEE Wireless Communications* from 2003 to 2005, *IEEE Transactions on Communications* from 2008 to 2011, and, at present, *IEEE Transactions on Cognitive Communications and Networking*. He served as a Member-at-Large of the ComSoc Board of Governors from 2009 to 2011, and as Director of Education and Training from 2014 to 2015.

“
*We learn by pushing ourselves
 and finding what really lies at the
 outer reaches of our abilities.”*

~ Josh Waitzkin

IEEE COMSOC
TRAINING
www.comsoc.org/training



Education and Training in ComSoc: Recent Achievements

Michele Zorzi

This article describes the recent achievements of the IEEE Communications Society in the area of education and training, and provides some background and description of how the Education and Training Board has been engaged in several activities that have led to results that position ComSoc at the forefront of educational activities and services within IEEE.

ABSTRACT

This article describes the recent achievements of the IEEE Communications Society in the area of education and training, and provides some background and description of how the Education and Training Board has been engaged in several activities that have led to results that position ComSoc at the forefront of educational activities and services within IEEE.

INTRODUCTION

The two signature areas of activities of ComSoc (and in fact of IEEE as a whole) have always been publications and conferences. IEEE and ComSoc are well known and highly respected for their top-quality and high-impact journals and magazines, and for their many influential meetings and conferences, which span all topics of interest within IEEE. The number of citations these attract is one piece of evidence of their impact, which puts them at the very top of all international rankings in comparable fields.

In an effort to better understand its role within IEEE and the scientific and technical community, and to identify effective ways to provide better and more useful services to its members, in recent years ComSoc has gone through a deep rethinking of its activities and its priorities. In particular, it tried to come up with some strategic guidelines about a way forward that promotes the excellence of its mission, attracts members through tangible benefits, and is compatible with the new challenges and technological opportunities of this new and exciting era.

In this context, education and training (now referred to as educational services) were identified as an area of great potential growth and significant opportunity for ComSoc, to the point that it has been called the “third pillar,” which, together with Publications and Conferences, will create value for the membership of ComSoc and sustain it in the years to come [1]. Strengthening Education and Training offerings as part of the benefits of membership in the Society is a key action to attract new members to ComSoc among students and professional engineers and practitioners (two groups that have been declining in recent years, and who would bring new life and new perspectives to ComSoc).

I was fortunate enough to be part of this piece of ComSoc’s history, as the Director of Education

and Training (and Chair of the Education & Training Board, ETB) in 2014 and 2015. These two years were arguably the time in which many initiatives in education and training came to life and became concrete, although many of them had been started well before I took office. The legacy I got from the previous Director, Dave Michelson (who has been serving on the Board to this day), and the work done by Stefano Bregni before him, were very rich and gave us a lot of ideas, and material that was brought to completion during my time.

Besides Dave, who, as the Past Director and initiator of many of the things we implemented, has been an invaluable help and tireless collaborator in all these initiatives, the current portfolio of education and training activities offered by ComSoc is the result of a team effort, to which the dedicated members of the ETB (many of whom had served multiple terms), with the constant support of ComSoc staff, contributed with enthusiasm and competence. Being the Chair of such a Board was really a pleasure, and seeing so many things being done by motivated people with little effort from my side was a wonderful leadership experience. We are now working with the new Director, Rulei Ting, who, with the help of the new Board, will consolidate these results and bring ComSoc’s Educational Services Board (ESB) to new heights. The ETB met regularly through face-to-face meetings at IEEE ICC and GLOBECOM twice a year, and in addition met (often on a monthly basis) via teleconference, to attend its business and to follow up on the many tasks.

In this article, I describe the various initiatives that were carried out in 2014–2015. Besides serving as recognition of the people who made these happen, I hope this report will also inspire others who care about educational activities and about the technical training of our members to engage in community service and to invest their talents for the furthering of our Society in such an important area, in a time where knowledge and continuous education have become a key asset in both research and the practicing engineering profession.

TELECOMMUNICATION ENGINEERING EDUCATION

The ETB was very active for many years in promoting telecommunications engineering as a distinct education discipline, and in providing

support activities for this to become practically implementable.

The most significant achievement in this area is undoubtedly the resolution of the Accreditation Board for Engineering and Technology (ABET) about telecommunications engineering. After more than six years of effort, primarily led on behalf of ComSoc by Tarek El-Bawab with the help of several people along the way, on November 1, 2014 the ABET Board finally approved the new accreditation criteria for “Electrical, Computer, Communications, and Similarly Named Engineering Programs” (<http://www.abet.org/eac-criteria-2015-2016/>). This approval brought to a fruitful conclusion ComSoc’s Telecommunications Engineering Education (TEE) Movement, which started quite a few years ago [2] and the long history of which is described in two articles that appeared in the November 21, 2014 issue of IEEE’s periodical *The Institute* [3] and in the November 2015 issue of *IEEE Communications Magazine* [4]. Also, an interview was posted on ComSoc Beats [5].

Of course, recognition of telecommunications engineering as a distinct discipline is only the first step toward developing curricula and programs in this specialty. Much effort is still needed from universities to introduce innovative courses of study in this field, as well as teaching material that focuses on modern telecommunications engineering. As part of this effort, a new textbook series was launched in 2015 by Springer to produce teaching material for telecommunications engineering courses and curricula. Proposals for textbooks in this series have been solicited, and at least three projects are already underway (Free-Space Wireless Optical Communication, Cognitive Radio, and Traffic Engineering).

Along these lines, and to promote some experimentation in teaching in telecommunications, the U.S. Department of Commerce’s National Institute of Standards and Technology (NIST) decided in August 2015 to fund a project to develop a course on the *Telecom Standards and Standardization Process*, as part of its effort to integrate standards education into science, technology, engineering, and mathematics (STEM) programs. The project, which is led by two Telecommunications Engineering Education (TEE) Work Group members, involves ComSoc’s Standards organization and leadership, and is technically supported by ComSoc. As mentioned, the next logical step is to launch an effort to design telecommunications engineering curricular models. While this work is underway, the NIST experience can be taken as an example to seek funding for other TEE-based initiatives and produce more telecommunications engineering courses. Both the National Science Foundation (NSF) and NASA have curricular development programs that can be utilized in this regard.

Finally, the ETB worked on moving the ABET work forward, and keeping the ComSoc Board of Governors (BoG) and IEEE BoG aware. Key people behind the TEE initiative are now bringing this experience to the IEEE at large, as members of the Educational Activities Board and of the Products and Services Committee at the IEEE level. An international angle for the TEE initiative has also been explored,

with focus on whether and how the ABET work can be extended beyond those countries where ABET operates, although this work is still in progress.

IEEE COMMUNICATIONS MAGAZINE SERIES ON EDUCATION AND TRAINING

The ETB has supported a series in *IEEE Communications Magazine* with a focus on Communications Education and Training. The series, edited by Dave Michelson and Wen Tong, is published twice a year. Four issues have been published so far: “Software Defined Radio” (May 2014) [6], “Expanding the Student Experience” (December 2014) [7], “Student Competitions” (May 2015) [8], and “Ethics and Professionalism” (November 2015) [9]. The present article is part of the November 2016 issue of the Series, and two Feature Topics are being advertised for 2017 (“Communications Standards in Education,” May 2017, and “Scholarship of Teaching and Supervision,” November 2017).

Having a regular presence in the most widespread of ComSoc publications gives visibility and prominence to the area of education and training, and, although the articles published in this series are somewhat different from the technical contributions that are typically found in IEEE periodicals, they contribute to spreading the notion that offering educational and training services to its members is a key mission for ComSoc. It also offers a discussion forum about how ComSoc can play a leadership role in furthering knowledge and helping its members to increase their own professional preparation and their competitiveness in the job market.

TECHNICAL SESSIONS ON EDUCATION AND TRAINING AT ICC AND GLOBECOM

In the discussion about making education and training a key area for ComSoc, it soon became apparent that a stronger and more direct connection with the members, as well as a forum to reach out to the community, were needed. For this reason, the ETB discussed the possibility of also having a presence on education and training at ComSoc’s flagship conferences. The main goal was to provide a forum in which to discuss issues such as telecommunications engineering education, availability of teaching and hands-on resources as a means to increase the effectiveness of education, and ways to bring together academics and practitioners so as to increase awareness of the needs and opportunities related to the continuous training of the workforce in the telecommunications industry.

This initiative started during Dave Michelson’s term as Director of Education and Training (2012–2013). The first Education sessions were held at GLOBECOM 2012 (“How Software Defined Radio Will Revolutionize Lab-Based Communications Courses,” on December 5, 2012, in Anaheim, California) and at GLOBECOM 2013 (“IF5: Hands-on Education and Training with Software-Defined Radio I” and “IF7: Hands-on Education and Training with Software-Defined Radio II,” on December 10, 2013, in Atlanta, Georgia). The success of these

In the discussion about making education and training a key area for ComSoc, it soon became apparent that a stronger and more direct connection with members, as well as a forum to reach out to the community, were needed. For this reason, the ETB discussed the possibility of also having a presence on education and training at ComSoc’s flagship conferences.

early initiatives gave rise to the proposals for a feature series in *IEEE Communications Magazine* and for the Hands-on Lab Exchange (discussed in other sections of this article).

Encouraged by the success obtained by these sessions at GLOBECOM 2012 and 2013, the ETB decided to make these sessions a regular event. More recently, panel sessions were held at the Industry Forum at GLOBECOM 2014 (“Hands-on Education and Training with Software Defined Radio,” session IF-27 of the Industry Program on December 11, 2014 in Austin, Texas, <http://globecom2014.ieee-globecom.org/indforum.html>), at ICC 2015 (“Preparing the Next Generation of Communications Designers with Education and Training,” Industry Panel #11 on June 10, 2015 in London, United Kingdom, <http://icc2015.ieee-icc.org/content/industry-panels>), at GLOBECOM 2015 (“Education and Training for the Next Generation of Communications Engineers,” session IF-18, on December 9, 2015 in San Diego, California, <http://globecom2015.ieee-globecom.org/program/industry-program/panels#IF18>), and at ICC 2016 (“Integrating Knowledge of Telecom Standards into Engineering Education,” session IF-05, on May 25, 2016 in Kuala Lumpur, Malaysia, <http://icc2016.ieee-icc.org/content/industry-panels#IF-05>).

A discussion has also been ongoing within the ETB about how to extend this initiative and make it part of the ICC/GLOBECOM technical program. In particular, ETB and GTC members have been discussing this issue and working on a proposal to shape this initiative in the future. The goal is to identify the “one have-to-attend event” that will push an

education/industry focus at these two ComSoc annual events to ultimately engage a greater participation from industry.

HANDS-ON LAB EXCHANGE: A PLATFORM FOR SHARING EDUCATIONAL MATERIAL

As part of the new ComSoc initiatives in the area of online content, we pursued the implementation of a pilot project the inception of which goes back to Dave Michelson and the Education session held at GLOBECOM 2012, but the implementation of which was spearheaded by Erik Luther in 2015. The goal was to make publicly available some education material on hands-on software defined radio related to the articles published in the May 2014 issue of the *IEEE Communications Magazine* Series on Communications Education and Training, as an example of a new way for ComSoc to provide support to teachers and professionals. After several discussions with ComSoc about how to put such content on a ComSoc platform, and after some testing, the site went live at the end of 2015 and is now operational.

The Hands-on Lab Exchange, accessible through the link <http://labs.comsoc.org>, hosts material that has been peer-reviewed for quality. Currently, it hosts two courses (“Coursework for Introducing Matlab to Communications Engineers” and “Introduction to Communication Systems Using Software Defined Radio”), and has different sections with different levels of access privileges in order to provide some free content to all visitors and possibly some restricted content only for members. An illustration of the current interface is given in Fig. 1, and a more extensive description of this initiative, including some history on how it came about, is given in another article in this issue of the series [10].

COURSE DEVELOPMENT

ComSoc has an impressive history in offering training programs. When I became the Director of Education and Training, we had a rich offering in the wireless communications area, thanks to the Wireless Communications Engineering Technologies (WCET) Certification program. As part of this program, developed thanks to the tremendous support provided by ComSoc staff, several courses were already in existence, ranging from basics of wireless communications to more advanced topics.

In order to bring this offering to a broader audience, especially to people who may not necessarily be interested in being certified but want to get quick, focused, and effective training on hot topics and new emerging areas, we worked hard to renew this offering, adding new courses that covered new topics that will become important in the near future. Figure 2 offers an example of our current offering, available at <http://www.comsoc.org/training/training-calendar>.

Three new two-day courses were developed in 2014–2015, with the support of the IEEE FDC Incubator Funds: Ted Rappaport of NYU taught an online course on 5G Millimeter-Wave Cellular Systems in December 2014; Ashutosh Dutta of AT&T Labs taught an online course

The screenshot shows the website interface for the Hands-on Lab Exchange. At the top, there is a navigation bar with links for HOME, ABOUT, SUBMIT YOUR COURSE, COURSE LIBRARY, and AUTHOR PROFILES, along with a search function. The main content area is divided into two columns. The left column features the course title, author information (Dr. Robert G Mauder), and a detailed description of the course content, which includes Matlab syntax, design, and simulation. Below the description is a 'PROJECT FILES' section with two categories: 'OPEN ACCESS' and 'INSTRUCTORS ONLY'. The 'OPEN ACCESS' section contains links to 'CourseworkInstructions.pdf (182.87 KB)' and 'AccompanyingMatlabCode.zip (5.5 KB)'. The 'INSTRUCTORS ONLY' section contains a link to 'AutomatedMarking.zip (65.7 KB)'. At the bottom of the page, there is a 'SHARE THIS:' section with social media icons for Twitter, Facebook, LinkedIn, and Google+. The footer includes the IEEE ComSoc logo and copyright information for the IEEE Communications Society.

Figure 1. Example of the interface for Hands-on Lab Exchange.

on Mobility Protocols in June 2015; Amarnath Gupta of the University of California, San Diego developed a combined in-person/online course on Big Data. In July 2015, thanks to some additional support provided by IEEE TAB, the ETB approved seven additional new courses that were developed and taught in 2015 and 2016: a new Internet of Things (IoT) course by Lee Vishloff; an Intro to WiFi by Daniel Wong; a new satellite course by Bruce Elbert; A Battle in Unlicensed Spectrum: The Path of LTE and WiFi in 3.5 GHz and 5 GHz Bands, by Jonathan Levine; Wireless Positioning in 4th Generation Cellular and Beyond, by Alan Bensky; The Tactile Internet by Gerhard Fettweis and Frank Fitzek; Designing the Green Internet: Energy Efficient Design of Access, Backbone and the Cloud, by Fabrizio Granelli.


We strive to keep updating existing courses and adding new ones, also relying on participants' feedback, and plan to also expand the scope of the offering to include all areas of interest to ComSoc members. I truly believe that ComSoc, thanks to the continuous effort of its staff and its volunteers, is doing a great job in this area, and can be looked on as a model for other societies and for the IEEE as a whole. ComSoc's members are working with the IEEE Education Activities Board and with the IEEE Products and Services Committee to bring this experience and track record to a larger community.

STUDENT SUMMER SCHOOL

At the beginning of my term, we were solicited by the ComSoc BoG to consider the possibility of creating a Summer School for Ph.D. students in communications. The example that was brought to our attention was that of the IEEE Magnetics Society, which had been running a very successful such program for many years. After consulting with our colleagues from that society and also running a feasibility study of what could be done and in what form, the ETB decided to proceed with such an initiative and to run an experimental edition of a ComSoc Summer School.

The first ComSoc Summer School was organized by Fabrizio Granelli and successfully held in Trento, Italy, on July 6–9, 2015. The event got extensive coverage in *IEEE Communications Magazine* and the *Global Communications Newsletter*. Informal feedback by speakers and attendees was extremely positive, with several requests to continue this initiative and make it a yearly event. The School had 43 attendees, of which 28 got free lodging and meals. A second edition was run in Trento in July 2016, to consolidate the event and rely on its well tested format and organization. This second edition was also very successful, and now the ComSoc Summer School initiative is mature enough to be brought around the world, thereby making it more accessible to people from various places. A more extensive description of this initiative is given in another article in this issue of the Series [11].

Starting this project was really a wonderful intuition, as confirmed by how well it was received by the community. It also sends a clear message that ComSoc wants to invest in students, who represent its future.



Training Calendar

Date	Course Format	Title
July 27, 2016	Online Course	Network Function Virtualization (NFV), Software-Defined Networking (SDN) and the Road to 5G
August 17, 2016	Online Course	Broadband Mobile Satellite Communications
August 24, 2016	Online Course	An Introduction to Wi-Fi
September 14, 2016	Online Course	High Throughput Satellites
September 20, 2016 - September 23, 2016	Online Course	2 for 1 Deal: September 4 Day Intensive Wireless Training Course PLUS Fall 2016 WCET Exam Seat
September 20, 2016 - September 23, 2016	Online Course	4 Day Intensive Wireless Communications Course
September 28, 2016	Online Course	Wireless for the Internet of Things (IoT)
October 19, 2016	Online Course	Machine-to-Machine (M2M) Essentials

Figure 2. Recent schedule of ComSoc's training activities.

OPPORTUNITIES FOR COMSOC EDUCATION ACTIVITIES WORLDWIDE

Although the education programs of ComSoc have been very effective in serving North American audiences, there is a clear desire and need to expand this reach well beyond these boundaries, and to reach out to different countries and regions of the world.

In these two years, we have explored different opportunities, and worked on two concrete possibilities to offer ComSoc Education products internationally. One has to do with a request that was received from the Nile University in Egypt, a recently established university that has strong ties with industry and may be interested in promoting ComSoc's WCET program locally. Some promotional material was sent, and the discussion is still ongoing. The other opportunity is about the possibility to offer WCET courses in India. A discussion is ongoing with the IEEE India office, and in November 2015, a ComSoc delegation visited Bangalore and met with local people interested in promoting such a project. One item in the discussion was an initiative to customize the training program for Indian needs, which looks like a very good opportunity.

While these initiatives have their own challenges (including financial issues as well as the required background knowledge for students to benefit from the current course offerings), it is clear that the future of ComSoc Education and Training must deal with the global angle. We are lucky to have dedicated and motivated volunteers all around the globe, providing invaluable points of contact (and points of view) to start discussions, and eventually to run programs in their own environment and cultural ambience.

THE WAY FORWARD

I was lucky to receive such a rich heritage from my predecessors, Dave Michelson and, before him, Stefano Bregni. They started what turned out to be a long process of renewal and revitalization of the Educational Services a society like ComSoc is able to provide to its members and to the community at large. I was fortunate enough to be part of this process at a stage where a lot of ideas were beginning to flourish. Although this meant that a lot of implementation work was needed, this also allowed me to participate first-

Although the education programs of ComSoc have been very effective in serving North American audiences, there is a clear desire and need to expand this reach well beyond these boundaries, and to reach out to different countries and regions of the world.

hand in the fruits of all these efforts, and to see many of these initiatives come to completion.

As of January 2016, I myself left this legacy to my successor, Rulei Ting, who had served as a member of the ETB for several years and is now the new Director. In wishing him all the best, and in guaranteeing him all the support I can provide, I am sure he will bring Educational Services to even higher standards, making education really the “third pillar” of ComSoc for many years to come.

ACKNOWLEDGMENTS

I would like to thank the people in ComSoc who entrusted me with the role of Director of Education and Training at such a critical time, then President Sergio Benedetto and Vice-President of Technical Activities Khaled Ben Letaief. I would also like to thank all members of the Education and Training Board during my term (Koichi Asatani, Periklis Chatzimisios, Michael Devetsikiotis, Ashutosh Dutta, Tarek El-Bawab, Fabrizio Granelli, Yves Lostanlen, Erik Luther, David Matolak, Neelesh B. Mehta, Dave Michelson, and Rulei Ting), the staff support (John Pape, Marilyn Catis, and Tara Gallus), and ComSoc’s Executive Directors under whom I served (Jack Howell and Susan Brooks).

If this was not sufficiently clear from the above description of the various activities and initiatives that were carried out during these two years, I want to make sure they get the credit they entirely deserve for taking ownership of the various projects and bringing them to brilliant completion, with dedication, competence, and diligence that show how our people (both volunteers and staff) are our most precious resource.

REFERENCES

- [1] S. Benedetto, K. Ben Letaief, and M. Zorzi, “Education and Training: The Third Pillar of ComSoc,” *IEEE Commun. Mag.*, vol. 53, no. 5, May 2015, pp. 6–8.

- [2] T. S. El-Bawab *et al.*, “Commentary: Toward Specialized Undergraduate Telecommunication Engineering Education in the U.S.,” *IEEE Commun. Mag.*, vol. 50, no. 9, Sept. 2012, pp. 14–16.
- [3] “Telecommunications Engineering Is Now a Distinct Education Discipline,” *IEEE Institute*, Nov. 21, 2014.
- [4] T. S. El-Bawab, “Telecommunication Engineering Education (TEE): Making the Case for a New Multidisciplinary Undergraduate Field of Study,” *IEEE Commun. Mag.*, vol. 53, no. 11, Nov. 2015, pp. 35–39.
- [5] T. El-Bawab on IEEE ComSoc Beats with Michele Zorzi, <https://www.youtube.com/watch?v=Q1gdmioqPgY>
- [6] D. G. Michelson, D. W. Matolak, and W. Tong, “Communications Education and Training: Software Defined Radio,” *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 182–83.
- [7] D.G. Michelson, M. Trocan, and W. Tong, “Communications Education and Training: Expanding the Student Experience,” *IEEE Commun. Mag.*, vol. 52, no. 12, Dec. 2014, pp. 96–97.
- [8] D.G. Michelson, W. Tong, and L. Reeves, “Communications Education and Training: Student Competitions,” *IEEE Commun. Mag.*, vol. 53, no. 5, May 2015, pp. 194–95.
- [9] D. G. Michelson, W. Tong, B. L. Shoop “Communications Education and Training: Ethics and Professionalism,” *IEEE Commun. Mag.*, vol. 53, no. 11, Nov. 2015, pp. 14–16.
- [10] R. Bowley, “The ComSoc Hands-on Lab Exchange,” *IEEE Commun. Mag.*, this issue.
- [11] F. Granelli, “Training and Networking for Young Society Members: the ComSoc Summer School Program,” *IEEE Commun. Mag.*, this issue.

BIOGRAPHY

MICHELE ZORZI [F’07] (zorzi@dei.unipd.it) received his Laurea and Ph.D. degrees in electrical engineering from the University of Padova in 1990 and 1994, respectively. During academic year 1992–1993 he was on leave at the University of California San Diego (UCSD). After being affiliated with the Dipartimento di Elettronica e Informazione, Politecnico di Milano, Italy, the Center for Wireless Communications at UCSD, and the University of Ferrara, in November 2003 he joined the faculty of the Information Engineering Department of the University of Padova, where he is currently a professor. His present research interests include performance evaluation in mobile communications systems, random access in mobile radio networks, ad hoc and sensor networks and IoT, energy constrained communications protocols, 5G millimeter-wave cellular systems, and underwater communications and networking. He was Editor-in-Chief of *IEEE Wireless Communications* from 2003 to 2005, Editor-in-Chief of *IEEE Transactions on Communications* from 2008 to 2011, and is currently the founding Editor-in-Chief of *IEEE Transactions on Cognitive Communications and Networking*. He has been a Guest Editor for several Special Issues of *IEEE Personal Communications*, *IEEE Wireless Communications*, *IEEE Network*, and *IEEE JSAC*. He served as a Member-at-Large in the Board of Governors of the IEEE Communications Society from 2009 to 2011, and as its Director of Education from 2014 to 2015.

The ComSoc Hands-on Lab Exchange

Rhys Bowley, Erik Luther, and David G. Michelson

ABSTRACT

Laboratory assignments and tutorials have long been recognized as an important component of science and engineering education, akin to the tradition of apprenticeship. The ComSoc Hands-on Lab Exchange (labs.comsoc.org) is designed to provide a formal mechanism for sharing course materials, projects, and best practices relevant to lab-based communications education. Materials shared through the Exchange provide an opportunity for peer review against specified criteria. Accordingly, the Exchange benefits both the contributor, who can gain both helpful feedback and third party endorsement of the work's value, and future users of the material, who are assured of the work's completeness and quality.

INTRODUCTION

In recent years, the Scholarship of Teaching and Learning (SoTL) movement has sought to promote application of principles and processes normally reserved for research activities to teaching and learning. Foremost among these principles and processes are the notions that the teaching and learning experience should be shared so that others may build on and benefit from previous efforts [1]. The *IEEE Communications Magazine* Feature Series on Education and Training has been created to provide a forum for sharing such experiences in the area of communications education.

Laboratory assignments and tutorials have long been recognized to be an important component of science and engineering education. By providing direct exposure to the equipment, processes, principles, and outcomes associated with a particular discipline, they provide engineering students with opportunities to develop intuition and appreciation for the field that cannot be matched by lectures or simulation. However, the effort required to develop suitable experiments and associated documents for communications education can be considerable. Historically, there have been few formal mechanisms to facilitate sharing of laboratory assignments so that others may build on and benefit from previous efforts. Anecdotes suggest that most sharing of laboratory assignments between instructors to date has been informal.

The ComSoc Hands-on Lab Exchange has been designed to provide a formal mechanism for sharing laboratory or project assignments relevant to communications education. The assign-

ments that are shared through the Exchange are published in two stages. In the first stage, submitted content is immediately made available to the user base so that it may be peer reviewed and helpful feedback provided. In the second stage, once the peer review process is complete, the submission is marked as such. Accordingly, the process benefits the academic contributor, who can receive useful feedback and claim this review in their teaching dossier; the industrial contributor, who can receive useful feedback and gain third party endorsement of the work's value; and future users of the material, who are assured of the work's completeness and quality. The remainder of this article is concerned with the background and motivation for the Exchange, a description of its essential elements, a summary of the review process and governance issues, and anticipated next steps.

BACKGROUND AND MOTIVATION

The ComSoc Hands-on Lab Exchange was inspired by:

- ComSoc Education & Training Board (ETB) (now the ComSoc Educational Services Board [ESB]) interest in promoting SoTL in communications education
- Opportunities presented by various ComSoc ETB activities designed to popularize the use of software defined radio (SDR) in communications education
- Previous efforts to share electrical engineering laboratory experiments

ComSoc ETB/ESB interest in SoTL is described in detail in a companion article in this Feature Topic [2]. Between 2012 and 2014, the ComSoc ETB organized a series of events within the Industry Forums and Exhibitions (IF&E) at IEEE ICC and IEEE GLOBECOM that focused on the role of SDR in communications education. Presenters at these events included both developers of SDR hardware and software from National Instruments, MathWorks, Ettus Research, and university professors who have developed SDR-based lab courses. In May 2014, the first edition of the new Feature Series on Communications Education and Training in *IEEE Communications Magazine* shared four case studies of lab-based courses based on SDR.

While the case studies presented at conferences or in the magazine help prospective instructors appreciate the opportunities and pitfalls of SDR in education, and gain a use-

The ComSoc Hands-On Lab Exchange (labs.comsoc.org) is designed to provide a formal mechanism for sharing course materials, projects, and best practices relevant to lab-based communications education. Materials shared through the Exchange provide an opportunity for peer review against specified criteria.

The ETB considered the matter carefully and concluded that a more flexible solution was required. Given the fast moving nature of the field, compiled volumes were likely to become outdated fairly quickly. A successful case was made that a web-based solution would allow more timely publication and would likely encourage more instructors to participate.



Figure 1. A mockup of a typical course page on the Hands-On Lab Exchange.

ful appreciation of both the instructor and student experience, there was a clear need to share the actual course documents: lab and pre-lab assignments, source code, scoring rubrics, and solutions or sample results. One possibility was to collect SDR and other communications lab assignments from an open call and publish them in a freely distributed document as the IEEE EMC Society had done in their EMC Experiments Manual years before [3, 4]. Another was to support publication of entire lab courses through IEEE Press or another publisher, following the lead of [5]. In both cases, ComSoc would add value by arranging for third party review, and ensuring the completeness and quality of the finished product.

The ETB considered the matter carefully and concluded that a more flexible solution was required. Given the fast moving nature of the field, compiled volumes were likely to become outdated fairly quickly. A successful case was made that a web-based solution would allow

more timely publication and was likely to encourage more instructors to participate. This led to the design, development, and deployment of the ComSoc Hands-on Lab Exchange website [6]. Some aspects of its design were inspired by the IEEE Real World Engineering Projects website [7] but tailored to meet the needs of lab-based courses.

THE HANDS-ON LAB EXCHANGE

The essential elements of the Hands-on Lab Exchange are depicted in the mockup in Fig. 1. The title, author(s), abstract, and representative image are presented in the upper half of the page. Submissions are clearly identified as either academic or industry contributed. Materials intended for students or trainees are presented in an open access area on the lower left side of the page. Solutions and other materials intended for instructors are presented in a protected area on the lower right side of the page. Access to instructor materials is only granted to instructors

who are IEEE members and whose instructional role can be verified, for example, against a public university website.

Acceptable materials fall into two categories: Lab Experiments and Lab Courses. A Lab Experiment is a standalone activity with materials to assist in student assessment. A Lab Course is a collection of Lab Experiments. Acceptable topics include, but are not limited to, hands-on lab assignments, projects, or tutorials on:

- Software tools useful in communications
- Hardware tools useful in communications
- Test and measurement equipment useful in communications
- Analog and digital communications
- Advanced physical (PHY) later digital communications
- The medium access control (MAC) layer in digital communications
- SDR
- Software defined networking
- Optical communications networks

The minimum submission to the Exchange is a Lab Experiment. A Lab Experiment submission must include:

- A list of the required hardware, software, and test and measurement equipment
- Lab instructions (pre-lab, hands-on activity, assessment tool)
- Instructor resources (instructor hints, assessment rubrics, solutions, and sample results)

The submission must include evidence that the materials have been used in a course setting at least twice and a brief description of how the materials were refined or improved based on that experience.

THE REVIEW PROCESS

The ComSoc ESB manages the submissions approval process. The first step is a checklist review to ensure that all of the requested materials have been supplied. The second step is a content review by anonymous peer reviewers to ensure completeness and correctness. It will usually result in helpful requests for clarification and friendly suggestions for revision. If all is in good order, the revised materials will be approved for publication and marked with an "IEEE ComSoc peer-reviewed" badge. As noted earlier, the process benefits the academic contributor, who can claim this review in their teaching dossier; the industrial contributor, who can gain third party endorsement of the work's value; and future users of the material, who are assured of the work's completeness and quality.

COPYRIGHT AND INDEMNIFICATION

Usually, IEEE requires that authors transfer copyright of their manuscripts before the work is published. In the case of the Hands-on Lab Exchange, the author retains ownership of the copyright but indemnifies IEEE and retains responsibility for any liability that may arise. Some authors may choose to release their materials under a Creative Commons license in order to promote sharing. Upon submission, the author would be required to grant unrestricted worldwide rights to the IEEE Communications Society

to disseminate the course materials in whole or in part at the Society's sole discretion. However, if the author chooses to revoke permission, the Society will remove the content from the website within 30 days. Under the agreement, the Society is not responsible for notifying users of updates, revisions, corrections, or termination of available content.

NEXT STEPS

The Hands-on Lab Exchange will benefit the communications education and training community by:

- Encouraging the development and dissemination of high-quality lab-based instructional materials
- Providing a mechanism for sharing the experience and insights gained in the development of such materials

With the website operational, the next steps are to encourage instructors to consider submitting their communications-oriented lab assignments, projects, and tutorials to the site, or make use of posted materials to inform and advance their own laboratory teaching practices.

REFERENCES

- [1] K. McKinney, *Enhancing Learning Through the Scholarship of Teaching and Learning*, Anker Publishing, 2007, 201 pp.
- [2] D. G. Michelson, "Integrating the Scholarship of Teaching, Learning and Research Supervision into Communications Education," *IEEE Commun. Mag.*, vol. 54, no. 11, Nov. 2016.
- [3] Education Committee, "EMC Education Manual," IEEE EMC Soc., 1992.
- [4] Education and Student Activities Committee, "EMC Experiments and Demonstrations Guide," IEEE EMC Soc., revised May 2015.
- [5] R. W. Heath, Jr., *Digital Wireless Communication: Physical Layer Exploration Lab Using the NI USRP*, Nat'l. Tech. and Science Press, 2012.
- [6] IEEE Communications Society, Hands-on Lab Exchange, <http://labs.comsoc.org/>.
- [7] IEEE, "Real World Engineering Projects," <http://www.realworldengineering.org>.

ADDITIONAL READING

- [1] D. G. Michelson, D. W. Matolak, and W. Tong, Eds., "Communications Education and Training: Software Defined Radio," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 182–209.

BIOGRAPHIES

RHYS BOWLEY (rhys.bowley@ni.com) works as a product marketing engineer at National Instruments (NI) as part of the academic team based in Austin, Texas. A graduate of Cardiff University, he holds a Master's in electronics and communications engineering and has been involved in delivering engineering education through a mixture of volunteer and professional efforts since 2007.

ERIK LUTHER (KF5LTV), senior group manager – 5G Prototyping Solutions, leads the 5G product marketing team at NI focused on accelerating next generation wireless research. Over the last five years he has managed product marketing for NI and Ettus Research software defined radio solutions including product roadmaps, outbound marketing, and collaborations with leading industry, academic, and government wireless research teams. Early in his career, he pioneered NI's efforts to support universities with curriculum and textbooks, launching NI's independent textbook publishing arm NTS Press. He worked closely with the IEEE Communication Society Education & Training Board to establish <http://labs.comsoc.org>, a community focused on establishing best practices for hands-on education and teaching resources for wireless communications.

DAVID G. MICHELSON leads the Radio Science Lab at the University of British Columbia where his research interests focus on wireless communications. His service to ComSoc has been recognized by ComSoc Chapter Awards, IEEE MGA's Outstanding Large Section award, and IEEE Canada's E. F. Glass Award. He completed the UBC Faculty Certificate Program on Teaching and Learning in Higher Education in 2011, served as ComSoc's Director of Education and Training from 2012 to 2013, and has served as a member of the ComSoc Educational Services Board from 2014 to present. He is also an elected member of the ComSoc Board of Governors (2013–2015 and 2017–2019).

With the website operational, the next steps are to encourage instructors to consider submitting their communications-oriented lab assignments, projects, and tutorials to the site, or make use of posted materials to inform and advance their own laboratory teaching practices.

Training and Networking for Young Society Members: The ComSoc Summer School Program

Fabrizio Granelli

Education represents the third pillar of the IEEE Communications Society and a key issue in training the next generation of communications engineers. With the aim of providing up-to-date knowledge about the hot topics and recent developments in the field of communications, ComSoc started a new initiative: ComSoc Summer Schools for Ph.D. students. This article provides a report on the implementation of the first two editions of the ComSoc Summer School and some hints as to what will come next.

ABSTRACT

Education represents the third pillar of the IEEE Communications Society and a key issue in training the next generation of communications engineers. With the aim of providing up-to-date knowledge about the hot topics and recent developments in the field of communications, ComSoc started a new initiative: ComSoc Summer Schools for Ph.D. students. The core idea is to enable the most promising young members of the Communications Society to get in touch with top-level experts in the field, as well as to network with one another. This article provides a report on the implementation of the first two editions of the ComSoc Summer School and some hints as to what will come next.

INTRODUCTION

With the introduction of education and training as the “third pillar” of the IEEE Communications Society and the definition of the Sub-Committee on Education Programs [1], Content and Services, the IEEE Communications Society has aimed to offer a world class training and professional education program, addressed primarily to ComSoc’s members (both current and prospective), providing high-quality instruction, at a reasonable cost and with easy access, to address the career needs of working professionals in communications and related fields.

There is no doubt that students are the future of our Society. It is therefore of fundamental importance that we provide special membership development opportunities for them. To do so, the IEEE Communications Society launched a new Summer School program, the objectives of which are to:

- Provide high-quality courses on selected topics in our field
- Engage local chapters in membership development and educational activities
- Link distinguished lecturers to relevant membership development activities
- Potentially develop high-quality tutorial materials that can be disseminated online to the larger ICT community

The idea of organizing a Summer School for Ph.D. student members was proposed by IEEE

ComSoc’s former President Prof. Sergio Benedetto, Prof. Khaled Letaief, Vice-President of Technical Activities, Prof. Michele Zorzi, Director of Education, and Prof. Stefano Bregni, Vice-President for Member Relations.

This article provides an overview of the current status of the ComSoc Summer School initiative and information about possible future developments of the program.

The article is structured as follows. The next section provides a general description of the contents provided by attendees of the Summer School. Then we provide a summary of the first two implementations of the program, in 2015 and 2016, respectively. After that we provide some lessons learned, and the final section concludes the article.

THE COMSOC SUMMER SCHOOL PROGRAMME

In order to achieve its goals — to bring together the best young members of the Society, enable them to network, and expose them to the top experts in the field of communications — the program of the ComSoc Summer Schools should incorporate the following features:

- Lectures of hot topics in communication by top international experts
- “Hands-on” sessions
- A poster session for the attendees
- Social events and opportunities for networking among participants

The next subsections provide additional details on the items above.

HOT TOPICS COVERAGE

The field of communications is evolving at an extremely fast pace, and high-level education institutions struggle to provide effective material and courses on recent advances coming from research activities. The ComSoc Summer School program is exactly aimed at bridging the gap between the solid background knowledge provided by traditional education and the best of what research offers. To achieve this goal, the Summer School requires support from the top expert members of the IEEE Communications Society to expose attendees to the current hot topics and trends, as well as possible subjects for further investigation within their research activities.



Figure 1. Opening of the first ComSoc Summer School.

The majority of the Summer School activities fall within this category in order to enable the attendees to capture the latest results and possible next steps forward in communications. Topics are presented in such a way as to enable fruitful exposure by attendees with a general communications background.

“HANDS-ON” SESSIONS

In several situations, researchers — especially during their first years — focus completely on their own research topic or issues, with the risk of losing contact with the day-to-day practices of the communication professional. To bridge this gap is extremely complex and surely not possible in a short period of time, but nevertheless the Summer School is expected to include some sessions related to more practical or “hands-on” activities.

POSTER SESSIONS

A key aspect in the work and research in communications is networking, that is, the capability to professionally interact and work with other member of the community.

As all attendees are registered Ph.D. students addressing specific topics in the field of research in communications, one of the requirements for participation in the Summer School is the presentation of a poster with current results or research plans on a research subject.

Presenting their own papers allows the participants to better know each other and explain their work to their colleagues, and provides a promising venue for building collaborations and additional opportunities for networking.

Moreover, the poster session is advertised to local Ph.D. schools and departments to enable local researchers to interact with the Summer School participants.

SOCIAL PROGRAMME

For the purpose of allowing high degree of interactions, the program is designed in such a way as to have social dinner(s) as well as free slots for attendees to freely organize and converse.

PREVIOUS EDITIONS

The first two editions of the ComSoc Summer School program were held in Trento, Italy.

The city of Trento was selected for testing the first editions of the IEEE ComSoc Summer School. The city, in the area of northeast Italy, was chosen because Trento is a lively place with several international initiatives in the field of education and research. Indeed, the city of

Trento is consistently in the first position in its country for its high quality of life and hosts the University of Trento, one of the youngest Italian universities (founded in the 1960s), as well as advanced research centers and renowned international cultural institutions.

The events were hosted at the Department of Information Engineering and Computer Science, the second highest ranked ICT Department in Italy, in the hills above the city.

For both events, Communications Society covered participation costs (lodging and meals) for around 30 participants, while the remaining ones paid a registration fee of €200 to cover organization costs.

2015 EDITION

The first IEEE ComSoc Summer School [2] was held in Trento on July 6–9, 2015. A total of 43 international participants attended the Summer School at the University of Trento (Figs. 1 and 2).

Participants were selected during the Spring term (between March and April) through an open call posted on the IEEE Communications Society website. More than 100 Society members applied to the call, and participants were selected based on scientific merits, and enabling both early stage and mature Ph.D. students to attend and interact.

The topics selected for the first edition of the ComSoc Summer School were future wireless networks, software defined-networking, and big data. To this goal, the following seminars were invited:

- “Collaborative Near-Capacity Wireless System Design” by Prof. Lajos Hanzo, University of Southampton, United Kingdom
- “From Dumb to Smarter Switches in Software Defined Networks: Towards a Stateful Data Plane” by Prof. Giuseppe Bianchi, University of Rome Tor Vergata, Italy
- “The Next Wave in Wireless Communications” by Prof. Andrea Goldsmith, Stanford University, United States
- “Networking for Big Data” by Prof. Nelson L.S. da Fonseca, State University of Campinas, SP, Brazil.

The first seminar (“Collaborative Near-Capacity Wireless System Design”) provided an effective introduction to future wireless communications systems. The first topic discussed was multiple-input multiple-output (MIMO) technology, underlining its limitations and proposing virtual antenna arrays (VAAs) as a way to maintain MIMO properties but exploiting distributed

As all attendees are registered Ph.D. students addressing specific topics in the field of research in communications, one of the requirements for participation to the Summer School is the presentation of a poster with current results or research plans on a research subject.

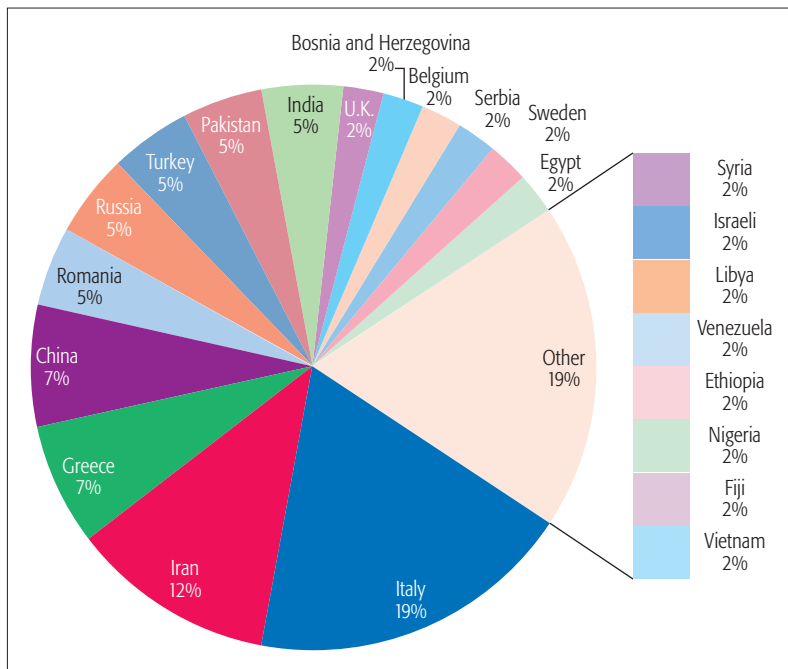


Figure 2. Participants in the first ComSoc Summer School by country of origin.

elements. Nevertheless, VAAs require the design of specific amplify-forward and decode-forward protocols for effective channel coding, leading to more sophisticated three-stage-concatenated iterative channel coding schemes. In this framework, the seminar demonstrated that near-capacity solutions were possible through the introduction of EXIT-chart-aided designs, and was closed with a discussion on future research directions and open problems.

The program continued the following day with the second seminar, focusing on the transition between traditional networks and software defined networks (SDNs): “From Dumb to Smarter Switches in Software Defined Networks: Towards a Stateful Data Plane”. Indeed, the seminar analyzes a major issue in today’s SDN architectures, which is the requirement to mandate the execution of all control tasks to a remote controller, leading to potential challenges in terms of latency and signaling overhead. The first section of the seminar provides an introduction to SDN principles as well as a review of Openflow, currently the main SDN protocol. After discussing recent OpenFlow developments (recent OpenFlow extensions, reconfigurable match tables, protocol oblivious forwarding, etc.), the seminar discussed OpenState, a recent proposal designed to allow the introduction of stateful programmable tasks. This provided the attendees with an effective overview on the current status of SDN and possible near future development.

A poster session was organized on July 10 in the afternoon, to enable attendees to introduce their profile and current or future research plans. The session was designed to allow attendees to interact among themselves and get to know other participants better. To encourage discussion, the session was open to all students of the local doctoral school as well as to the professors and researchers of the ICT Department.

Future wireless communications were at the center of the following seminar, “The Next Wave in Wireless Communications.” The lecturer, Prof. Andrea Goldsmith, started the seminar by introducing the fundamentals of wireless communications and providing an effective survey on the design principles of current cellular systems. Then the speech described the possible future scenarios for wireless network deployment, by introducing the concepts of device-to-device communications and the Internet of Things. The scarcity of communication resources for such emerging applications leads to the definition of cognitive radio networks as a potential paradigm to provide additional flexibility in the management of the transmission spectrum.

The seminar closed with the novel concepts related to object connectivity to the Internet and power harvesting communications, as well as the interdisciplinary potential of the communications theory results (e.g., in the field of neural sciences).

The last seminar was focused on “Networking for Big Data.” The speech introduced the basic concepts of the big data paradigm and vision, and discussed its ecosystem and related requirements. Then the speaker, Prof. Nelson L.S. da Fonseca, focused on the communication aspects related to the big data scenario and identified network virtualization as a fundamental requirement to provide effective support to big data applications. Finally, research subjects in both the framework of networking for big data and big data for networking were introduced and discussed. In this case, the seminar content was complemented by lab exercises that were provided to attendees for additional practice offline.

As previously discussed, the lecture-style component of the ComSoc Summer School was complemented by more interactive and “hands-on” sessions. In addition to the poster session already described, the school program included two visits:

- A visit to the main data center of the University of Trento
- A visit to the local infrastructure and network provider, Trentino Network, and its network operation center

The purpose of the events was two-fold:

- To enable participants to understand the challenges, design goals, and dimensioning methods of a real network and network infrastructure
- To encourage interaction with networking experts operating on intranet and Internet infrastructure

2016 EDITION

The second edition of the ComSoc Summer School was held in Trento, Italy, on June 20–23, 2016.

This edition was intended to consolidate the event, and demonstrated consistent figures with respect to the previous year, with more than 100 applicants and 41 participants from all the world (Fig. 3).

The topics selected for the second edition of the ComSoc Summer School were the Green Internet, and future and smart networks. The following seminars were invited:

- “Network Coding and Compressed Sensing Enabling the Tactile Internet” by Prof. Frank Fitzek of the University of Dresden, Germany
- “Recent Advances in TCP” by Prof. Reuven Cohen from Technion, Israel
- “Data Center Networking” by Prof. Suresh Subramaniam, from George Washington University, United States
- “Design and Performance of a Smarter Infrastructure: Smart Energy, Smart Buildings and Electric Vehicles” by Prof. Michael Devetsikiotis North Carolina State University, United States
- “Energy Efficiency in Cloud Networks” by Prof. Jafaar Elmirghani from the University of Leeds, United Kingdom

The program started on June 20 with the seminar on “Network Coding and Compressed Sensing Enabling the Tactile Internet” by Prof. Frank Fitzek (Fig. 4). The focus of the seminar was the tactile Internet, which is expected to enable a global network to control and steer the Internet of Things in real time. The fifth generation (5G) will be the first generation of mobile communication systems that will enable the tactile Internet. The course addressed the need for the tactile Internet and described the need for new technologies such as network coding and compressed sensing to break with the commonly accepted trade-offs between throughput, latency, resilience, and security. The course also highlighted the main enabling technologies such as network coding and compressed sensing.

On the same day, Prof. Reuven Cohen discussed “Recent Advances in TCP.” As is well known, TCP, the Internet’s transport protocol, has played a critical role in the evolution of the Internet, and in the last 20 years, more than 100 RFCs have offered ways to improve its performance. This seminar analyzed the most recent advances in TCP, including the new congestion control scheme used by Android (TCP CUBIC) and the new MPTCP extension used by iPhone devices, which allows a client to establish multiple connections to the same server over different network adapters. QUIC, a new transport protocol that may replace TCP in the near future, developed by Google, was also introduced.

“Data Center Networking” was the topic of the seminar of the second day of the event, presented by Prof. Suresh Subramaniam. The tutorial focused on data center networking by introducing data centers and their requirements. Then a detailed presentation of several data center network architectures and their relative merits was provided. Conventional electrical switching architectures such as fat tree, flattened butterfly, and VL2 were presented and compared. Subsequently, the tutorial examined several recent research proposals on incorporating optical switching in the data center network. The presentation concluded with a discussion of protocol and performance issues and some emerging topics.

On June 22, the Summer School hosted Prof. Michael Devetsikiotis for his seminar on “Design and Performance of a Smarter Infrastructure: Smart Energy, Smart Buildings and Electric Vehicles.”



Figure 3. Participants in the second ComSoc Summer School.



Figure 4. A moment in the lecture by Prof. Fitzek during the second ComSoc Summer School.

The seminar was focused on designing a larger and smarter infrastructure. This includes the modeling of smart buildings, interactive spaces, and smart cities, enabled by the Internet of Things. The presentation adopted a “cyber-physical” viewpoint for designing, analyzing, and simulating smart building communications, energy, and transportation systems, focusing attention on the performance evaluation of socio-technical systems of multi-layered scope, such as virtual collaboration and indoor localization settings, as well as smart grid communications, intelligent buildings, and electric vehicles. The seminar included some specific quantitative models, including queueing models for EV charging stations; performance modeling of indoor-localization smart spaces; IoT and high-performance cloud systems; joint optimization of communications and energy flows in smart buildings; energy storage in campus buildings; and strategies for intelligent EVs on wirelessly-charging highways.

The last seminar of the program was held on June 23, 2016, by Prof. Jafaar Elmirghani on the topic “Energy Efficiency in Cloud Networks.”

The tutorial introduced and discussed a number of measures that can be used to reduce the power consumption of cloud networks. Then network optimization was presented through the use

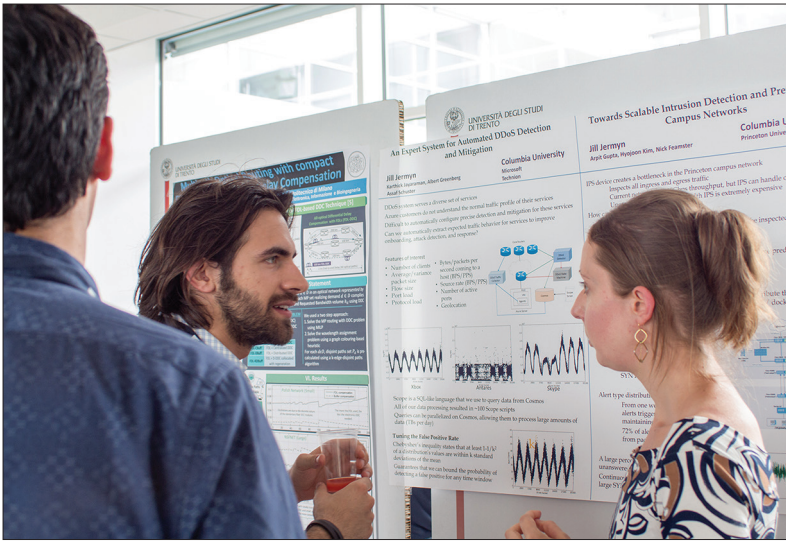


Figure 5. During the poster session of the second ComSoc Summer School.

of mixed integer linear programming (MILP), giving a short tutorial on MILP, and building on this and heuristics inspired by it to explore a number of energy and carbon footprint reduction measures including:

- Optimum use of time varying renewable energy in cloud networks
 - Physical topology design considering operational and embodied energies
 - Elastic optical networks using mixed line rates and optical orthogonal frequency-division multiplexing (OFDM)
 - Optimum resource allocation and green network design with data centers
 - Dynamic energy-efficient content caching
 - Energy efficiency through data compression
 - Energy-efficient peer-to-peer content distribution
 - Energy-efficient distributed clouds
 - Energy-efficient network virtualization
- As in the previous event:
- On Tuesday afternoon, the participants had the chance to present their ongoing and future plans during a two-hour poster session, open to all students of the local doctoral school as well as to the professors and researchers of the ICT department (Fig. 5).
 - Practical sessions were held on Tuesday and Wednesday, and included visits to the data center of the University of Trento and to the local network provider, Trentino Network, and its network operation center.

During the 2016 edition, invited lectures were experimentally recorded and broadcast in real time through the Internet. The number of remote accesses to the live stream of the event were around 70 in the first two days of the event, and then 50 in the third and 25 in the last day. Recorded lectures will be made available to the IEEE Communications Society before the end of the year.

LESSONS LEARNED AND NEXT STEPS

Overall, the ComSoc Summer School program achieved good results, in terms of interest in the program, attendance, and feedback from partic-

ipants (both instructors and attendees). Major points of strengths underlined by participants were the quality of the speakers, the quality of the location, the effective economic support by the Society, and visa/travel support by the local university.

Therefore, the initiative is worthy of becoming a permanent event sponsored by ComSoc. The Summer School would bring benefits to the Society by involving young members and increasing outreach to young researchers; to the participating students by providing them up-to-date training; and to the hosting institutions through international visibility and interaction with international experts.

However, for the organization of future events, it is important to underline relevant issues that should be addressed.

Geographical Location: The incubation of the initiative was done in Trento, Italy, as a location characterized by reasonable living costs, strong support by the local university, and good services. Surely, future events should be organized in different locations worldwide, with the aim of enabling young researchers from different countries to reach the event venue easily.

Visa: One of the issues that reduced the number of actual participants was problems with visas. Such problems surely depend on the hosting country, but effective support to help visa requests is necessary and should be carefully taken into account.

Travel Costs: Another issue is related to travel costs for participants coming from remote areas. Budget-wise, it is not feasible to offer travel reimbursement to participants; therefore, this could make it difficult for attendees from remote areas to participate in person. However, live broadcasting could represent a solution to enable an increase in participation without requiring additional funding or personal expenses for interested young researchers.

Number of Events: With the current model, it is expected to organize one Summer School each year in different locations worldwide. However, an alternative could be represented by organizing a set of smaller events in different regions to minimize the travel distance for potentially interested attendees.

CONCLUSIONS

As education is becoming central for the IEEE Communications Society, several efforts are being implemented to provide value-added services to its members. Within this scenario, this article has described the main idea and current implementation of the ComSoc Summer School program for Ph.D. students, designed to provide up-to-date knowledge about the hot topics and recent developments in the field of communications.

Considering the success of the initiative, the program will be continued in the near future to give continuity to this effort and to enable the most promising young members of the Communications Society to get in touch with top-level experts in the field, as well as to network with one another.

REFERENCES

- [1] S. Benedetto, K. Ben Letaief, and M. Zorzi, "Education and Training: The Third Pillar of Comsoc," *IEEE Commun. Mag.*, President's Page, vol. 53, no. 5, May 2015, pp. 6–8. DOI: 10.1109/MCOM.2015.7105632; URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7105632&isnumber=7105629>
- [2] F. Granelli, "The First IEEE Communications Society Summer School: Trento, Italy, July 6–9, 2015," Society News, *IEEE Commun. Mag.*, vol. 53, no. 8, Aug. 2015, pp. 28–29.

BIOGRAPHY

FABRIZIO GRANELLI [SM] is an associate professor and Delegate for Education at the Department of Information Engineering and Computer Science (DISI) of the University of Trento, Italy, and IEEE ComSoc Director for Online Content. He received his Laurea (M.Sc.) degree in electronic engineering from the University of Genoa, Italy, in 1997, with a thesis on video coding, awarded with the TELECOM Italy prize, and his Ph.D. in telecommunications from the same university in 2001. Since 2000 he has carried on his research and didactical activities at the Department of Information Engineering and Computer Science University of Trento, and coordinator of the Networking Laboratory. Between 2004 and 2015 for a total of six months, he was a visiting professor at the State University of Campinas, Brazil. In 2016, he was a visiting professor at the University of Tokyo, Japan. He was an IEEE ComSoc Distinguished Lecturer

for the period 2012–2015 (two terms). He is an author or co-author of more than 170 papers published in international journals, books, and conferences, focused on networking, with particular reference to network performance modeling, cross-layering, wireless networks, cognitive radios and networks, green networking, and smart grid communications. He was a Guest Editor of the *ACM Journal on Mobile Networks and Applications* Special Issues on WLAN Optimization at the MAC and Network Levels, and Ultra-Wide Band for Sensor Networks, and Co-Chair of 10th and 13th IEEE Workshop on Computer-Aided Modeling, Analysis, and Design of Communication Links and Networks. He was General Vice-Chair of the First International Conference on Wireless Internet and General Chair of the 11th, 15th and 18th IEEE Workshop on Computer-Aided Modeling, Analysis, and Design of Communication Links and Networks '06, '10, '13. He was TPC Co-Chair of GLOBECOM 2007–2009, 2012 Symposia on Communications QoS, Reliability and Performance Modeling, and TPC Co-Chair of GLOBECOM 2014 Symposium on Selected Areas in Communications – Green Communications Track. He is the Chair of the IEEE Standardization Research Group on Software Defined and Virtualized Wireless Access. He was a voting member and contributor of IEEE SCC41 for standards IEEE P1900.1 and IEEE P1900.2. He was an officer (Secretary 2005–2006, Vice-Chair 2007–2008, Chair 2009–2010) of the IEEE ComSoc Technical Committee on Communication Systems Integration and Modeling (CSIM) and is currently Secretary of IEEE ComSoc TC on Transmission, Access and Optical Systems. He is Associate Editor of *IEEE Communications Surveys & Tutorials*, the *IEEE JSAC Green Communications and Networks Series*, and the *International Journal on Communication Systems*.

Integrating the Scholarship of Teaching, Learning, and Research Supervision into Communications Education

David G. Michelson

The author considers how the principles of SoTL can help inform teaching, learning, and research supervision within the engineering disciplines, recommends best practices for planning, conducting, and presenting SoTL-based research in communications education and training, and reveals how the IEEE Communications Society proposes to help promote SoTL activity in communications education and training.

ABSTRACT

Although the SoTL movement has gathered a large following within the academic community in recent years, adoption of SoTL by communications and other engineering disciplines is still in its early stages. While many works have been devoted to SoTL as a scholarly process, relatively few have considered the institutional or discipline-specific context within which SoTL is practiced. Here, we consider how the principles of SoTL can help inform teaching, learning, and research supervision within the engineering disciplines; recommend best practices for planning, conducting, and presenting SoTL-based research in communications education and training; and reveal how the IEEE Communications Society proposes to help promote SoTL activity in communications education and training.

INTRODUCTION

By most measures, today's graduating engineering students are better prepared for career success than any cohort in history. They are confident, communicate effectively, work well together, and can plan ahead. They are well prepared to apply their solid grasp of the fundamentals to the pursuit of practical goals and objectives. Co-op work placements in industry and government have broadened their views of both themselves and their intended profession. Many have benefited from support and encouragement from parents who work in the technology sector. However, the quest to understand the current limitations of engineering education and ensure continuous improvement is undiminished. A recent special focus on the future of engineering and technology education in *IEEE Canadian Review* provides an interesting collection of recent viewpoints on the topic from academic, industry, and student perspectives [1].

For many years, engineering education practice was driven almost exclusively by a combination of institutional edict and instructor experience. During the 1990s, two fundamental shifts in thinking occurred. First, engineering accreditation boards began to take a more aggressive approach in challenging engineering schools to switch their focus from course con-

tent to student outcomes. Second, the notion that the effectiveness of teaching and learning methods could be usefully subjected to scholarly research and inquiry and shared with the community took root and gave rise to the Scholarship of Teaching and Learning (SoTL) movement. Although SoTL has gathered a large following within the academic community in recent years, adoption of SoTL by the engineering disciplines is still in its early stages. While many works have been devoted to SoTL as a scholarly process, relatively few have considered the institutional or discipline-specific context within which SoTL is practiced.

Here, we consider how the principles of SoTL can be usefully applied to establishing best practices for teaching and learning within communications and other engineering disciplines, and its potential contribution to both the accreditation process and research supervision. First, we present a brief history of SoTL. Next, we summarize the essential aspects of planning, conducting, and reporting SoTL research. We then describe how the IEEE Communications Society proposes to help promote SoTL activity in communications education and training. Finally, we summarize the key issues and recommend next steps.

A BRIEF HISTORY OF THE SCHOLARSHIP OF TEACHING AND LEARNING

The manner in which electrical engineering education has evolved over the past century and the best practices that are recommended for engineering educators have been well documented in the engineering literature. Almost 30 years ago, IEEE Press reprinted almost 40 education-themed papers that were published between 1958 and 1983 in a variety of IEEE and non-IEEE publications. *Teaching Engineering: A Beginners Guide* [2] was a landmark volume covering the gamut of relevant topics from the history and goals of engineering education to the learning process and teaching strategies to career development for engineering teachers. The value of the compendium is enhanced by the range of disciplines, perspectives, and timeframes that are presented. Ample space in the volume is allocated to specific educational tasks, including labora-

tory instruction, examinations and grading, and thesis supervision.

The approach taken by the IEEE Press reprint volume was clearly explained in the preface: “Teaching engineering at the college level does not appear to be amenable to recipes and style manuals. [The reprint volume] allows the reader to benefit from the collective wisdom of a number of authors accumulated over the years, and from the resulting variety of viewpoints.” The cited rationale for the volume included:

- The lack of organized training in teaching
- The versatile role of a book
- The different needs of beginning teachers
- The time pressures on engineering teachers

A decade later, IEEE Press followed up with a guide for aspiring and new faculty members [3] that is more tutorial in nature but addresses many of the same issues.

A revolution was brewing, however. In the early 1990s, ABET began the almost 10 years of development that would lead to the release of Engineering Criteria 2000 (EC2000) and an era in which focus shifted from course content to learning outcomes [4]. EC2000 made U.S. engineering schools responsible for developing assessment processes to ensure that their programs produce graduates with the technical and professional skills required by industry. In the years following, most other engineering accreditation bodies around the world have followed suit and developed similar guidelines. Traditional delivery and examination methods have been downplayed in favor of activities involving teamwork and design, and action-based outcome assessment strategies. Many of the new approaches add considerably to the faculty members’ workload. The need to determine the actual value returned and whether the additional level of effort is justified has become increasingly apparent.

Also in the early 1990s, Ernest Boyer of the Carnegie Foundation for the Advancement of Teaching published his landmark volume, *Scholarship Reconsidered: Priorities of the Professoriate* [5]. Boyer presented a powerful case for breaking down the traditional barriers between teaching, research, and service. Among other things, he advocated for the application of a rigorous and scholarly approach to understanding teaching and learning processes and for assessing the improvement gained by introducing new methods and techniques. Many heeded Boyer’s call, and soon a new movement devoted to the Scholarship of Teaching and Learning (SoTL) emerged.

By the early 2000s, books, journals, and conferences devoted to the topic were thriving. While early works focused on the philosophy of SoTL and the manner in which individual researchers can address questions regarding educational practices in a scholarly way [6, 7], more recent works have begun to address its impact at the institutional, national, and disciplinary levels [9–11]. Others have demonstrated how scholarly approaches can be extended to research supervision [12–13]. Some have described this as the Scholarship of Research and Supervision (SoRS). Such efforts can provide helpful guidance to both supervisors and students, and inform the design

of programs designed to accelerate student mastery of research skills and perspectives.

The Scholarship of Teaching and Learning is not exclusively focused on academic education; it can usefully inform industry training practices as well. At the individual level, SoTL can provide educators with a means to test assumptions and verify hypotheses concerning the merits of alternative or improved delivery and assessment methods. This allows educators to both develop their intuition and convince others of the effectiveness of alternative approaches much more quickly. At the institutional level, SoTL can provide administrators with a more effective method to assess the effectiveness of their program decisions and to solicit feedback from individual educators than was previously possible. At the national or disciplinary level, SoTL can provide accreditation bodies with an objective framework within which to engage both educators and administrators, validate or refine policy, and avoid bad decisions. By promoting the sharing of the results of scholarly inquiries into teaching, learning, and research supervision, SoTL helps build an education community in the same way that sharing the results of technical inquiries helps build a research community.

PLANNING, CONDUCTING, AND REPORTING SoTL RESEARCH

Like any other scholarly activity, SoTL research generally seeks to:

- Identify the limitations of past work
- Propose novel, clever, and useful techniques for overcoming these limitations
- Provide convincing evidence, arguments, and insights that the proposed techniques are useful and effective
- Show how the outcomes of the work can be generalized to inform current practice or policy, or suggest future research

Ideally, the evidence will reveal either causal relationships between the proposed teaching and supervision methodologies and desired student outcomes or at least insights concerning the perceived effectiveness of the proposed methodologies. Insights concerning the processes and mechanisms that relate the teaching or supervision methodologies and desired student outcomes or behaviors are also valuable outcomes.

Suggestions and hints for conducting SoTL research have been presented by Weimer [8] and McKinney [9], among others. McKinney’s work is particularly accessible to those beginning SoTL and is strongly recommended. The eight steps in conducting SoTL research usually involve:

1. Identifying a significant limitation of past work taken either from the literature or the educator’s own practice and possible improvements
2. Surveying past work to establish the context for the perceived limitation and the proposed solution
3. Identifying a venue within which to conduct the research, for example, a course, a class, a cohort, an assignment, or even a database
4. Selecting appropriate methods for collecting evidence and data

By promoting the sharing of the results of scholarly inquiries into teaching, learning and research supervision, SoTL helps build an education community in the same way that sharing of the results of technical inquiries helps build a research community.

Informed consent, the right to privacy and protection from harm, for example, potentially ill effects that may arise from deception methods, are key issues that will be considered. However, the vast majority of SoTL research is considered to be low risk and the review panel is not likely to impose many restrictions on the activity.

5. Submitting the project for ethics review and approval by the institution
6. Collecting and analyzing the evidence and data
7. Sharing the preliminary results with peers and inviting critical commentary and evaluation
8. Making the results publicly available for use by others through publication or presentation

Those planning to conduct such work should not be deterred by the relatively modest resources generally available for the purpose. Careful selection of the research question, insightful review of past work, and appropriate choice of research methodology can more than compensate. Nor should prospective SoTL researchers be concerned that research involving human subjects automatically requires institutional review by an ethics review panel. Informed consent, and the right to privacy and protection from harm (e.g., potentially ill effects that may arise from deception methods) are key issues that will be considered. However, the vast majority of SoTL research is considered to be low risk, and the review panel is not likely to impose many restrictions on the activity.

SoTL research may be conducted by a variety of methods [9]. Many SoTL research projects are *case studies* that focus on a single assignment, class, or course, often using multiple research methods, including instructor reflection and student outcomes and/or feedback, over a long period of time. Research methods divide into several categories including direct engagement, evidence review, and experiments and quasi-experiments.

Direct engagement includes interviews and questionnaires that are designed specifically for the purpose of the SoTL research project. *Interviews* may be conducted with individuals or focus groups. They may be used to obtain student perspectives and reactions, and give the possibility of follow up questions to clarify responses and solicit additional information. *Questionnaires* or surveys are simpler and faster to administer than interviews and focus groups, especially where large numbers of respondents are involved, but the opportunities for follow-up are limited. They may be conducted on paper or online, and can easily be anonymized. Such methods can be used to understand or interpret the outcomes observed during evidence review.

Evidence review includes review of activities and materials generated during routine conduct of the course. *Observational methods* may be useful when assessing student behavior or mastery of techniques, but privacy, informed consent and reactivity (potential impact of the process of observation on observed behavior) must be considered. *Content analysis* involves review of student work and identification of trends or patterns in that work. *Secondary analysis* involves use of data that was collected for another purpose that may be available for use in the SoTL research project.

Experiments and quasi-experiments include review of activities and materials generated during portions of the course that are specif-

ically designed to reveal causal relationships between methods and outcomes. True experiments are difficult to design in a classroom setting for both practical and ethical reasons. Quasi-experiments in which the relevant factors are not completely random or uncorrelated are often easier to arrange and frequently substituted.

Reporting of SoTL research follows the same conventions as other scholarly work. For the most part, however, SoTL is an evidence-based activity. Proposals for new courses, assignments or pedagogy by themselves are usually insufficient to qualify as SoTL research. Some indication of the effectiveness of the innovation or the significance of the insights gained from classroom practice must usually be present. Evidence to support formulation of new policy or modification of existing policy would be acceptable. When formulating both the research question and the conclusions, consideration should be given to showing how the outcomes of the work can be generalized to inform current practice or policy, or suggest future research to resolve outstanding questions or issues.

INTEGRATING SOTL INTO COMMUNICATIONS EDUCATION

While there is clear value in adopting SoTL within the engineering disciplines, to either inform the practice of teaching or influence institutional or accreditation policy, efforts in this regard are still in their early stages compared to other academic disciplines. Many works on SoTL tend to be abstract and philosophical or specific to non-engineering disciplines. Until a critical mass of SoTL research in communications has been produced, it will likely be challenging for individual researchers to identify and select good research questions or research methodologies. Individual educators may not have access to enough students, groups, or classes to produce sufficient performance or survey data for reduction and interpretation. Venues for sharing SoTL results with other members of the specific discipline may not be common.

The IEEE Communications Society's Educational Services Board is committed to taking a leadership role in helping to advance the practice of SoTL and SoRS in communications education and training by helping to address the above issues. On the practical side, the Board has established the ComSoc Hands-on Lab Exchange, <http://labs.comsoc.org>, to promote sharing of instructional materials and experiences concerning lab-based courses. On the scholarly side, the Board will promote:

- The formulation and sharing of questions or issues relevant to communications education and training that may be suitable for SoTL-based inquiry
- The formulation and sharing of questions or issues relevant to communications research supervision that may be suitable for SoRS-based inquiry
- The formulation and sharing of questions or issues relevant to accreditation that may be suitable for SoTL-based inquiry

- Cooperation in the planning and execution of multi-institutional SoTL-based studies in order to build critical mass and ensure samples of adequate size to draw useful conclusions
- The presentation and publication of the results of SoTL-based research in venues that are readily accessible to, and promote discussion between, communications educators and trainers

IEEE Communications Magazine has issued a Call for Papers for a Feature Topic on the Scholarship of Teaching and Supervision to be published in November 2017. It is intended to hasten the incorporation of SoTL and SoRS into communications engineering curricula by providing educators and supervisors with an opportunity to share their experience, best practices, and case studies.

CONCLUDING REMARKS

The Scholarship of Teaching and Learning provides engineering educators in both academia and industry with a valuable tool in the quest to improve teaching, learning, and research supervision. By helping to promote the development and assessment of new teaching and research supervision techniques, SoTL serves the community at the individual, institutional, and national or disciplinary levels. By promoting the sharing of the results of scholarly inquiries into teaching, learning, and research supervision, SoTL helps to build a strong education community in the same way that sharing of the results of technical inquiries helps to build a strong research community. The IEEE Communications Society's Educational Services Board is committed to taking a leadership role in helping to advance the practice of SoTL and SoRS in communications education and training.

REFERENCES

- [1] M. Luiken, Ed., "Special Focus on the Future of Engineering and Technology Education," *IEEE Canadian Review*, Summer 2013, pp. 1–31.
- [2] M. S. Gupta, Ed., *Teaching Engineering: A Beginner's Guide*, IEEE Press, 1987.
- [3] R. M. Reis, *Tomorrow's Professor: Preparing for Careers in Science and Engineering*, Wiley-IEEE Press, 1997.
- [4] L. R. Lattuca, P. T. Terenzini, and J. F. Volkwein, *Engineering Change: A Study of the Impact of EC2000*, ABET, 2006.
- [5] E. Boyer, *Scholarship Reconsidered: Priorities of the Professoriate*, The Carnegie Foundation for the Advancement of Teaching, 1990.
- [6] W. E. Becker and M. L. Andrews, Eds., *The Scholarship of Teaching and Learning in Higher Education: Contributions of the Research Universities*, Indiana Univ. Press, 2004.
- [7] T. Hatch, *Into the Classroom: Developing the Scholarship of Teaching and Learning*, Jossey-Bass, 2006.
- [8] M. Weimer, *Enhancing Scholarly Work on Teaching and Learning: Professional Literature that Makes a Difference*, Jossey-Bass, 2006.
- [9] K. McKinney, *Enhancing Learning through the Scholarship of Teaching and Learning*, Anker Publishing, 2007.
- [10] P. Hutchings, M. T. Huber and A. Ciccone, *The Scholarship of Teaching and Learning Reconsidered: Institutional Integration and Impact*, Jossey-Bass, 2011.
- [11] G. D. Poole. "Using the Scholarship of Teaching and Learning at Disciplinary, National and Institutional Levels to Strategically Improve the Quality of Post-Secondary Education," *Int'l. J. Scholarship of Teaching and Learning*, vol. 1, no. 2, 2007, pp. 1–16.
- [12] M. Pearson and A. Bew, "Research Training and Supervision Development," *Studies in Higher Education*, vol. 27, no. 2, 2002, pp. 135–50.
- [13] K. Grant, R. Hackney, and D. Edgar, "Postgraduate Research Supervision: An 'Agreed' Conceptual View of Good Practice through Derived Metaphors," *Int'l. J. Doctoral Studies*, vol. 9, 2014, pp. 43–60.

BIOGRAPHY

DAVID G. MICHELSON (davem@ece.ubc.ca) leads the Radio Science Lab at the University of British Columbia where his research interests focus on wireless communications. His service to ComSoc has been recognized by ComSoc Chapter Awards, IEEE MGA's Outstanding Large Section award, and IEEE Canada's E. F. Glass Award. He completed the UBC Faculty Certificate Program on Teaching and Learning in Higher Education in 2011, served as ComSoc's Director of Education and Training from 2012 to 2013, and has served as a member of the ComSoc Educational Services Board from 2014 to present. He is also an elected member of the ComSoc Board of Governors (2013–2015 and 2017–2019).

While there is clear value in adopting SoTL within the engineering disciplines, either to inform the practice of teaching or to influence institutional or accreditation policy, efforts in this regard are still in their early stages compared to other academic disciplines.

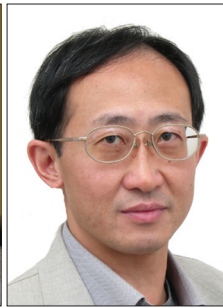
GREEN COMMUNICATIONS AND COMPUTING NETWORKS



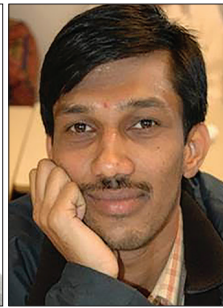
Jinsong Wu



John Thompson



Honggang Zhang



RangaRao Venkatesha Prasad



Song Guo

During the period from 23 May to 28 May 2016, the second United Nations Environment Assembly (UNEA-2) was held in Nairobi, Kenya, and was attended by over 2500 delegates from 174 countries, including 123 ministerial-level participants as well as 230 representatives of business and 400 representatives from accredited major groups and stakeholders [1]. With the theme of delivering on the environmental dimension of the 2030 Agenda for Sustainable Development, UNEA-2 concluded with a high level of agreement among environment ministers of the world on the need to strengthen the environmental dimension in the process of implementing the sustainable development goals (SDGs), although there were no concrete proposals agreed on how UNEP can ensure that the environmental dimension is included at the heart of all development policies. The first session of UNEA in June 2014 discussed major issues such as illegal trade in wildlife, air quality, environmental rule of law, financing the green economy, and the SDGs. The establishment of UNEA shows that the environment has recently moved from the margins to the center of the sustainable development agenda of the world. UNEA-2 adopted 24 resolutions addressing diverse topics, such as sustainable consumption and production, biodiversity, illegal wildlife trade, desertification, and marine issues [1].

The scope of the IEEE Technical Committee on Green Communications and Computing (TCGCC) has included environmental sustainability topics since it was the IEEE Technical Sub-Committee on Green Communications and Computing (TSCGCC) in 2011 [2]. In two recent papers published in May 2016, the relevance of environmental sustainability to information and communication technologies has been addressed specifically [3, 4]. This Green Series would like to call for more high-quality submissions addressing more general environmental issues. The fifth, November 2016, issue of the IEEE Series on Green Communications and Computing Networks includes nine articles relevant to green ICT.

The article “Green Touchable Nanorobotic Sensor Networks” explores a broader aspect of green communications for nano-robots. The article describes how these robots may

be useful in a wide range of applications, particularly for monitoring the human body and detecting illness.

The article “Simultaneous Information and Energy Flow for IoT Relay Systems with Crowd Harvesting” addresses energy-efficient sensor networks for the Internet of Things.

The article “LIFETEL: Managing the Energy-Lifetime Trade-off in Telecommunication Networks” discusses the important issue of device lifetime.

The article “Dynamic Energy Trading for Wireless Powered Communication Networks” introduces a dynamic energy trading mechanism to improve the energy supply reliability and performance of wireless powered communication networks.

The article “Power-Saving Methods for Internet of Things over Converged Fiber-Wireless Access Networks” takes advantage of converged fiber-wireless access networks to design a shared communication infrastructure for supporting both IoT applications and traditional services.

The article “Sustainability Information Model for Energy Efficiency Policies” models energy efficiency policies extending the Internet Engineering Task Force (IETF) Policy Core Information Model.

The article “Software Defined Networking, Caching, and Computing for Green Wireless Networks” develops a software-defined networking-caching-computing integrated architecture for next generation green wireless networks.

The following two articles were independently handled through the Open Call by the Associate Editors-in-Chief, Zoran Zvonar.

The article “Green Data Path for TCAM-Based Software-Defined Networks” proposes a green architecture for SDN networks using the dynamic voltage and frequency scaling (DVFS) technique. The DVFS-enabled controller and switch can generate the frequency and voltage configurations according to a certain energy efficiency policy for TCAM chips.

The article “Toward the Development of a Techno-Social Smart Grid” proposes a new techno-social framework for smart grids. Knowing that the social aspects, interactions, and behavior of people and their energy usage/demands are intertwined these days, new forms of collecting qualitative information is becoming important.

ACKNOWLEDGMENTS

We would like to acknowledge the great support from Osman S. Gebizlioglu, the current Editor-in-Chief, as well as Zoran Zvonar, the just retired Associate Editor-in-Chief, of *IEEE Communications Magazine*, Peggy Kang, the Managing Editor of IEEE Communications Society Magazines, Jennifer Porcello, Production Specialist, and Joseph Milizzo, Assistant Publisher, and the other IEEE Communications Society publication staff. We also highlight the great support of this Green Series from the members of the IEEE Technical Committee on Green Communications and Computing (TCGCC).

REFERENCES

- [1] The United Nations Environment Assembly (UNEA); <http://web.unep.org/unea>
- [2] IEEE TCGCC; <http://www.comsoc.org/committees/technical-committee/green-communications-computing>
- [3] J. Wu *et al.*, "Big Data Meet Green Challenges: Big Data Toward Green Applications," *IEEE Systems J.*, vol. 10, no. 3, Sept. 2016; first published in May 2016.
- [4] J. Wu *et al.*, "Big Data Meet Green Challenges: Greening Big Data," *IEEE Systems J.*, vol. 10, no. 3, Sept. 2016; first published in May 2016.

BIOGRAPHIES

JINSONG WU [SM] (wujs@ieee.org) is an associate professor in the Department of Electrical Engineering, Universidad de Chile, Santiago. He is the founder and founding Chair of the IEEE Technical Committee on Green Communications and Computing. He is an Editor of the *IEEE Journal on Selected Areas in Communications* Series on Green Communications and Networking. He was

the leading editor and co-author of the comprehensive book *Green Communications: Theoretical Fundamentals, Algorithms, and Applications* (CRC Press, 2012).

JOHN THOMPSON [SM] (john.thompson@ed.ac.uk) currently holds a Personal Chair in Signal Processing and Communications at the School of Engineering in the University of Edinburgh, United Kingdom. He was deputy academic coordinator for the recent Mobile Virtual Centre of Excellence Green Radio project and now leads the UK SERAN project which studies spectrum issues for 5G wireless. He also currently leads the European Marie Curie Training Network ADVANTAGE which trains 13 Ph.D. students in the area of smart grid technology. He was also a Distinguished Lecturer on green topics for ComSoc in 2014–2015.

HONGGANG ZHANG [SM] (honggangzhang@zju.edu.cn) is a full professor at Zhejiang University, China. He was the International Chair Professor of Excellence for Université Européenne de Bretagne and Supélec, France (2012–2014). He served as the Chair of the Technical Committee on Cognitive Networks (TCCN) of ComSoc during 2011–2012. He has been the Lead Guest Editor of *IEEE Communications Magazine* Feature Topics on Green Communications. He served as the General Co-Chair of IEEE GreenCom 2010 and the Co-Chair of IEEE Online GreenComm 2015. He is the co-editor/co-author of *Green Communications: Theoretical Fundamentals, Algorithms and Applications* (CRC Press).

RANGARAO VENKATESHA PRASAD [SM] (R.R.VenkateshaPrasad@tudelft.nl) received his Ph.D. from the Indian Institute of Science, Bangalore. During his Ph.D. research, a scalable VoIP conferencing platform was designed. Many new ideas including a conjecture were formulated and tested by developing an application suite based on the research findings. Part of the thesis led to a startup venture, Esquebe Communication Solutions. In 2005, he joined the Technical University of Delft (TUDelft). He has worked on personal networks (PNs), IoT, CPS, and energy harvesting networks. His work at TUDelft has resulted in 180+ publications. He is a Senior Member of ACM.

SONG GUO [SM] (song.guo@polyu.edu.hk) is a full professor at the Department of Computing, Hong Kong Polytechnic University. He has published over 300 papers in refereed journals/conferences and received multiple IEEE/ACM best paper awards. He is an Editor of *IEEE Transactions on Green Communications and Networking* and the Secretary of the IEEE Technical Subcommittee on Big Data. He is a Senior Member of the ACM and an IEEE Communications Society Distinguished Lecturer.

"Do not get obsolete like an old technology, keep innovating yourself."

– Sukant Ratnakar

IEEE COMSOC
TRAINING
www.comsoc.org/training

Green Touchable Nanorobotic Sensor Networks

Yifan Chen, Tadashi Nakano, Panagiotis Kosmas, Chau Yuen, Athanasios V. Vasilakos, and Muhamad Asvial

The authors define the in-messaging and out-messaging interfaces for nanorobots to interact with a macro-unit. They describe the propagation and transient characteristics of nanorobots based on the existing experimental results. They discuss the planning of nanorobot paths by taking into account the effectiveness of region-of-interest detection and the period of surveillance.

ABSTRACT

Recent advancements in biological nanomachines have motivated the research on nanorobotic sensor networks (NSNs), where the nanorobots are *green* (i.e., biocompatible and biodegradable) and *touchable* (i.e., externally controllable and continuously trackable). In the former aspect, NSNs will dissolve in an aqueous environment after finishing designated tasks and are harmless to the environment. In the latter aspect, NSNs employ cross-scale interfaces to interconnect the *in vivo* environment and its external environment. Specifically, the in-messaging and out-messaging interfaces for nanorobots to interact with a macro-unit are defined. The propagation and transient characteristics of nanorobots are described based on the existing experimental results. Furthermore, planning of nanorobot paths is discussed by taking into account the effectiveness of region-of-interest detection and the period of surveillance. Finally, a case study on how NSNs may be applied to microwave breast cancer detection is presented.

INTRODUCTION

BACKGROUND AND MOTIVATION

Recent progress in bio-nanomachines have motivated the research on bio-inspired, biocompatible, and biodegradable nanorobots such as flagellated magnetotactic bacteria (MTB) with nanometer-sized magnetosomes. These bacteria can be utilized as efficient carriers of nanoloads, and thus can serve as diagnostic and therapeutic agents for tumor targeting applications [1, 2]. For example, the experiments in [2] demonstrated MTB targeted in the interstitial region of a tumor from the blood vessels. Moreover, it has been shown that MTB can be maneuvered by a magnetic field generated in custom-made MRI systems [3, 4]. Other examples of bio-nanorobots include motor proteins reconstructed for transportation of molecules, cells genetically modified for active moving, and so on [3].

These emerging technologies have motivated the design of nanorobotic sensor networks (NSNs) for detection and examination of regions of interest (ROIs) in an *in vivo* environment such as the internal space of the human body [4]. An NSN comprises a swarm of nanorobots,

which are designed, engineered, and controlled to perform specific tasks including adjusting the direction of movement based on an external propulsion-and-steering gradient, releasing signaling particles to form a concentration gradient, binding to particular cell receptors on the ROIs, and so on. The presence of ROIs is assumed to be a potential threat to the *in vivo* environment (e.g., tissue malignancies). The primary concern of NSNs is therefore to coordinate the movement of nanorobots and plan their paths, detect ROIs and identify their locations, and examine the properties (e.g., size and shape) of ROIs.

Information exchange between an NSN and a macroscale monitoring device can be realized through cross-length-scale communication interfaces as presented in the current work. Furthermore, the degradability of nanorobots, which contributes to “green” systems by allowing for benign integration into life, is also discussed.

MAIN CONTRIBUTIONS

The main contributions of this work are as follows. First, due to size and power constraints, reliable direct communications between multiple nanorobots are difficult to achieve. Thus, one of the key strategies we propose here is to realize *touchable* NSNs by establishing interfaces to interconnect the small-scale aqueous environment and the large-scale non-aqueous environment. Such interfaces should allow an external macro-unit to control the timing of *in vivo* sensing processes taking place (i.e., start to release nanorobots) and the pathways of nanorobots, which expands the capability of NSNs. Subsequently, the cross-scale in-messaging interface (IMI) and out-messaging interface (OMI) for nanorobots to interact with a macro-unit are presented, which facilitate the control, tracking, and sensing operations. The term *touchable* means that the sensing process can be controlled and tracked. This is similar to controlling through simple or multi-touch gestures by touching the screen with a finger through a touchscreen, where the finger here refers to the external guiding field. The second important strategy we propose here is to realize *green* NSNs. This paradigm requires that each nanorobot dissolves into biofluids in the human body without creating harmful byproducts after carrying out the medical operations. Based

This work was supported by the National Natural Science Foundation of China (61550110244), the Guangdong Natural Science Funds (S2013050014223, 2016A030313640), the Shenzhen Development and Reform Commission Funds ([2015]863, [2015]1939), and the Shenzhen Science, Technology and Innovation Commission Funds (KQX2015033110182368).

Digital Object Identifier:
10.1109/MCOM.2016.1500626CM

Yifan Chen is with Southern University of Science and Technology; Tadashi Nakano is with Osaka University; Panagiotis Kosmas is with King's College London; Chau Yuen is with Singapore University of Technology and Design; Athanasios V. Vasilakos is with Luleå University of Technology; Muhamad Asvial is with the University of Indonesia.

on this strategy, empirical models of nanorobot bioresorbability based on experimental findings reported in [1, 2] are presented. Furthermore, planning of nanorobot paths is discussed by considering the effectiveness of ROI detection and the period of surveillance. The path optimization process contributes to green NSNs by reducing the time nanorobots are present in the human body.

It is worth noting that classical green networks are mostly focused on energy-relevant green issues. The current work, on the other hand, extends this conventional concept to include reduction of the use and generation of toxic substances. Green NSNs may also find important applications in environmental protection and monitoring, where system degradability allows for benign integration into the environment.

Green touchable NSNs have the potential to alleviate some of the limitations in the existing nanoscale communication paradigms [5, 6]. For example, nanosensors cannot store and process a large amount of data, change signaling particles at will, or control the particle release and reception processes accurately for the message encoding and decoding. Furthermore, there is a great deal of uncertainty during the propagation process caused by random Brownian motions, temperature variation, chemical reactions, molecule decomposition, and so on. The proposed system, on the other hand, only requires simple functionalities including sensing and manoeuvring at nanorobots, and moves most operations to a macro-unit. Moreover, remote controllability and trackability of nanorobots reduce propagation delay and eliminate environmental uncertainty.

The remainder of the article is organized as follows. First, we give an architecture of touchable NSNs. Next, we describe the propagation and transient (i.e., “green”) characteristics of nanorobots, while we present the path planning criteria. We then illustrate the approach with an example of NSN for microwave breast cancer detection. Finally, some conclusions are drawn.

SYSTEM ARCHITECTURE OF TOUCHABLE NSNs

An architecture of touchable NSNs consists of remotely controllable and trackable nanorobots in an *in vivo* aqueous environment (e.g., blood vessels in the human body), and a macro-unit located in an external environment (e.g., a hybrid operating room equipped with medical imaging devices). A macro-unit implements two types of interfaces: IMI and OMI (I/OMI), as shown in Fig. 1. It uses IMIs to transmit in-messages to nanorobots and OMIs to receive out-messages from nanorobots. Note that the communication interface for direct interaction between nanorobots is not implemented due to the limited size and computational power of nanoscale entities. The key components of touchable NSNs are described as follows.

MACRO-UNIT AND NANOROBOT

A macro-unit is a large-scale device that relies on conventional means of controlling and imaging, which may be composed of materials incompatible with the *in vivo* environment and may be larger than nanorobots by orders of magnitude. A macro-unit can generate in-messages

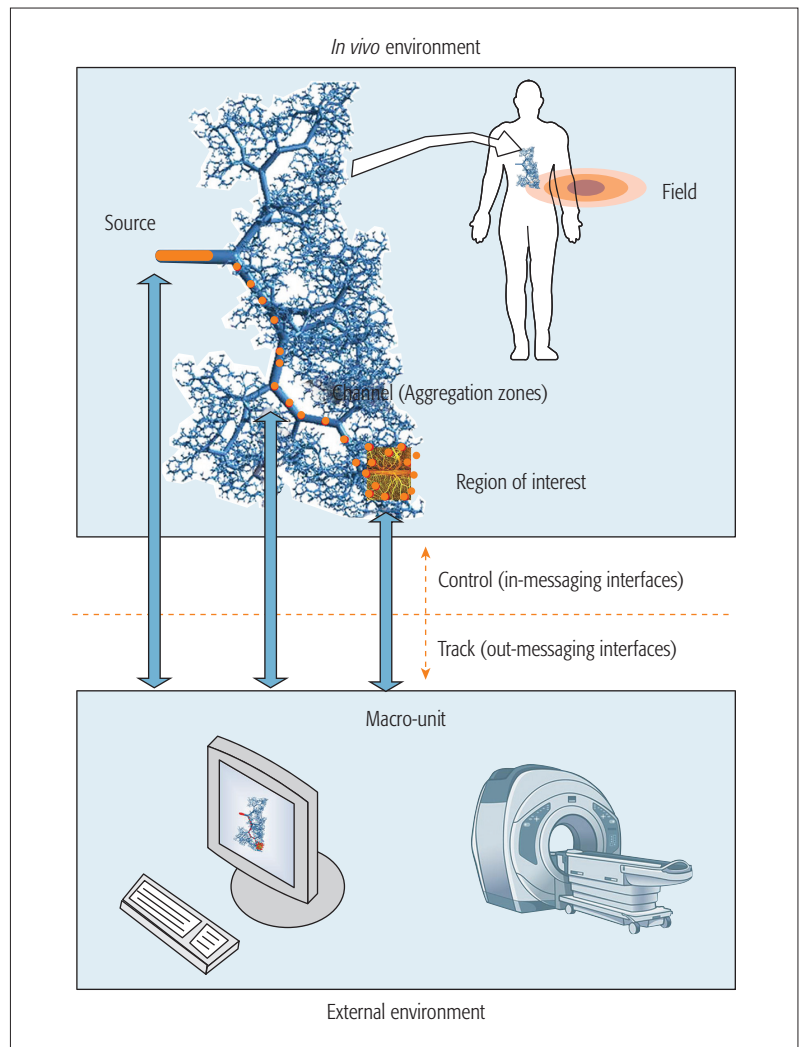


Figure 1. An architecture of touchable nanorobotic sensor networks with in-messaging interfaces and out-messaging interfaces.

to and receive out-messages from nanorobots that reside in the *in vivo* environment. It consists of two main modules: the tracking module uses a medical imaging platform such as an optical microscope, a fluoroscopic system, an MRI machine, or a microwave imaging device [1, 2, 7] to monitor the movement and aggregation of nanorobots; the controlling module uses a traditional electronic, magnetic, or optical system to release, propel, or steer nanorobots.

A nanorobot is composed of biocompatible and biodegradable materials (e.g., MTB [1, 2]), and has a size ranging from the size of a macromolecule to that of a biological cell [3, 8]. A nanorobot implements the following basic functionalities [8]: *steering wheel* and *propeller*, controlling the swimming direction and speed of the nanorobot, respectively; a *fuel* unit harvesting energy from the surroundings to provide power to the nanorobot; a *sensor* unit acting as the interface between the aqueous medium and the nanorobot; a *nanoload* made of nanocomposites such as diagnostic agents and therapeutic drugs attached to the nanorobot; and a *navigator* tracking the movement of the nanorobot. Examples of nanorobots include unicellular organisms, genetically modified cells, and biological cells

Propagating molecules themselves are not nanorobots. However, similar fluorescent technique can be applied to nanorobots for tracking purposes. In addition, differential microwave imaging can be employed for tracking of nanorobots, where the nanoload attached to nanorobots is a contrast agent such as carbon nanotubes.

[1–3], which are able to swim in the *in vivo* environment under the maneuvering of an external field, carrying a cargo such as nanoparticles, acting as biosensors detecting specific molecules, and being integrated with cell-native components (e.g., magnetosomes in MTB) or coated with artificial materials (e.g., fluorescent molecules) for controlling and tracking purposes.

IMIS AND OMIS

A macro-unit implements cross-scale IMIs and OMIs to interact with nanorobots as shown in Fig. 1. The IMIs must convert conventional electronic, magnetic, or optical signals used by the large-scale device into commands to which nanorobots respond by performing subsequent small-scale operations, while the OMIs must convert motion signals generated by nanorobots (e.g., releasing, swimming, and targeting of nanorobots) to externally detectable and interpretable messages. Examples of IMIs and OMIs are as follows (Fig. 1).

The source IMI controls the release of nanorobots at a specified location and time from the nanorobot source [1–3], which initializes the sensing process. For example, a laser emits light at a specific frequency, which may cause caged compounds to release encapsulated molecules through bond breaking [3]. An MRI device generates a magnetic gradient, which may maneuver a swarm of MTB confined within an aggregation zone (AZ) toward a targeted destination by magnetotaxis [1, 2]. The source OMI tracks the release of nanorobots. For example, fluorophores such as derivatives of rhodamine attached to nanorobots may emit light at a specific frequency in response to excitation from an external macro-unit, and the macro-unit may detect emitted fluorescence [3]; an X-ray scanner may provide an angiogram showing the blood vessels for monitoring of catheter placements where a catheter is used to release nanorobots at a site near the target [1, 2].

The channel IMI controls the movement of nanorobots in the *in vivo* environment. First, an external macro-unit creates a propelling-and-steering field in the surveillance area as illustrated in Fig. 1. For example, in [1, 2], the patient was positioned inside the bore of an MRI system where MTB were released from the tip of a catheter. The maneuver of MTB required three-dimensional steering magnetic coils, which induced a torque on the chain of magnetosomes in MTB. In [1, 2], an agglomeration of MTB were controlled to move along pre-designed microchannels mimicking the human microvasculature. The channel OMI tracks the path of nanorobots through various imaging modalities such as MRI (for MTB propagating in microvasculature) [1, 2] and fluorescence microscopy (for molecules propagating through gap junction channels) [3]. It is worth noting that propagating molecules themselves are not nanorobots. However, similar fluorescent technique can be applied to nanorobots for tracking purposes. In addition, differential microwave imaging can be employed for tracking of nanorobots, where the nanoload attached to nanorobots is a contrast agent such as carbon nanotubes [7, 9].

The ROI IMI controls the targeting, absorp-

tion, and dissolution of nanorobots at the ROI, while the ROI OMI tracks these processes occurring at the ROI, which also indicates the presence of ROI. First, the nanorobots are not located until they reach an intermediate AZ. This is to avoid interference caused by electromagnetic fields generated during the propelling-and-steering phase and the tracking phase, as well as to reduce the complexity of the macro-unit. Furthermore, nanorobots in a swarm swim at different velocities during traveling to an AZ and as such they disperse, making the density too low for tracking (see, e.g., [1, 2]). The distribution and number of AZs are dependent on the performance requirements of the ROI sensing. Denser placement of AZs gives rise to more effective ROI detection and more accurate localization at the cost of a longer surveillance period. Consequently, the measured nanorobot footprints corresponding to these AZs result in snapshots of the actual nanorobot paths. If a swarm of nanorobots fail to detect an ROI (i.e., the corresponding path does not cross the ROI), they will keep navigating until being maneuvered out of the surveillance area or dissolving in the medium. On the other hand, if these nanorobots reach an ROI, the nanoload will be removed and assimilated into the ROI, allowing the nanoload to carry out medical tasks. The attachment of the load to the ROI turns off the navigator module in the nanorobots. This may be realized by load discharging where the load itself also serves as the navigator (e.g., only fluorescent molecules are appended to the load). Consequently, the nanorobots will become invisible to the macro-unit. The final location of the nanoload will result in a “sink” on the preplanned nanorobot pathway, as shown in Fig. 1.

Nanorobots can also be administered to target a detected and localized ROI from various directions through the ROI IMI [7]. In this case, the size and shape of the ROI can be estimated via registration of all the final nanorobot footprints, as also shown in Fig. 1. This additional information will help reduce false alarms for ROI inspection. The information about the properties of an ROI is thus achieved by requiring the macro-unit to observe the footprints of nanorobots. This *seeing-is-sensing* strategy forms the ROI OMI of the touchable NSN. It is worth noting that the sensing operation can be combined with targeted transportation of nanocomposites to the ROI, while preventing them from affecting other sites within the surveillance area. An important application is targeted drug delivery, which enhances locoregional therapies for cancer treatment without causing toxic effects to non-targeted sites in the human body.

PROPAGATION CHARACTERISTICS AND TRANSIENT KINETICS OF NANOROBOTS FOR GREEN NSNS

PROPAGATION MODELS OF NANOROBOTS

Velocity-Jump Random Walk Model: Nanorobots under the guidance of externally exerted magnetic fields result in trajectories similar to random walks as demonstrated in [10]. Various random walk models for different molecular

propagation scenarios have been discussed in [11]. In [12], a velocity-jump random walk process was employed for simulating the nanorobot movement in a liquid along the propelling field lines. According to this model, nanorobots move forward with a time-varying velocity for a random step, which follows a Poisson process of turning frequency λ . To model changes in direction, a von Mises distribution of mean θ and concentration κ is employed. The mean of this distribution depends on the propelling field at the present location and thus introduces a directional bias. The width of the angular spread increases as κ reduces. The two parameters λ and κ characterize the effects of microenvironmental mechanisms (e.g., chemotaxis and aerotaxis) and the propagation environment. A large λ indicates a small mean run length time, which is applicable to a homogeneous fluid medium that lacks a flux structure, imposing constraints on the feasible nanorobot paths. In this case, $1/\lambda$ is mainly controlled by the curvature of the external field line. On the other hand, a small λ corresponds to a structured fluid network in which nanorobots will travel through the fluid vessels and will not change their direction frequently. In this scenario, $1/\lambda$ is related to the average length of each straight section of the flow network, such as a branch in the vascular tree. Furthermore, a large κ indicates that the propelling force predominates over other microenvironmental gradients. In general, a flux structure also leads to a small κ .

Fractal-Based Statistical Model: The pathway of nanorobots may also be described using the fractal-based statistical model [8], which characterizes the movement of nanorobots in a vascular tree when the capillaries can be imaged. First, the vascular network distributes the blood stream through the human body by successive bifurcations, which can be described by a fractal model due to the self-similar character of a bifurcating tree. Second, the hemodynamic response leads to increased cross-sectional areas of blood flow, which results in blood velocity low enough for the exchange of metabolic substances across the capillary wall. Hemodynamics explains the empirical Murray's laws that characterize the connection between the diameter of the parent segment and the diameters of two daughter branches at each bifurcating node [8]. The angle between two daughter vessels after division and the lengths of the vessels are also determined by Murray's formula. Finally, when the microvasculature becomes angiographically unresolvable, it may be simplified as a structureless fluid medium, and the velocity-jump random walk mentioned above may be used to describe the pathway of nanorobots as suggested in [8].

Deterministic Model: The pathway of nanorobots may also be described using the deterministic topology of large arteries collected from anatomical data [13]. The model in [13] was divided into two parts: the cardiovascular network model and the drug propagation network model. The former was obtained by solving the Navier-Stokes equation, which computed the blood velocity at an arbitrary location through the cardiovascular network given the blood pressure. In this case, the transmission line analysis

was applied to characterize the arterial interconnection. The latter was obtained by solving the advection-diffusion equation, which computed the concentration of drug particles at an arbitrary location through the cardiovascular network given the blood velocity. In this case, the harmonic transfer matrix analysis was applied to characterize the transfer function of each blood vessel and dichotomous division.

TRANSIENT KINETICS OF NANOROBOTS FOR GREEN NSNs

The degeneration of bacterial nanorobots (e.g., MTB) results in decrease in their terminal velocity. In [1, 2], Martel *et al.* measured the velocity of a large agglomeration of MTB by using a video camera mounted on an optical microscope. The velocity exhibited two-stage kinetics with respect to the immersion time for 37°C blood temperature. If the threshold speed below which a bacterium is considered degenerating is 100 $\mu\text{m/s}$, MTB would function stably for a duration of 15 min. Over the following 25 min, MTB started to deteriorate with gradual reduction in their speed. Overall, MTB remained active for 40 min and then slowed down due to high blood temperature. Moreover, it has been demonstrated that a swarm of MTB could be controlled by an external computer like a single organism to swim along a predetermined path [1, 2]. The amount of MTB in each swarm is usually large, and therefore, it is useful to look into their concentration at various locations of blood vessels. As such, the transient behavior of MTB is also related to diffusion of bacteria in the medium. The diffusion effect has been commonly modeled using Fick's laws for space of various dimensions [8, 11]. Then the overall transient characteristic of MTB (ranging between 0 and 1 for no resorption and complete resorption, respectively) is given by the percentage of diffused MTB on arrival at the ROI if the total immersion time is less than the lifespan of MTB. Otherwise, the MTB are fully degenerated.

NANOROBOT PATH PLANNING CRITERIA

Two criteria for nanorobot path planning are considered in this section, which are based on the effectiveness of ROI detection and the period of surveillance.

Consider a number of *sequential* and *directional* nanorobot routes in the *in vivo* environment to cover the entire surveillance area. Each route utilizes one set of nanorobots to cut through the surveillance area. If an ROI is identified, and assuming that its spatial distribution is known a priori (e.g., via a lower-resolution tomographic image), the probability of the ROI being present within the area traveled by nanorobots between any two adjacent AZs, A_1 , and A_2 , can be estimated. The ROI detection between these two AZs aims to distinguish between the null hypothesis \mathcal{H}_0 ("ROI is absent") and the alternative hypothesis \mathcal{H}_1 ("ROI is present"). If \mathcal{H}_1 is true, a portion of the fluorescent nanoload will be successfully discharged, and the rest will be transported to A_2 . Subsequently, the indicator quantity signifying presence or absence of nanoload at A_2 is the sum of the transient characteristic at A_2 without encountering any ROI, and the increase in transient characteristic due to nanoload unloading at the ROI between A_1 and A_2 .

The sensing operation can be combined with targeted transportation of nanocomposites to the ROI, while preventing them from affecting other sites within the surveillance area. An important application is targeted drug delivery, which enhances locoregional therapies for cancer treatment without causing toxic effects to non-targeted sites in the human body.

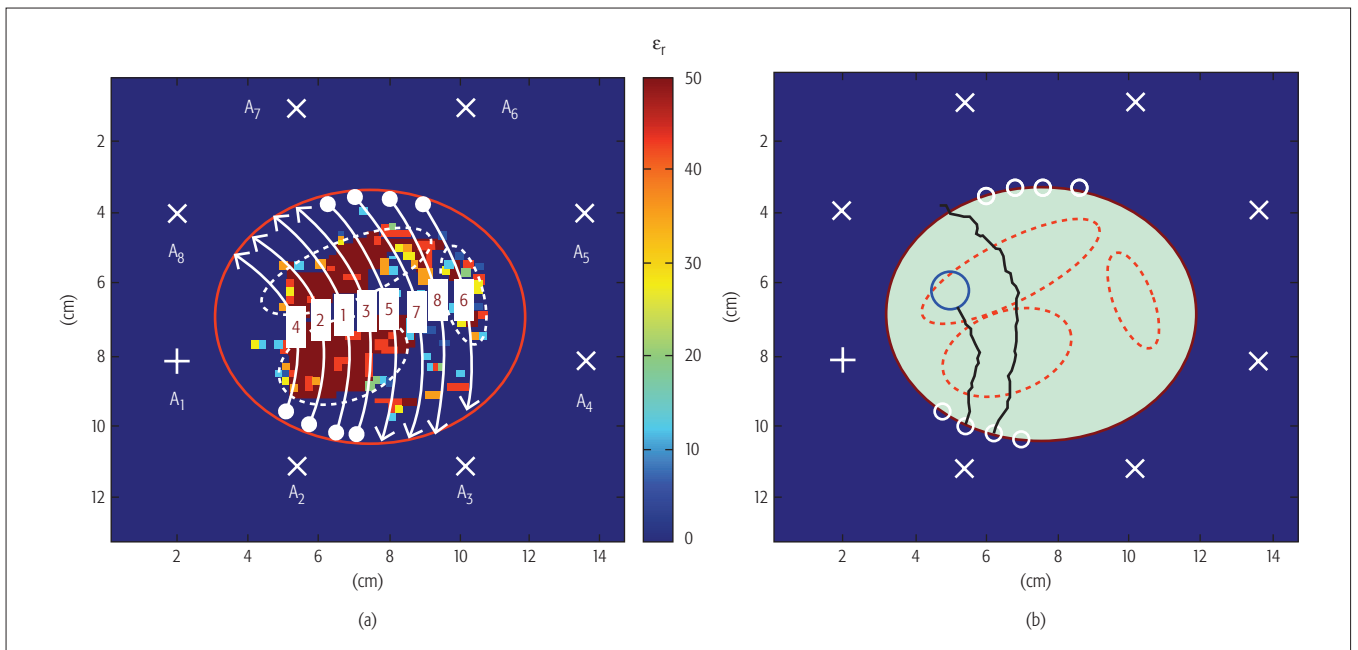


Figure 2. a) Dielectric profile of the breast cross-section at 1 GHz. A cancer is assumed to be uniformly distributed within the regions encircled by the three dotted ellipses. A dc source (marked with “+”) generates concentric magnetic fields. The nanorobot survey routes (marked by white curved arrows) cover the surveillance area, which are designed by applying the path planning principles stated earlier. The arrow and number indicate the direction and sequence of the routes, respectively; b) simulated tumor location (marked with blue circle) and nanorobot trajectories (marked with thick black curves) following the optimized survey plan in a).

Then the probability of ROI detection is given by the probability that the indicator is no less than a pre-specified threshold beyond which nanorobot tracking would fail. If there is no ROI between A_1 and A_2 , the indicating metric is given by the transient characteristic at A_2 without encountering any ROI. The probability of false alarm is obtained as the probability that the indicator is no less than the same threshold. Subsequently, the overall probabilities of detection and false alarm can be obtained. The Youden’s index, which is the difference between the probabilities of detection and false alarm, can thus be calculated, and the corresponding path planning criterion is to identify the optimal route such that the Youden’s index is maximized. In general, optimal path planning rules should ensure that AZs where the ROI detection performance is poorer (i.e., smaller Youden’s indices) should also correspond to areas with lower probabilities of ROI presence. Furthermore, to achieve more effective ROI detection, the direction of nanorobot movement for each path should be designed such that the location-variant nanorobot transient characteristics “match” the spatial distribution of ROI in the surveillance area (i.e., the denser the distribution of the ROI, the smaller the transient characteristic of nanorobots). In other words, the “higher-risk” areas should be surveyed when the nanorobots deployed are under more reliable operating conditions.

The second criterion for optimization of nanorobot paths is based on the time to ROI detection. Our goal is to identify the desirable set of paths resulting in the maximum cumulative distribution function of time to detection across all times. There are two general strategies to obtain the solution. First, if the spatial distribution of

the ROI in the interior of the surveillance area is available, the preplanned nanorobot routes should “match” the probability map, such that the AZ begins with the location of the highest probability of ROI detection and shifts successively to locations having descending detection probabilities. This protocol is consistent with the strategy suggested previously because the transient characteristic is a non-decreasing function with the operation time. Second, the velocity should also “match” the probability map, where the nanorobots navigate across the areas having higher probabilities of ROI presence with higher speeds. Similarly, this condition agrees with the other strategies mentioned above because the velocity is a non-increasing function with the operation time [1, 2].

It is worth emphasizing that the path planning strategies also contribute to green NSNs by reducing the duration of the nanorobots’ presence in the *in vivo* environment.

CASE STUDY: NSN FOR MICROWAVE BREAST CANCER DETECTION

Microwave medical imaging for early-stage breast cancer diagnosis attempts to discriminate breast tissues based on their dielectric properties [7, 9]. However, its effectiveness is compromised by clutter interference due to healthy tissue inhomogeneities. This problem can be solved by using a contrast agent to change the tumor tissue dielectric values. The agent is transported selectively to cancer cells via systemic administration for specific and non-invasive diagnosis of tissue malignancies. However, the existing targeting techniques have limited efficacy as a result of intratumoral penetration limit. Consequently,

only a small amount of a contrast agent is able to enter cancer cells.

NSN can provide a potential remedy for the aforementioned problem. For example, suppose that MTB are used to effectively deliver a contrast agent (i.e., nanoload) to a tumoral region (i.e., ROI) in the human breast. Each time an agglomeration of MTB loaded with a contrast agent are injected into the breast (i.e., surveillance area) from a predefined injection site, the MTB will navigate in the direction of an externally exerted magnetic field. Their movement can be tracked by using a differential microwave imaging system (i.e., macro-unit) [7, 9]. When a particular swarm trajectory meets a tumor, the contrast agent is discharged from the nanorobots and attached to cancer cell receptors (i.e., nanoload discharging). Subsequently, the unloaded nanorobots may no longer be trackable by the differential imaging system, which only detects difference in the tissue dielectric properties caused by the agent. Therefore, a nanorobot footprint “sink” inside the breast would emerge at the tumor location where the contrast agent eventually accumulates (i.e., seeing-is-sensing).

To elaborate on the approach, consider a “heterogeneously dense” numerical phantom in the breast model repository in [14], which also includes a tumor having the dielectric values of cancerous breast tissue. Similar to [7, 9], an antenna array consisting of 40 dipole elements arranged over 5 circular rings is employed as the macro-unit. Two-dimensional tumor sensing along the cross-sectional planes of the rings is implemented using the differential microwave imaging data. For illustrative purposes, consider the cross-sectional plane slicing through the tumor. The corresponding dielectric profile is shown in Fig. 2a. It is supposed that the presence of a contrast agent in a tissue region will substantially change its dielectric values [9]. Simulation studies on the movement, tracking, and location estimation of MTB have been presented in [7].

Assuming that the distribution of a cancer is given by the profile indicated in Fig. 2a (i.e., uniformly distributed within the high-dielectric-property areas encircled by the three dotted ellipses), the optimal MTB routes following the principles stated earlier are depicted in Fig. 2a and explained in the figure caption. Note that the sequence and direction of the MTB paths are determined by the criteria based on the detection effectiveness and surveillance period. More specifically, *the higher the distribution of ROI, the larger the Youden’s index, the smaller the MTB transient characteristic, and the higher the MTB speed*. Figure 2b shows a typical trajectory following the optimized survey route in Fig. 2a. The velocity-jump random walk model is considered. An MTB footprint sink indicating the tumor location is registered. Figure 3 shows the probability distributions of the tumor sensing time for both the optimized and non-optimized scenarios when the number of simulation runs is 1000. In the latter case, the same survey paths are used but with a non-selective sequence and direction of traveling (i.e., left-to-right, up-to-down). As can be seen from the figure, in general

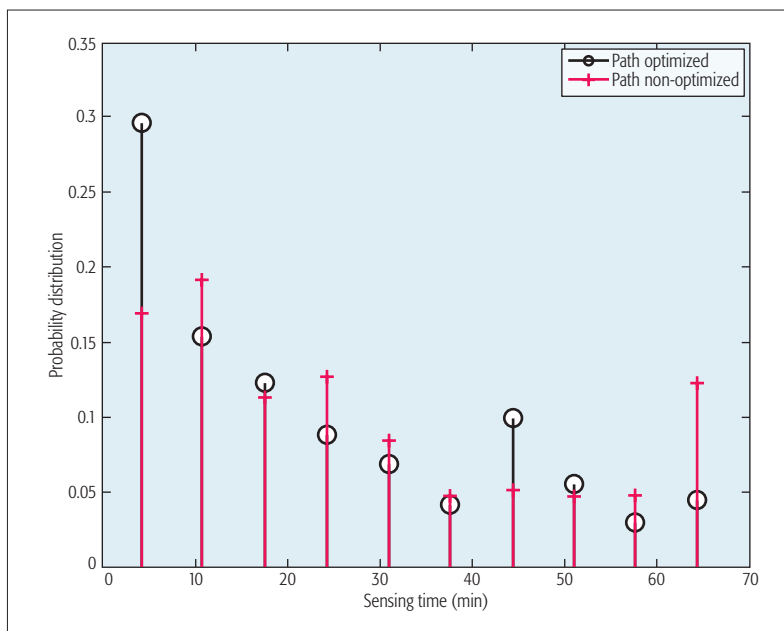


Figure 3. The probability distributions of the sensing periods for the optimized and non-optimized nanorobot routes.

the probability distribution when the MTB path is optimized decreases with the sensing time, whereas the distribution for the non-optimized situation is more uniformly distributed over the entire range of surveillance period. With path optimization, the mean and median values have been reduced from 27 min to 24 min and from 23 min to 20 min, respectively.

CONCLUSIONS

A novel system of green touchable NSNs has been introduced in this article. The system makes use of cross-scale IMIs and OMIs to facilitate the control, sensing, and transportation operations at the bottom. Nanorobots would disappear in the fluid medium after completing designated tasks, which eliminates the costs and environmental/health risks incurred by the retrieval of waste materials. The fundamental principles, propagation and degradation models, and design case study of the proposed system have been presented. Both environmental friendliness and path planning strategies help to realize green NSNs. The analytical framework presented here would pave the way for system-level implementations and experimental studies of NSNs in future.

REFERENCES

- [1] S. Martel *et al.*, “Flagellated Magnetotactic Bacteria as Controlled MRI-Trackable Propulsion and Steering Systems for Medical Nanorobots Operating in the Human Microvasculature,” *Int’l. J. Rob. Res.*, vol. 28, Apr. 2009, pp. 571–82.
- [2] S. Martel *et al.*, “MRI-Based Medical Nanorobotic Platform for the Control of Magnetic Nanoparticles and Flagellated Bacteria for Target Interventions in Human Capillaries,” *Int’l. J. Rob. Res.*, vol. 28, Sept. 2009, pp. 1169–82.
- [3] T. Nakano *et al.*, “Externally Controllable Molecular Communication,” *IEEE JSAC*, vol. 32, no. 12, Dec. 2014, pp. 2417–31.
- [4] Y. Okaie *et al.*, “Autonomous Mobile Bionanosensor Networks for Target Tracking: A Two-Dimensional Model,” *Nano Commun. Net.*, vol. 5, 2014, pp. 63–71.
- [5] S. F. Bush, *Nanoscale Communication Networks*, Artech House, 2010.
- [6] I. F. Akyildiz, J. M. Jornet, and M. Pierobon, “Nanonetworks: A New Frontier in Communications,” *Commun. ACM*, vol. 54, Nov. 2011, pp. 84–89.

- [7] Y. Chen, P. Kosmas, and S. Martel, "A Feasibility Study for Microwave Breast Cancer Detection Using Contrast-Agent-Loaded Bacterial Microbots," *Int'l. J. Antennas Propag.*, vol. 2013, article ID 309703, 11 pages.
- [8] Y. Chen *et al.*, "A Touch-Communication Framework for Drug Delivery Based on a Transient Microbot System," *IEEE Trans. Nanobiosci.*, vol. 14, no. 4, June 2015, pp. 397–408.
- [9] Y. Chen and P. Kosmas, "Detection and Localization of Tissue Malignancy Using Contrast-Enhanced Microwave Imaging: Exploring Information Theoretic Criteria," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 3, Mar. 2012, pp. 766–76.
- [10] I. S. M. Khalil *et al.*, "Closed-Loop Control of Magnetotactic Bacteria," *Int'l. J. Rob. Res.*, vol. 32, 2013, pp. 636–48.
- [11] T. Nakano *et al.*, "Molecular Communication and Networking: Opportunities and Challenges," *IEEE Trans. Nanobiosci.*, vol. 11, no. 2, June 2012, pp. 135–48.
- [12] Y. Chen, P. Kosmas, and R. Wang, "Conceptual Design and Simulations of a Nano-Communication Model for Drug Delivery Based on a Transient Microbot System," *Proc. 2014 8th Euro. Conf. Antennas and Propag.*, The Hague, The Netherlands, Apr. 2014, pp. 63–67.
- [13] Y. Chahibi *et al.*, "A Molecular Communication System Model for Particulate Drug Delivery Systems," *IEEE Trans. Nanobiosci.*, vol. 60, no. 12, Dec. 2013, pp. 3468–83.
- [14] E. Zastrow *et al.*, "Development of Anatomically Realistic Numerical Breast Phantoms with Accurate Dielectric Properties for Modeling Microwave Interactions with the Human Breast," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 12, Dec. 2008, pp. 2792–2800.

BIOGRAPHIES

YIFAN CHEN (chen.yf@sustc.edu.cn) is presently a professor and head of the Department of Electrical and Electronic Engineering with Southern University of Science and Technology, Shenzhen, China. His current research interests include transient communications, small-scale and cross-scale communications and sensing, microwave medical imaging and diagnosis, and propagation channel modeling. He is a key contributor to IEEE Standard 1906.1 on Nanoscale and Molecular Communication Framework, and an Editor of *IEEE ComSoc Best Readings in Nanoscale Communication Networks*.

TADASHI NAKANO (tadasi.nakano@fbs.osaka-u.ac.jp) is an associate adjunct professor of the Institute of Academic Initiatives, Osaka University, Japan, and

a guest associate professor of the Graduate School of Biological Sciences, Osaka University. His research interests are in the areas of network applications and distributed computing systems with strong emphasis on interdisciplinary approaches. His current research is focused on biological ICT including design, implementation, and evaluation of biologically inspired systems and synthetic biological systems.

PANAGIOTIS KOSMAS (panagiotis.kosmas@kcl.ac.uk) is currently a senior lecturer at King's College London, Department of Informatics. He is also a co-founder of Mediwise Ltd, an award-winning U.K.-based SME focusing on the use of electromagnetic waves for medical applications. His research interests include bio-electromagnetics with application to wave propagation, sensing and imaging, antenna design, physics-based detection methods, and inverse problem theory and techniques.

CHAU YUEN (yuenchau@sutd.edu.sg) was a postdoctoral fellow at Lucent Technologies Bell Labs during 2005. During the period of 2006–2010, he was at the Institute for Infocomm Research as a senior research engineer. He joined Singapore University of Technology and Design as an assistant professor in June 2010. He also serves as an Associate Editor for *IEEE Transactions on Vehicular Technology*. In 2012, he received the IEEE Asia-Pacific Outstanding Young Researcher Award.

ATHANASIOS V. VASILAKOS (vasilako@ath.forthnet.gr) is currently a professor with the Lulea University of Technology, Sweden. He has served or is serving as an Editor for many technical journals, such as *IEEE Transactions on Network and Service Management*, *IEEE Transactions on Cloud Computing*, *IEEE Transactions on Information Forensics and Security*, *IEEE Transactions on Cybernetics*, *IEEE Transactions on NanoBioscience*, and the *IEEE Journal on Selected Areas in Communications*. He is also General Chair of the European Alliances for Innovation.

MUHAMAD ASVIAL [M] (asvial@ee.ui.ac.id) is currently a researcher and lecturer in the Electrical Engineering Department, Universitas Indonesia. His research interests include mobile communication (terrestrial and satellite communication), HAP networks, genetic algorithm applications, and broadband and ultra wideband communication systems. He has published over 90 papers in several international journals and conferences. He is a member of the IET and the AIAA.

Simultaneous Information and Energy Flow for IoT Relay Systems with Crowd Harvesting

Weisi Guo, Sheng Zhou, Yunfei Chen, Siyi Wang, Xiaoli Chu, and Zhisheng Niu

ABSTRACT

It is expected that the number of wireless devices will grow rapidly over the next few years due to the growing proliferation of the Internet of Things. In order to improve the energy efficiency of information transfer between small devices, we review state-of-the-art research in simultaneous wireless energy and information transfer, especially for relay-based IoT systems. In particular, we analyze simultaneous information and energy transfer from the source node, and the design of time-switching and power-splitting operation modes, as well as the associated optimization algorithms. We also investigate the potential of crowd energy harvesting from transmission nodes that belong to multiple radio networks. The combination of source and crowd energy harvesting can greatly reduce the use of battery power and increase the availability and reliability for relaying. We provide insight into the fundamental limits of crowd energy harvesting reliability based on a case study using real city data. Furthermore, we examine the optimization of transmissions in crowd harvesting, especially with the use of node collaboration while guaranteeing quality of service.

INTRODUCTION

Today, wireless services are dominated by packet data transfer over the cellular and Wi-Fi networks. Cellular networks account for the majority of the world's wireless power consumption, with 6 million macrocells worldwide consuming a peak rate of 12 billion W. While video demand drives most of the data consumption, machine-to-machine (M2M) data is the fastest growing driver. The rapid growth in the Internet of Things (IoT) sector is set to increase the energy consumption of small devices dramatically. While many such small devices are sensor nodes with power consumption that is in the order of 1 W or less, the sheer number of such devices (25 billion) is set to consume more power than cellular networks worldwide. Therefore, it is important to address the emerging challenge of energy efficiency for connected small devices. In order for small devices to communicate without a tether, wireless relaying has

been widely employed in current and emerging wireless systems. For example, M2M-R relaying has been proposed as a suitable heterogeneous architecture for either the European Telecommunications Standards Institute (ETSI) M2M, or Third Generation Partnership Project (3GPP) machine type communications (MTC) or 802.16p IoT architectures [1].

Conventionally, relaying operations cost the relaying nodes extra energy and therefore may prevent battery-operated nodes from taking part in relaying. Thus, RF powered relaying is a promising solution, whereby the relay nodes can harvest energy from either the source node directly [2] or external sources [3] (i.e., *crowd harvesting* from external radio transmissions) for sustainable operation, as shown in Fig. 1. In fact, the feasibility of crowd harvesting is supported by the dramatic increase in the density of RF transmitters in cities. For example, the global cellular infrastructure consists of more than 6 million base stations (BSs) and serves more than 7 billion user equipments (UEs), and the number of Wi-Fi access points (APs) has reached 350/km², with many metropolitan areas reaching over 700 per km².¹

This article addresses several challenges faced by energy harvesting relay systems attempting to optimize throughput. We break down the problem into two domains:

- The source of the RF energy
- Optimizing the transfer of information subject to different energy harvesting scenarios

In particular, we answer two important research questions: how to optimally schedule information and energy transmission from a common transmitter node, and how much energy can be crowd-harvested from ambient RF signals for relaying.

The organization of the article is as follows. We first review state-of-the-art research on RF powered relaying systems. We also review the potential of harvesting a crowd of transmission nodes that belong to multiple heterogeneous networks. We then discuss the optimization of transmissions with crowd harvesting, especially with node collaboration, while guaranteeing a certain QoS. Finally, we discuss open challenges and research opportunities in this area.

The authors review state-of-the-art research in simultaneous wireless energy and information transfer, especially for relay-based IoT systems. In particular, they analyze simultaneous information and energy transfer from the source node, and the design of time-switching and power-splitting operation modes, as well as the associated optimization algorithms.

This work is sponsored in part by the Nature Science Foundation of China under the Grant No.61571265, 61461136004 and 61321061, the Young Scholar Programme of the Jiangsu Science and Technology Programme under the Grant No.BK20150374.

¹ <http://www.smallcellforum.org/press-releases/small-cells-outnumber-traditional-mobile-base-stations>

Weisi Guo and Yunfei Chen are with the University of Warwick; Sheng Zhou and Zhisheng Niu are with Tsinghua University; Siyi Wang is with Xi'an Jiaotong-Liverpool University; Xiaoli Chu is with the University of Sheffield.

Digital Object Identifier: 10.1109/MCOM.2016.1500649CM

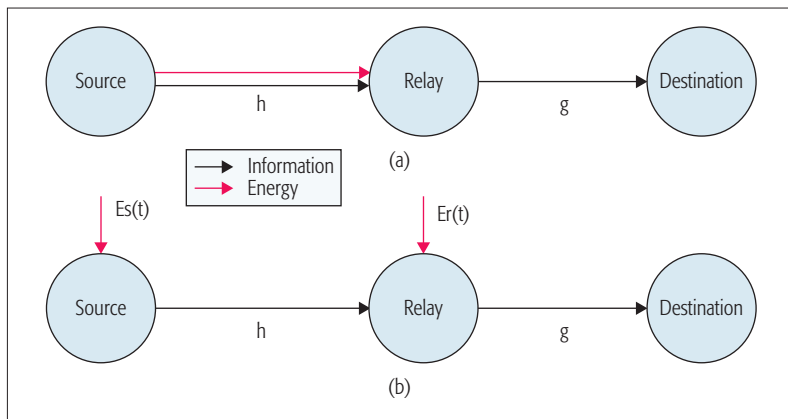


Figure 1. Diagram of wireless powered relaying harvesting the source node or harvesting external sources.

RF POWERED RELAYING

SWIPT: HARVESTING FROM THE SOURCE NODE AND RELAYING

If the relay node harvests energy from the source node's RF transmissions, the relaying becomes a simultaneous wireless information and power transfer (SWIPT) system [2], as the relay node also receives information from the source node. In this case, there are two main algorithms to implement wireless powered relaying (WPR): time switching (TS) and power splitting (PS). In TS, the source node transmits energy to the relay node for αT s, and the remaining $(1 - \alpha)T$ s are used for information delivery, where $0 \leq \alpha \leq 1$, and T is the total duration of transmission. In PS, a portion ρ of the received power from the source is used for energy harvesting, while the rest, $1 - \rho$, is used for information decoding, where $0 \leq \rho \leq 1$.

Both TS and PS relaying require a separate energy harvester in addition to the data transceiver at the relay. On one hand, PS is slightly more complicated in hardware, as it uses a power splitter, which is not a trivial hardware component. On the other hand, TS requires a dedicated harvesting time, which adds complexity to synchronization and also reduces the effective throughput of the system. Consequently, under similar conditions, TS often has a smaller throughput than PS. Another issue with PS is that it uses the same signal for energy harvesting and information delivery in the broadcast phase. This could be a problem for amplify-and-forward (AF) relaying, as it uses the harvested energy to forward the rest of the signal directly without any further processing. If the value of ρ is small, a small amount of harvested energy would be used to forward a strong signal; and if the value of ρ is large, a large amount of harvested energy would be used to forward a weak signal, both leading to a weak forwarded signal. Thus, PS often has a smaller transmission range than TS in AF relaying.

Research studies in TS or PS WPR systems have mainly focused on the analysis of the relaying performance and its optimization with respect to α and ρ . Figures 2a and 2b show the scheduling of SWIPT relaying. The SWIPT relaying system in [4] is different in that the relay node can harvest energy from large-scale

network interference or from self-interference in full duplex nodes. Multiple antennas can be employed at either the source node or the relay node to perform beamforming or antenna selection for diversity gain. These are not discussed in detail here.

NON-SWIPT: HARVESTING FROM EXTERNAL SOURCES

Another type of RF powered relaying system harvests energy from external sources (e.g., crowd RF transmissions). As the energy and information come from different sources, they do not assume a standard SWIPT structure. However, since the transmission times of the nodes in the same network are normally scheduled by a network-level central controller, the relay can use the transmission times of all other nodes as dedicated harvesting time, and energy is still controlled and correlated with information. Traffic prediction and knowledge of the nodes' transmission schedule will help a harvesting node to build up a statistical understanding of the energy arrival intensity and frequency.

To achieve efficient scheduling, one needs to model the energy arrival or the energy profile as a function of time. This model could be a random process, for example, a Markov process that considers energy state transition [3]. It could be a probabilistic model, for example, a Bernoulli energy injection model with a probability of p that an energy of E is harvested and a probability of $1 - p$ that no energy is harvested, or a simple model with a probability of $1/3$ that no energy is harvested, a probability of $1/3$ that an energy of E is harvested, and a probability of $1/3$ that an energy of $2E$ is harvested [3]. It could also be a deterministic model where the amount and arrival time of the energy are known in advance before scheduling [5].

Using these energy models, scheduling can be formulated as an optimization problem that searches for the best transmission time and the best transmission power. Depending on whether the source node and/or the relay node conduct energy harvesting at the K th time slot ($k = 1, 2, \dots, K$), the available transmission energy $E_s(k)$ and $E_r(k)$ at the source node and the relay node, respectively, will be variable. The optimization is bounded by the energy causality, where the transmission power $P_s(k)$ and $P_r(k)$ at the source and relay nodes, respectively, must be smaller than the available energy from energy harvesting. Moreover, there is an information causality constraint for relaying systems, where the information must be transmitted from the source node before it can be forwarded by the relay node. The available energies at the source and the relay are updated after each transmission. Existing optimization techniques, such as convex optimization, non-convex mixed-integer non-linear program, directional water-filling, and dynamic programming, can be used to find the optimum transmission times and power levels for the source and relay nodes to maximize the sum throughput.

The optimization can be performed for online algorithms that only have causal knowledge of the channel state information (CSI) and energy state information (ESI) as well as for offline algorithms that assume knowledge of the CSI

and energy for all transmissions. For relaying systems, offline algorithms are very complicated, as they require knowledge of transmissions at each of the hops. It can be performed by energy-constrained relay nodes that have limited energy as well as energy-unconstrained relay nodes that have unlimited energy. The size of energy storage can be limited or infinite, which may impose an averaging-out effect on the available energy that has been harvested. Even though not practical, the offline algorithms can provide a performance upper bound for the energy harvesting relay system, or can be applied when the energy arrival can be predicted to some extent. For online algorithms, the optimal scheduling policy can be obtained via policy iteration with the Markov decision process (MDP) formulation of the problem, especially for nodes with finite energy storage. While the general policy iteration suffers from the curse of dimensionality (i.e., high complexity), heuristic algorithms like threshold-based transmission policies are more practical, and in many cases, their performance can be qualitatively analyzed. The optimization can cover various quality of service (QoS) requirements of the traffic being delivered. For example, it can be performed for the delay-constrained case when the relay node must forward the information upon reception, or for the non-delay-constrained case when the relay node does not need to forward the information immediately after reception. In the non-delay-constrained case, the binary indicator $d_s(k) = 0$ ($d_r(k) = 0$) means no information transmission in the K th frame, and $d_s(k) = 1$ ($d_r(k) = 1$) means information transmission in the K th frame at the source (relay) node. Alternatively, one can minimize the total relay transmission time instead of maximizing its throughput.

Furthermore, the SWIPT and non-SWIPT modes can be combined. For example, a SWIPT mode could be activated when insufficient energy is harvested from the non-SWIPT mode. However, this combination brings challenges too. First, since the relaying system and the ambient systems may not operate on the same frequencies, multiple energy harvesters may be required with increased hardware complexity. Second, since the activity of one mode depends on the other mode, the optimization of α and ρ in SWIPT and the optimization of transmission power and time in non-SWIPT will be technically difficult.

CROWD HARVESTING FROM EXTERNAL RADIO TRANSMISSIONS

There has been rapid growth in the density of fixed and mobile wireless devices globally. While the increase in transmitter density will undoubtedly increase the amount of RF energy available in urban environments, it is difficult to quantify the amount of energy reliably available for any given energy harvesting device over some arbitrary time period. The lack of certainty is due to three complexities:

- The random nature of RF transmitter locations
- Randomness in the RF propagation channel
- Variations in spectrum utilization due to traffic patterns

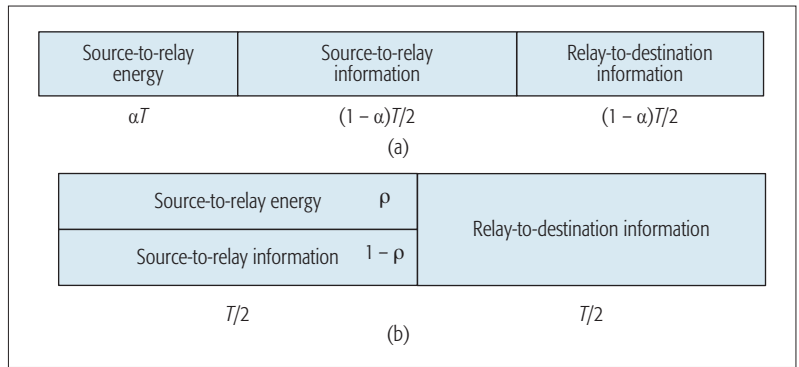


Figure 2. TS and PS for SWIPT wireless powered relaying: a) TS for SWIPT wireless powered relaying; b) PS for SWIPT wireless powered relaying.

STATISTICAL MODELING: DISTANCE DISTRIBUTIONS

Considering energy harvesting from a large number of fixed RF transmitters, it is possible to calculate the deterministic path loss for each channel (i.e., ray-tracing) and predict the energy harvesting performance. However, several challenges exist, two of which are:

- *Computational complexity* grows linearly with the number of energy harvesting devices.
- *Imperfect knowledge of transmitters' locations* is a problem when private small cells and mobile user equipments (UEs) are considered as energy harvesting sources.

Hence, deterministically predicting the performance is challenging, and statistical approximations may be useful as bounds. Stochastic geometry studies random spatial patterns. The underlying principle is that the locations of network transmitters are random in nature, but their spatial density and relative distances follow stable distributions. This can be used to create tractable statistical frameworks for analyzing energy harvesting performance [6, 7].

In order to estimate the energy received from a large number of RF transmitters, one needs to know the probability distribution of the distance between the n th nearest RF transmitter and the energy harvesting device. To demonstrate this, let us consider for a moment that all the transmitters follow a certain spatial distribution with node density Λ . For example, there is a strong body of evidence that macro-BSs are distributed following a Poisson point process (PPP), that is, each macro-BS is deployed independent of others. In the literature, the probability density function (pdf) of the distance r between the energy harvesting device at an arbitrary location and the n th nearest macro-BS is given by [6] $f_{BS,R_n}(r; n, \Lambda)$. The top of Fig. 3b shows the $n = 1$ case for PPP distributed macro-BSs, which follows a Rayleigh distribution.

As for small cells (e.g., Wi-Fi APs and home femto-BSs), there is a lack of knowledge about the distance distribution due to the lack of large-scale empirical data. Two types of small cells exist:

- Operator deployed
- Privately deployed

Existing research has largely focused on the former, whereby it has been proposed that the distribution of small cells can follow Poisson cluster processes (PCPs). Using U.K. Census data as a

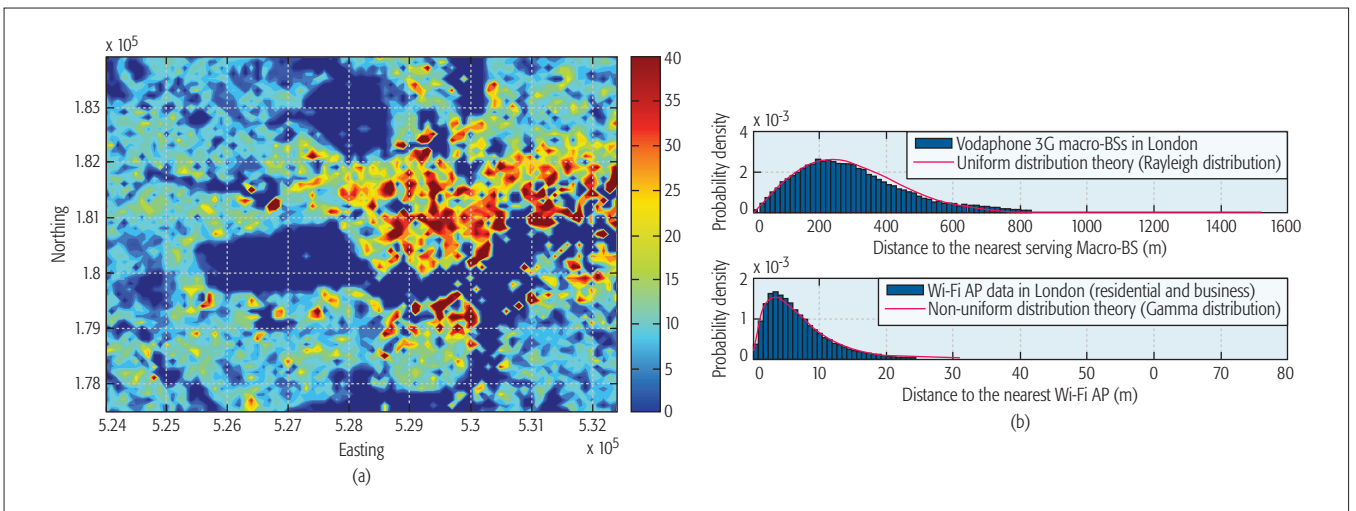


Figure 3. Case study of central London’s 3G cellular and Wi-Fi networks: a) density map of residential and business Wi-Fi APs; b) distance distributions from a random point to the nearest ($n = 1$) macro-BS and Wi-Fi AP transmitter. Data obtained from the U.K. Census (2011–2015) and the Office of Communications (Ofcom).

RAT	Peak power density
Macro-BS downlink	11 fW/Hz
Femto-BS downlink	24 fW/Hz
Wi-Fi downlink	9 fW/Hz
TV broadcast	7550 fW/Hz
RAT	Peak power (LOS)
Macro-BS downlink	0.21 μ W
Femto-BS downlink	0.47 μ W
Wi-Fi downlink	0.18 μ W
TV broadcast	151 μ W

Table 1. Case study results with different RATs.

proxy, we infer small cell locations from household and business population and location data (U.K. Wi-Fi penetration is 89.8 percent). The evidence gathered from a statistically significant amount of data indicates that the small cell distance distribution is a non-uniform clustered spatial process, whereby the nearest distance distribution closely matches a Gamma distribution (Fig. 3b). It remains an open area of research to explore the precise spatial process behind small cells and the impact it has on crowd energy harvesting.

ENERGY HARVESTING SCALING LAWS AND RELIABILITY

In this subsection, we consider the aggregated RF power density (Watts per Hertz) over a bandwidth of B Hz and from an area with an average transmitter density of Λ .

Upper-Bound (Full Spectrum Utilization) with Dual-Slope Path Loss: In the upper bound analysis, we assume that: all transmitters i) are transmitting across the whole spectrum available, and ii) are emitting with the maximum allowable power spectrum density on all frequency bands. Leveraging the spatial distributions of RF transmitters found in the previous subsection, research in [6] found that the received power density P_{RX} is linearly proportional to the bandwidth B and the transmit power density P_{TX} , and has a complex monotonic relationship with the distance-depen-

dent path loss exponent α and the node density Λ . Given that the energy from each transmitter will vary significantly depending on whether the propagation is largely line-of-sight (LOS, $\alpha = 2$) or non-LOS (NLOS, $\alpha > 2$), one can consider a dual-slope approach (as shown in Fig. 4), where two sets of path loss exponents are considered [8]: typically $\alpha = 2$ for LOS free-space propagation and $\alpha > 4$ for NLOS urban propagation. As a result, the total power (averaged over distance) harvested from K RF transmitters (each following its own spatial distribution) follows the following scaling rules [6]:

- Linear with transmit power: $P_{RX} \propto P_{TX}$
- Polynomial with transmitter density: $P_{RX} \propto (\Lambda)^{a/2}$

As a case study, we consider the central London area (60 km^2 as shown in Fig. 3a) with network parameters for multiple radio access technologies (RATs):

- Cellular macro-BS downlink (20 MHz bandwidth, 40 W, $0.3\text{--}5/\text{km}^2$, real locations)
- Cellular femto-BS downlink (20 MHz, 1 W, $15\text{--}200/\text{km}^2$, PCP distributed);
- Wi-Fi AP downlink (60 MHz, 100 mW, $50\text{--}1000/\text{km}^2$, proxy locations)
- TV broadcast (100 MHz, 1000 kW, $0.01\text{--}0.2/\text{km}^2$, real locations)

The path loss model considered is the WINNER model² with the appropriate shadow fading for different urban propagation environments.

Figure 4 shows the upper bound power harvested from different RATs as a function of transmitter density Λ and different path loss exponent values a . In particular, two exponent values are considered: (top) $a = 2$ (free-space), and (bottom) $a = 4.3$ (urban NLOS propagation in the WINNER model). Table 1 shows the peak harvested power (W) and power density (W/Hz) for different RATs (full spectrum usage). The results are obtained from extensive Monte Carlo simulations in a manner similar to [6]. It can be seen that the greatest opportunity for power harvesting lies in the TV broadcast channels. Given that the network traffic of small cells and TV is typically higher than that of macro-BSs, it is

² The Wireless World Initiative New Radio (WINNER) is a statistical radio propagation model (2–6 GHz) for link- and system-level simulations of a variety of short-range and wide area wireless communication systems.

advisable to focus crowd energy harvesting in these bands for relaying. In Fig. 4 (top), the free-space (LOS) results match those found in existing field test observations. For example, it was found that 100 μW can be achieved at a 20 km distance from a 150 kW TV station [9]. Looking ahead, we do not expect the node density for macro-BSs and Wi-Fi hubs to change over the coming years, but the estimates are that the femto-BS density will increase by at least 20-fold. Therefore, we expect a polynomial increase (exponent $a/2$, where the value of a is typically 2–4) in the power available for harvesting. This potentially will lead to femto-BSs acting as an alternative or complementary source of crowd harvesting RF energy.

As for NLOS energy harvesting, as shown in Fig. 4 (bottom), the values are several orders of magnitude lower due to the more significant distance-dependent path loss, affecting the long-range TV signals more than the short-range Wi-Fi signals. Therefore, it is better to use Wi-Fi bands for NLOS energy harvesting, as opposed to TV. A key consideration from these results is that a relay performing RF harvesting would also require a band for information delivery, which should be the bands that present the least energy harvesting potential. While the nearest node accounts for 89 percent of the energy from crowd harvesting, in realistic networks, the nodes may not transmit at full buffer continuously, and analysis that incorporates traffic patterns is necessary to understand the reliability advantage of crowd energy harvesting over nearest node harvesting.

Realistic Traffic Load (Variable Spectrum Utilization): In order to estimate the realistic time-dependent RF energy, it is important to consider the spectrum utilization over time for each RF transmitter, which depends on the traffic load of each transmitter. Leveraging a well-known statistical model of third generation (3G) high-speed packet access (HSPA) networks [10], one can infer the spectrum utilization pattern at each BS as a ratio of the traffic and the peak capacity of the BS. Given that the N BSs are independent and identically distributed in space and in spectrum utilization, the pdf of the power density (Watts per Hertz) harvested from all N RF transmitters is the linear combinations of random variables. This corresponds to the convolution of probability distributions if the traffic random variables are independent. Therefore, there is N -fold *continuous convolution* of the traffic load pdf (i.e., $f_{L1} * f_{L2} * f_{L3} * \dots * f_{LN}$). These foundation statistical results and future research will provide useful guidelines to designing crowd harvesting powered relaying systems under variable spectrum utilization.

OPTIMIZATION FOR CROWD HARVESTING

We have so far reviewed SWIPT and non-SWIPT relaying, and in particular how non-SWIPT relaying can benefit from crowd RF energy harvesting across different RATs, which is attractive, especially in the TV bands (LOS) and Wi-Fi bands (NLOS). However, what remains unclear is how a relay system, where the nodes are sufficiently apart (and hence have different energy harvesting potentials), can collaborate to achieve optimal relaying performance. In this section we discuss node col-

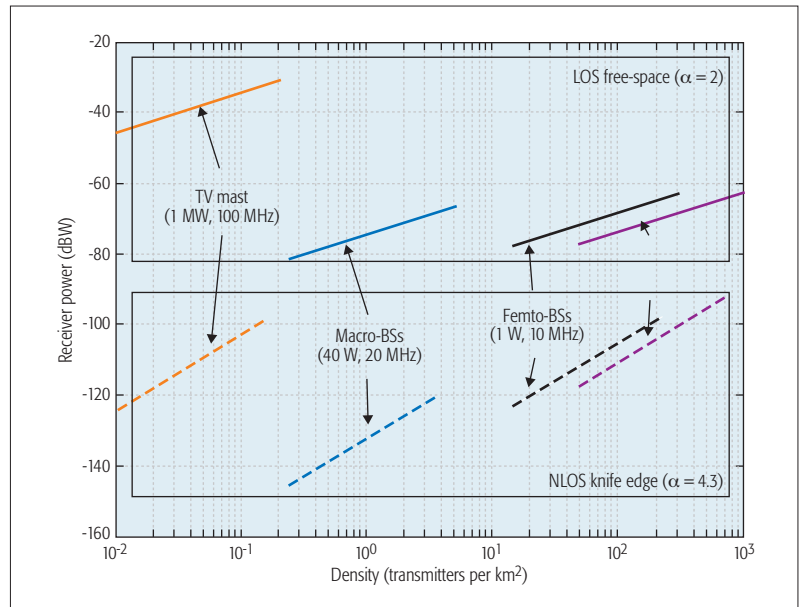


Figure 4. Upper bound power harvested from different RATs as a function of transmitter density Λ and path loss scenarios.

laboration and transmission scheduling with QoS guarantees for relaying with crowd harvesting.

NODE COLLABORATION

It has been revealed that the correlation distance of the traffic density is less than 80 m in urban areas [11], indicating that the RF energy harvesting process may follow similar spatial correlation. Two nodes that are more than 100 m apart tend to have almost independent energy harvesting processes, and thus node collaboration can be performed to exploit the independent relationship between energy profiles (e.g., to achieve energy harvesting diversity gains).

First, we illustrate the benefit of node collaboration via combining the SWIPT and ambient energy harvesting in order to compensate for the possible energy shortage at either source or relay as shown in Fig. 1. Assuming that the source can harvest more ambient RF energy than the relay node, one can introduce new TS parameters α_1 and α_2 ($0 \leq \alpha_1, 0 \leq \alpha_2, \alpha_1 + \alpha_2 \leq 1$), whereas the energy harvesting phase can be split further into two parts, with $\alpha_1 T$ s and $\alpha_2 T$ s, for, respectively:

- Source-to-relay energy delivery
- Ambient RF energy harvesting at relay

Note that the source can make use of the time when the relay forwards the message to harvest additional ambient RF energy. For PS, a similar protocol can be adopted, with power splitting factors ρ_1 and ρ_2 ($0 \leq \rho_1, 0 \leq \rho_2, \rho_1 + \rho_2 \leq 1$) for source energy delivery and ambient RF energy harvesting, respectively. While this may require another energy harvesting component, a mechanism of hybrid TS and PS can be a promising solution without additional hardware requirements.

Furthermore, for the scenario where the nodes have no SWIPT structure or have some common information to the same destination (e.g., the multiple relays in the second hop of a relaying transmission) or multiple sensors that sense the same target and need to deliver the sensing results to the sink, collaborative trans-

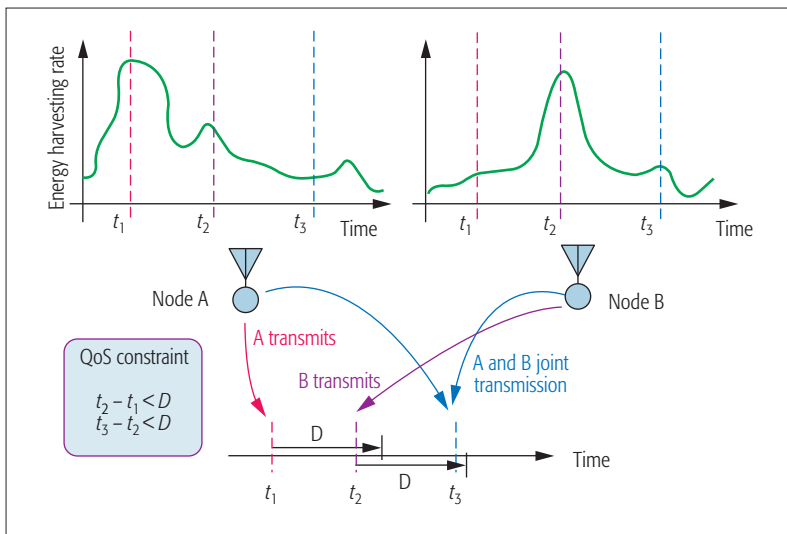


Figure 5. Transmission scheduling and node collaboration under independent energy arrival processes and inter-delivery time requirement.

mission can be used to address the uneven energy arrival rates (e.g., the energy arrival process illustrated in Fig. 5) for nodes A and B. As a simple example, a transmission frame can be divided into two subframes. In the first subframe, only one of the nodes can be scheduled to transmit in the conventional way. In the second subframe, multiple nodes can perform simultaneous joint transmission (JT) to the destination with distributed beamforming. To this end, the frame division portion ξ ($0 \leq \xi \leq 1$) and the node scheduling should be jointly optimized, taking into account the ESI of all nodes.

For both cases, to get the optimal system parameters α_i , ρ_i , and ξ , practical online algorithms should be designed based on the prediction of the mobile traffic that generates the crowd energy harvesting source. One can model the mobile traffic variation with a Markovian model, the transition probabilities of which can be trained based on real data as presented earlier, and then MDP policy iteration will provide the optimal transmission scheduling and system parameters. To reduce the complexity of MDP, one can do conventional optimization on a per-frame basis, with the energy arrivals and channel conditions of several future frames as known, given high traffic prediction precision.

DELAY QoS GUARANTEE

As a type of *delay-sensitive* traffic, the transmission of the sensing data is subject to a temporal regularity constraint (e.g., the inter-delivery time requirement [12]), as a large gap between updates can cause system instability. To guarantee such a delay QoS requirement, the transmission scheduling should utilize the *diversity* from the different energy harvesting processes of nodes. For instance, in Fig. 5, there are three consecutive delivery instants, t_1 , t_2 , and t_3 , and it is required that the inter-delivery time should be less than D . At time t_1 , node A has a high energy harvesting rate and can afford to deliver the sensing data to the sink, and likewise at time t_2 , node B has enough energy to deliver the data. However at time t_3 , both nodes have low energy arrival

rates and the inter-delivery time constraint D is about to break; thus, they use collaborative JT to ensure reliable transmission to the sink.

Some delay-insensitive applications require that important events are sensed or sampled with finer time granularity, but can be delivered in a less timely manner. In this case, the sensor node needs to act proactively according to the energy stored in its battery and the prediction of future energy arrivals. If the battery is about to overflow, the node can send the sensing results in a batch to the destination/sink, while maintaining enough energy for possible sensing efforts when important events happen. By optimizing the energy storage in the battery and the scheduling of transmission/sensing activities, the node can equivalently *match* the energy arrivals and the occurrence of sensing events over time.

OPEN CHALLENGES

WPR is a relatively new technique. It has strong potential for the emerging M2M communications in IoT systems, but a number of open challenges exist. Accurate *energy modeling* is important. Most existing work merely considers the energy used for transmission. However, the energy consumption for receiving is comparable to that of transmission and cannot be ignored [13]. Therefore, scheduling decisions should take into account the energy consumption of receiving as well, and a relay node should switch among receiving, forwarding, and keeping silent modes to save energy, based on its ESI, the CSI of the multiple hops, and the delay constraints. In terms of optimization, an area of interest that has not been mentioned in this article is *relay selection*, which provides a trade-off between complexity and performance. However, the relay selection criterion for information transmission may be different from that for energy transfer. For example, previous research has shown that the relay node offering the largest end-to-end signal-to-noise ratio may not be the one with the largest received power [14]. In SWIPT WPR, it is not clear which selection criterion should be used to achieve the best performance. In practice, the distances between nodes will significantly affect the system performance, as existing energy harvesters work efficiently over short distances. Therefore, a practical challenge is to decide whether it is better to use multihop with shorter distances.

REFERENCES

- [1] A. Lo, Y. Law, and M. Jacobsson, "A Cellular-Centric Service Architecture for Machine-to-Machine Communications," *IEEE Wireless Commun.*, vol. 20, 2013, pp. 143–51.
- [2] A. Nasir *et al.*, "Relaying Protocols for Wireless Energy Harvesting and Information Processing," *IEEE Trans. Wireless Commun.*, vol. 12, 2013, pp. 3622–36.
- [3] N. M. B. Medepally, "Voluntary Energy Harvesting Relays and Selection in Cooperative Wireless Networks," *IEEE Trans. Wireless Commun.*, vol. 9, 2010, pp. 3543–53.
- [4] I. Krikidis, "Simultaneous Information and Energy Transfer in Large-Scale Networks With/Without Relaying," *IEEE Trans. Commun.*, vol. 62, 2014, pp. 900–12.
- [5] C. Huang, R. Zhang, and S. Cui, "Throughput Maximization for the Gaussian Relay Channel with Energy Harvesting Constraints," *IEEE JSAC*, vol. 31, 2013, pp. 1469–79.
- [6] W. Guo and S. Wang, "Radio-Frequency Energy Harvesting Potential: A Stochastic Analysis," *Trans. Emerging Telecommun. Technologies*, vol. 24, 2013, pp. 453–57.
- [7] I. Flint *et al.*, "Performance Analysis of Ambient RF Energy Harvesting: A Stochastic Geometry Approach," *IEEE GLOBECOM*, 2014, pp. 1448–53.

- [8] X. Zhang and J. Andrews, "Downlink Cellular Network Analysis with Multi-Slope Path Loss Models," *IEEE Trans. Commun.*, 2015.
- [9] T. Ajmal *et al.*, "Design and Optimisation of Compact RF Energy Harvesting Device for Smart Applications," *Electronics Letters*, vol. 50, no. 2, 2014, pp. 111–13.
- [10] M. Laner *et al.*, "Users in Cells: A Data Traffic Analysis," *IEEE WCNC*, Apr. 2012, pp. 3063–68.
- [11] D. Lee *et al.*, "Spatial Modeling of the Traffic Density in Cellular Networks," *IEEE Wireless Commun.*, vol. 21, no. 1, Feb. 2014, pp. 80–88.
- [12] X. Guo *et al.*, "A High Reliability Asymptotic Approach for Packet Inter-Delivery Time Optimization in Cyber-Physical Systems," *ACM MobiHoc '15*, June 2015, pp. 1–10.
- [13] S. Zhou *et al.*, "Outage Minimization for A Fading Wireless Link With Energy Harvesting Transmitter and Receiver," *IEEE JSAC*, vol. 33, no. 3, 2015, pp. 496–511.
- [14] Y. Chen *et al.*, "Novel Partial Selection Schemes for af Relaying in Nakagami-m Fading Channels," *IEEE Trans. Vehic. Tech.*, vol. 60, 2011, pp. 3497–3503.

BIOGRAPHIES

WEISI GUO [M] (weisi.guo@warwick.ac.uk) received his M.Eng., M.A., and Ph.D. degrees from the University of Cambridge. He is currently an assistant professor and co-director of the Cities research theme at the School of Engineering, University of Warwick, United Kingdom. He was the recipient of the IET Innovation Award 2015 and a finalist in the Bell Labs Prize 2014. He is a co-inventor of the world's first molecular communication prototype and has published over 80 papers. His research interests are in the areas of heterogeneous networks, molecular communications, complex networks, and mobile data analytics.

SHENG ZHOU [M] (sheng.zhou@tsinghua.edu.cn) received his B.S. and Ph.D. degrees in electronic engineering from Tsinghua University, China, in 2005 and 2011, respectively. He is currently an associate professor in the Electronic Engineering Department, Tsinghua University. From January to June 2010, he was a visiting student at the Wireless System Lab, Electrical Engineering Department, Stanford University, California. His research interests include cross-layer design for multiple antenna systems, cooperative transmission in cellular systems, and green wireless communications.

YUNFEI CHEN [SM] (yunfei.chen@warwick.ac.uk) received his B.E. and M.E. degrees in electronic engineering from Shanghai Jiaotong University, P.R.China, in 1998 and 2001, respectively. He received his Ph.D. degree from the University of Alberta in 2006. He is currently working as an associate professor at the University of Warwick. His research interests include wireless communications, cognitive radios, relaying, energy harvesting, and fading channels.

SIYI WANG [M] (siyi.wang@xjtlu.edu.cn) received his Ph.D. degree in wireless communications from the University of Sheffield, United Kingdom, in 2014 when he was a research member of the project Core 5 Green Radio funded by the Virtual Centre of Excellence (VCE) and U.K. EPSRC. He is currently a lecturer at Xi'an Jiaotong-Liverpool University (XJTLU). He has published over 40 IEEE papers in the past four years. His research interests include molecular communications, indoor-outdoor network interaction, small cell deployment, device-to-device communications, stochastic geometry, theoretical frameworks for complex networks, and urban informatics.

XIAOLI CHU [SM] (x.chu@sheffield.ac.uk) is a lecturer at the University of Sheffield. She received her Ph.D. degree from the Hong Kong University of Science and Technology in 2005. From 2005 to 2012, she was with Kings College London. She has published more than 90 peer-reviewed journal and conference papers. She is the lead Editor/author of *Heterogeneous Cellular Networks* (Cambridge University Press, 2013). She was Co-Chair of the Wireless Communications Symposium at IEEE ICC 2015, and a Guest Editor for *IEEE Transactions on Vehicular Technology* and the *ACM/Springer Journal of Mobile Networks & Applications*.

ZHISHENG NIU [F] (niuzhs@tsinghua.edu.cn) graduated from Beijing Jiaotong University, China, in 1985, and got his M.E. and D.E. degrees from Toyohashi University of Technology, Japan, in 1989 and 1992, respectively. During 1992–1994, he worked for Fujitsu Laboratories Ltd., Japan, and in 1994 joined Tsinghua University, Beijing, China, where he is now a professor in the Department of Electronic Engineering. He is also a guest chair professor at Shandong University, China. His major research interests include queueing theory, traffic engineering, mobile Internet, radio resource management of wireless networks, and green communication and networks.

LIFETEL: Managing the Energy-Lifetime Trade-off in Telecommunication Networks

Luca Chiaraviglio, Lavinia Amorosi, Andrea Baiocchi, Antonio Cianfrani, Francesca Cuomo, Paolo Dell'Olmo, and Marco Listanti

The authors present a novel framework to optimize the lifetime of telecommunication networks exploiting sleep mode strategy as an outcome of the LIFETEL project. They first consider the main effects impacting the device lifetime, showing that the transitions between active mode and sleep mode may drastically reduce the device lifetime. They then propose a methodology to limit the impact on the device lifetime when SM states are exploited.

ABSTRACT

We present a novel framework to optimize the lifetime of telecommunication networks exploiting sleep mode (SM) strategy as an outcome of the LIFETEL project. We first consider the main effects impacting the device lifetime, showing that the transitions between active mode (AM) and SM may drastically reduce the device lifetime. We then propose a methodology to limit the impact on the device lifetime when SM states are exploited. Additionally, we report the main outcomes obtained from cellular and backbone scenarios. Results show that a lifetime-aware approach is of fundamental importance for limiting the impact of failures on network devices, while at the same time allowing adequate energy saving. Finally, we discuss the main challenges that still need to be faced for full adoption of lifetime-aware strategies in an operator network.

INTRODUCTION

In the last years, the energy consumption of the Internet has been constantly increasing due to the proliferation of connected devices, as well as the variety of offered services. Additionally, future trends forecast an increase of power consumption in backbone and cellular networks. In this context, the topic of “green” networking is being investigated by the research community, with several works targeting the reduction of energy consumption in backbone and cellular networks (e.g., see [1, 2] for detailed surveys).

Among the different solutions proposed so far in the literature to reduce power consumption in telecommunication networks, one of the most promising approaches is the application of a sleep mode (SM) state of devices. The main idea of SM-based solutions is to exploit the fact that current networks are dimensioned for peak traffic. However, traffic varies over time, and it is lower at night. Thus, it is possible to exploit this excess of capacity in order to put different network devices in SM when traffic is low.

Although the benefits of SM approaches in terms of energy saving are clear, their application in operator networks is posing additional challenges. In particular, one of the main concerns of network operators is the impact of SM on the device lifetime. More specifically, the network devices are designed to be always powered on, and their frequent activation/deactivation may

impact their lifetime [3, 4] (i.e., the amount of time between a device failure and the following one). A shorter lifetime leads to two negative effects: i) a quality of service (QoS) degradation experienced by users, and ii) an increase in the costs to repair the devices (or replace them in the worst case).

In this context, energy efficiency is not the only metric that should be considered, since high energy savings may lead to an unacceptable increase in failure rate [4]. Our goal is therefore to consider the energy-lifetime trade-off that emerges in telecommunication networks in order to propose a new approach. More in depth, we report the main outcomes of LIFETEL, an 18-month project funded by the University of Rome Sapienza. To the best of our knowledge, LIFETEL is the first project proposing solutions to reduce power consumption while at the same time considering the device lifetime. We face a multi-domain problem, considering both backbone and cellular networks to be lifetime-aware. Our results show that the proposed framework opens new challenges in operator networks. In particular, we claim for measurements of hardware (HW) parameters related to failures of devices with SM, as well as a deeper investigation of failure events in energy-efficient networks.

The rest of the article is organized as follows. The investigation on how SM impacts the device lifetime is reported. The LIFETEL framework is reported. A set of results (obtained from both backbone and cellular scenarios) is presented. We review the main challenges triggered by the proposed lifetime-aware approach. Finally, conclusions are drawn.

HOW SLEEP MODE IMPACTS DEVICE LIFETIME

We first consider a simple toy case example; then we give details on the main physical phenomena related to failures, and we introduce the lifetime models.

A SIMPLE TOY-CASE EXAMPLE

We provide a simple example, reported in Fig. 1, showing a general energy-saving algorithm and its comparison with a lifetime-aware solution. Figure 1a details the topology and link capacities. A single traffic relationship from node 1 to node 4 is considered (i.e., all the other node pairs are not exchanging traffic); the traffic variation over a period of 7 time slots (TSs) is reported in Fig. 1b. This trend may represent a typical day/night

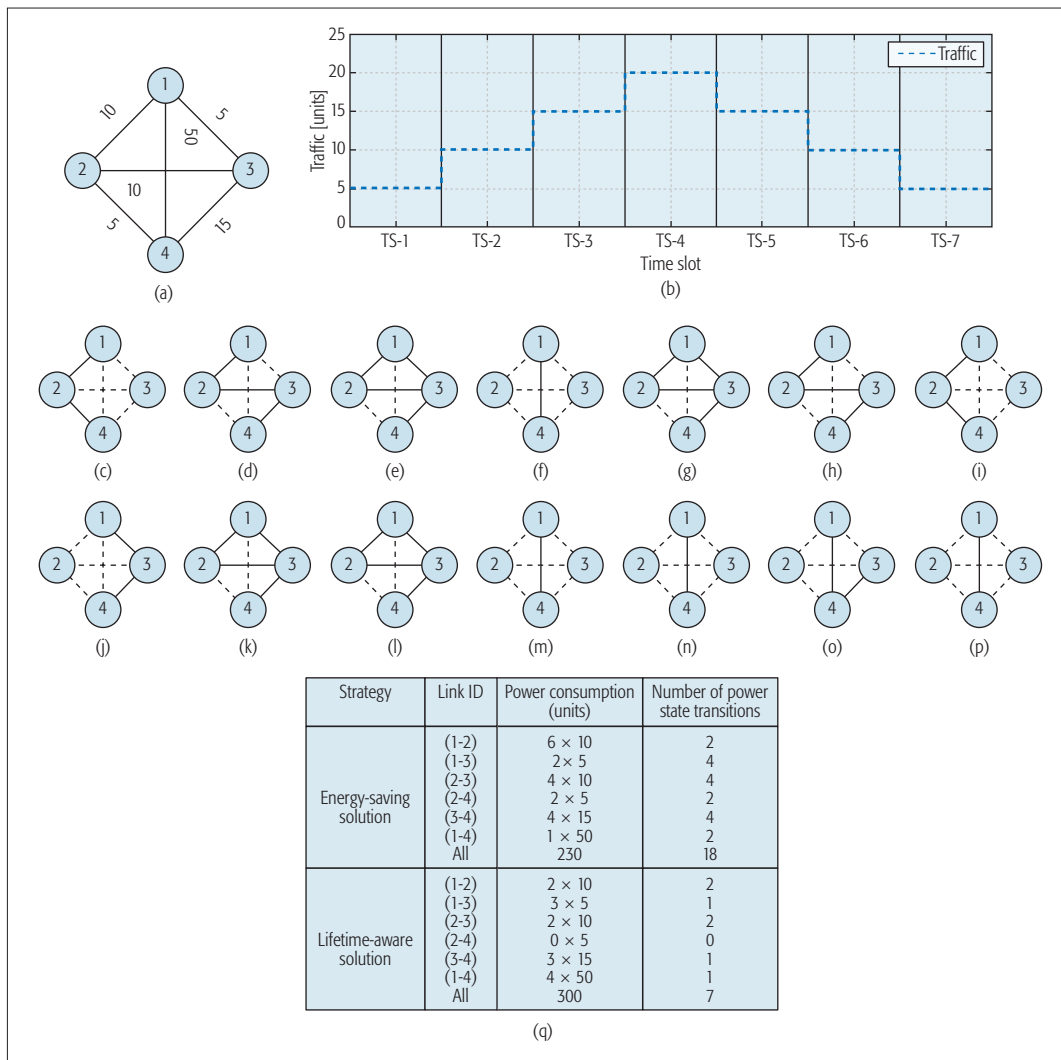


Figure 1. Comparison among an energy-saving solution and a lifetime-aware one in a toy-case scenario: a) network topology and link capacities (units); b) traffic for node pair 1-4 vs. time slot (TS) ID (the other node pairs are not exchanging traffic); c) energy-saving configuration for TS1; d) energy-saving configuration for TS2; e) energy-saving configuration for TS3; f) energy-saving configuration for TS4; g) energy-saving configuration for TS5; h) energy-saving configuration for TS6; i) energy-saving configuration for TS7; j) lifetime-aware configuration for TS1; k) [lifetime-aware configuration for TS2; l) lifetime-aware configuration for TS3; m) lifetime-aware configuration for TS4; n) lifetime-aware configuration for TS5; o) lifetime-aware configuration for TS6; p) lifetime-aware configuration for TS7; q) power consumption and number of power state transitions for each link in the network and in total for the energy-saving and lifetime-aware solutions.

behavior, that is, the traffic is lower at night and higher during the day. For the sake of simplicity, we assume that the link power consumption is proportional to the installed capacity. Moreover, the power spent in SM is negligible. In this scenario we apply a classical energy-saving approach: the amount of capacity in SM is maximized during each time slot while satisfying the traffic requirements. In this way, the power consumption is minimized. Moreover, we also consider a lifetime-aware approach. In this case, the resulting network configurations trade between:

- Increasing the amount of capacity in SM
- Limiting the number of power state transitions (more details about a general lifetime-aware algorithm are provided later).

Figures 1c–1p report the network configurations for both energy-saving and lifetime-aware solutions.

Figure 1q reports the comparison between energy-saving and lifetime-aware, provided in terms of:

- The resulting power consumption, considering that the power consumption of network links is proportional to their (installed) capacity and each time slot lasts for one unit
- The number of power state transitions

The main outcome of this evaluation is that the lifetime-aware approach is able to greatly reduce the number of power state transitions (more than 60 percent with respect to energy-saving), with a limited increase of the power consumption (about 30 percent). Moreover, the power saving capability of the lifetime-aware solution can be proven considering a classical approach with all network links always on: it can easily be verified that the lifetime-aware approach is able to reduce the power consump-

The main outcome of this evaluation is that the lifetime-aware approach is able to greatly reduce the number of power state transitions (more than 60 percent with respect to energy-saving), with a limited increase of the power consumption (about 30 percent).

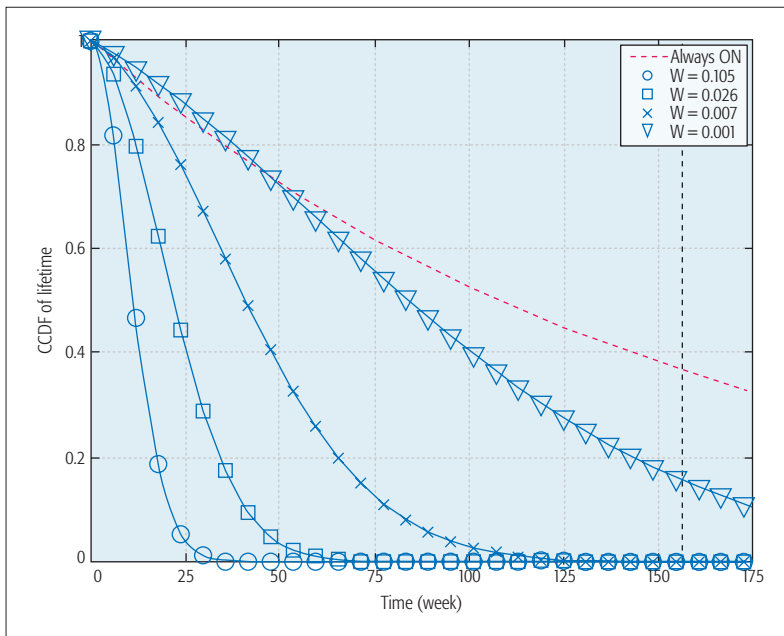


Figure 2. Results of the analytical model in [9] of an LTE base station lifetime under adaptive energy saving policies and fixed state operation (W is the ratio between the failure rate increase due to a single power transition and the failure rate in active mode).

tion by about 55 percent with respect to the always-on solution. A key question is, how is it possible to control the trade-off between power consumption and lifetime? Answering this is the purpose of the following sections.

PHYSICAL PHENOMENA AND SYSTEM-LEVEL LIFETIME MODELS

A comprehensive description of the main effects on the lifetime triggered by SM application is reported in [3, 4]. We refer to these works for more details, while here we report the main intuitions. In brief, when SM is applied, the device temperature tends to be decreased, and consequently its lifetime is increased. This behavior is taken into account by the Arrhenius model [5]. However, the power state transitions introduce temperature variations that may decrease the lifetime. This effect is captured by the Coffin-Manson model [6, 7] and confirmed by in-the-field experiments (like the ones reported by [8] for server machines). The overall long-term effect of the application of SM-active mode (AM) transitions is the result of the two contrasting effects on the failure rate [4].

In the following, we briefly review the system-level model of [4] for the lifetime of a device exploiting SM. The total lifetime is composed of three terms:

- The failure rate in SM, measured for the amount of time the device is in SM (normalized to the total period of time under consideration)
- The failure rate in AM, γ^{AM} ,¹ measured for the amount of time the device spends in this state (normalized again to the total period of time)
- The failure rate increase due to a single transition multiplied by the number of transitions

From this model, we define two adimensional parameters:

- K , which is the ratio between the failure rate in SM and the failure rate in AM
- W , which is the ratio between the failure rate increase due to a single transition and the failure rate in AM

In this model, γ^{AM} , K , and W are HW parameters that depend only on the components used to build the device. On the contrary, the amount of time spent in SM and the number of power state transitions depend on the SM policy adopted (which is governed by a network-wide process).

More insight can be gained with an analysis that accounts for the time dynamics of the interplay between energy saving policies and device stress due to transitions. A Markov chain model of the time dynamics of a device subject to both processes is derived in [9]. The state transitions are driven by an external control process, such as representing the outcome of energy saving policies. More in depth, illustrative results of the model are shown in Fig. 2, in the case of application to a Long Term Evolution (LTE) cellular access network, comprising LTE base station (BS) hotspots and an umbrella macrocell in an urban environment. A time varying user traffic profile and the operation of the energy-aware Least Load Algorithm (LLA) [10] have been simulated, and the state transition probability matrix of a typical LTE BS has been estimated. More details on the simulation model are provided. The probability distribution of the lifetime of the BS is derived by accounting for the state-dependent failure rate (AM and SM states) and for the accumulation of failure rate increments due to transitions along the history of the device operation.

Figure 2 shows the probability that the BS lifetime exceeds a time t (i.e., the CCDF) as a function of t for $1/\gamma^{AM} = 3$ years,² $K = 0.15$, and different values of W . The K parameter has been set by assuming that:

- The average temperature of the device in SM and AM states is equal to 290°K and 320°K, respectively.
- The activation energy entering the Arrhenius law is set to 0.5 eV.

Concerning W , we assume that the number of transitions that the device may sustain is set to 500, 1000, 2000, and 5000 for $W = 0.105$, 0.026, 0.007, and 0.001, respectively. Albeit the physical parameter numerical values are somewhat arbitrary (yet reasonable), the qualitative behavior of the lifetime is unaffected, and the model can obviously easily be adapted to other parameter values (e.g., provided by an experimental device characterization).

In the long term (i.e., Fig. 2, right), device stress due to state transitions wear out the BS statistically sooner with an energy saving policy compared to the case in which the BS is always in AM (“always on” label in the subfigure). In the short term (i.e., until week 50), the variable state operation can bring a slight lifetime statistical advantage for the smallest level of W , thanks to the beneficial effect of putting the BS in the SM state for some time. The curves shown in Fig. 2 point out that for bigger W , the higher the lifetime reduction. The main finding is that state transitions have a major impact on the lifetime of the device, even if the increment of the failure rate is 1/1000 of the baseline failure rate in the ON state (i.e., $W = 0.001$).

¹ This term is the average failure rate of a device typically in AM. It may also include any kind of normal operation conditions (e.g., software/hardware upgrades, maintenance checks, weather, power failures) except transitions to an SM state due to energy saving policies.

² This is a typical operational lifetime, considering technological evolution and amortization of equipment.

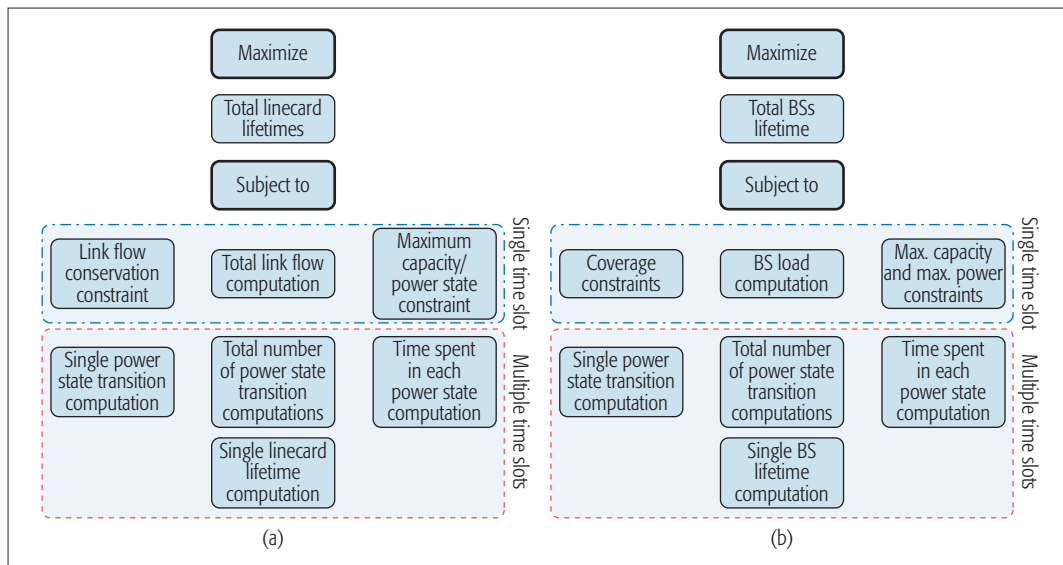


Figure 3. Optimization model for the BS lifetime in backbone and cellular networks: a) backbone networks — OPT-B [11]; b) cellular networks — OPT-C [12].

THE LIFETEL FRAMEWORK

LIFETEL tackles the energy-lifetime trade-off in backbone and cellular networks. We first adopt an optimization approach to formally define the problem. Then, as a second step, we develop efficient algorithms to practically solve it.

OPTIMAL FORMULATIONS

Figure 3 reports the optimal formulation schemes for backbone and cellular networks, respectively. We assume that linecards and BSs exploit SM state, since these devices are normally targeted by energy-efficient solutions [1, 2]. Moreover, the total period of time under consideration is divided in time slots, and for each of them the amount of traffic requested by users is assumed to be known. Focusing first on the backbone case (Fig. 3a), the goal of the problem is to maximize the lifetime of all linecards in the network. This function can be, for instance, the maximization of the average lifetime computed over the set of linecards. However, the objective can be pursued only under the problem constraints. In particular, the traffic in each time slot has to be routed in the network, and connectivity as well as maximum link constraints have to be satisfied. Additionally, the maximum link utilization constraint also means that if the flow on the link is larger than zero, the corresponding linecards are powered on. These constraints are applied for each time slot. Then the problem has to consider the long-term objective of lifetime, which is instead computed over the whole set of time slots (denoted as multiple time slots in Fig. 3a). As a consequence, the lifetime is a function of the amount of time every linecard spends in each power state, as well as the number of power state transitions. Both of these metrics have to be modeled as positive integer variables for each linecard. Then the total lifetime is computed, and it is represented in the objective function. The resulting problem is a mixed integer linear programming (MILP), which is very hard to solve due to the fact that the objective has to be evaluated over multiple time slots. Additionally,

the full knowledge of traffic variation over the whole set of time slots is required. We refer to this formulation as OPT-B.

When a cellular network is considered (Fig. 3b), the optimization goal becomes the minimization of the BSs' lifetime in a given scenario. In particular, each BS has to guarantee coverage, as well as serve the users' traffic, without exceeding the maximum transmitted power of the BSs.

Clearly, the computation of the transmitted power has to also consider the other BSs in the network in order to limit the interference caused to the other neighboring BSs. When a BS is not serving any traffic and coverage is guaranteed by the neighboring BS, the former can be put in SM.

All these constraints have to be satisfied in each time slot. Similar to the backbone case, the lifetime computation is performed over multiple time slots by computing the amount of time spent in each power state, and also the total number of transitions. Also, in this case the resulting problem is a MILP, which is referred as OPT-C in the rest of the article.

LIFETIME-AWARE ALGORITHMS

In order to practically solve the lifetime-aware problem, we have proposed the Acceleration Factor Algorithm (AFA) [13] for backbone networks and the Lifetime-Aware Algorithm (LIFE) [12] for cellular ones. The algorithms are applied for each time slot. The input is the traffic at the current time slot and the power state of devices (i.e., AM or SM) at previous time slots. The output is the current power state of devices, as well as the current lifetime. Initially, the lifetime is computed for all the devices. Additionally, the QoS constraints are checked with the current set of devices powered on. These constraints include connectivity and maximum link utilization for backbone networks, and coverage and maximum BS capacity for cellular ones (as reported earlier). If the QoS constraints are violated, the powering on procedure of Fig. 4a is executed. Otherwise, the SM procedure of Fig. 4b is run. The main idea of this part is to search for can-

The lifetime is a function of the amount of time every device spends in each power state, as well as the number of power transitions. The resulting problem is a mixed integer linear programming (MILP), which is challenging to be optimally solved.

For each device, a check on its lifetime is performed. If the device lifetime falls inside other two thresholds (defined for the lifetime of the single device), the current device is put in SM. Then, if the QoS constraints are satisfied, the device is kept in SM (otherwise the device is powered on).

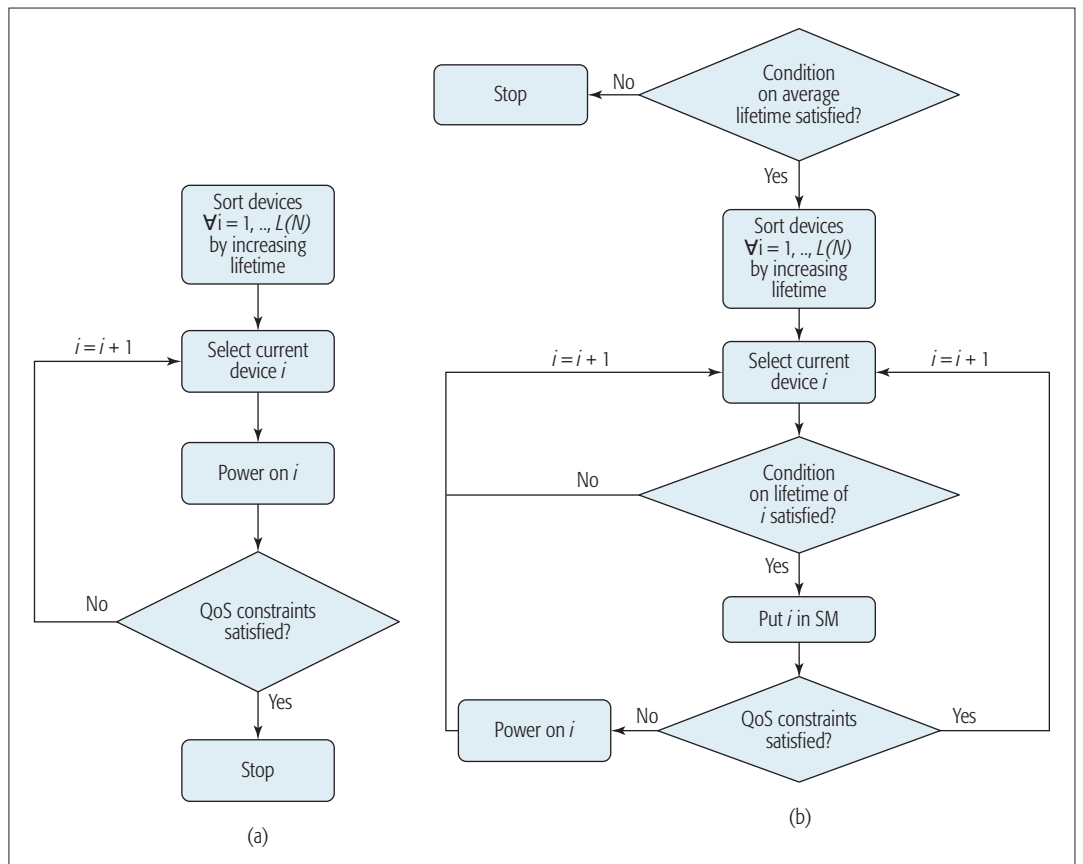


Figure 4. Powering on and SM procedures for AFA (LIFE). N is the number of BSs for LIFE, while L is the number of linecards for AFA: a) powering on procedure; b) SM procedure.

didate devices to put in SM only if the average lifetime (computed over all the devices) falls inside two thresholds, which can be set as input parameters. Then the devices are ordered with increasing lifetime. For each device, a check on its lifetime is performed. If the device lifetime falls inside the other two thresholds (defined for the lifetime of the single device), the current device is put in SM. Then, if the QoS constraints are satisfied, the device is kept in SM (otherwise the device is powered on).³ The entire procedure is then repeated for all the devices. The key idea of this approach is to perform SM decisions only if the lifetime thresholds are met. In this way, we avoid frequently changing the power state for a device that has already decreased its lifetime.

ENERGY-LIFETIME RESULTS

Figure 5 reports an overview of the LIFETEL results, obtained by solving the optimization problem with the CPLEX solver, and the AFA and LIFE heuristics with custom simulators. We have first considered a backbone network, and a time slot duration of one hour for a total period of time T equal to four days. We refer the reader to [14] for a more detailed description of this scenario. In brief, the network is composed of 38 nodes and 72 bidirectional links. The scenario includes link capacities, routing weights, and the 24 traffic matrices (one for each hour of a workday).⁴ The network is dimensioned to satisfy a maximum link utilization equal to 50 percent of the link capacity (during peak hours). Moreover, focusing on

power consumption, we have adopted the same power model of [13], in which each link consumes an amount of power corresponding to a pair of optical transponders and a pair of IP interface ports. Each 10 Gb/s transponder consumes 37 W, and each 1 Gb/s port consumes 10 W. Finally, we assume that the power consumption of a link is negligible when it is put in SM.

Figure 5a reports the normalized lifetime for OPT-B and AFA over the backbone scenario, considering the variation of the HW parameters K and W . The obtained average lifetime (computed over the whole set of linecards in the scenario) is then normalized to the average lifetime with all linecards always powered on. Interestingly, for low values of K and W , the normalized lifetime is higher than 1, meaning that it has been increased compared to the “always on” case. This case is representative of devices that are able to sustain frequent power state transitions, and therefore their lifetime is improved when the SM state is set. Then K and W are increased when moving from left to right in the figure. In these cases, the normalized lifetime is slightly decreased, due to the fact that the gain from putting devices in SM is lower than in the previous case. Moreover, the transitions tend to deteriorate the lifetime. Clearly, the lifetime obtained by AFA is lower than or equal to OPT-B, due to the fact that the latter optimizes the lifetime jointly considering all the time slots, and not one by one as in AFA.

To provide more insight, we have considered a total period of time equal to 15 days. In this case,

³ An alternative to this step may be to shift the traffic from the device and then check the QoS constraint before entering in SM.

⁴ The same set of traffic matrices is repeated across different days.

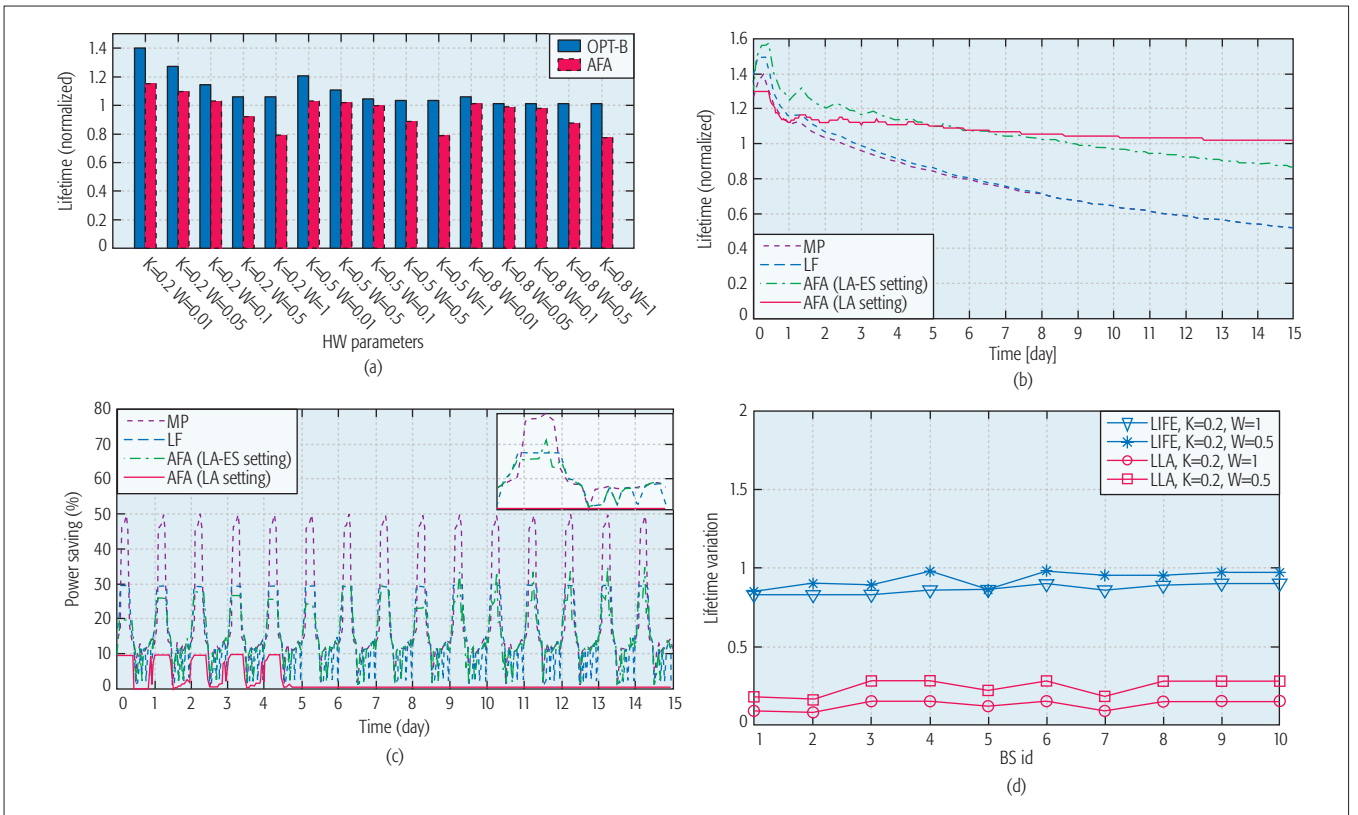


Figure 5. Results from backbone and cellular scenarios: a) backbone scenario: OPT-B and AFA normalized lifetime comparison for different values of HW parameters K and W ; b) backbone scenario: algorithms comparison in terms of normalized lifetime; c) backbone scenario: algorithms comparison in terms of power savings vs. time (the inset is the magnification during day 14); d) LTE cellular scenario: normalized lifetime variation for each BS for different values of HW parameters K and W .

we have set $K = 0.2$ and $K = 0.5$. In this way, the lifetime gain due to SM is high (low K), while at the same time there is a non-negligible cost paid when transitions are applied. Figures 5b and 5c report the comparison of AFA against two energy-aware algorithms, Least-Flow (LF) [15] and Most-Power (MP) [15]. Both these solutions do not consider the impact of SM on the failure rate. As a consequence, the average lifetime with LF and MP is decreased to nearly 40 percent at the end of the considered time period. On the contrary, AFA is able to limit this lifetime decrease, which can easily be controlled by proper setting of the lifetime thresholds. More in depth, with the lifetime-aware (LA) setting the set of candidate devices that may be put in SM is restricted only to those ones with lifetime higher than or equal to the one obtained by keeping the device powered on. Consequently, AFA stops putting most linecards in SM after day 4 in order to prevent their lifetime degradation. On the contrary, the lifetime-aware and energy-saving (LA-ES) setting refers to a case in which a maximum lifetime degradation equal to 50 percent is allowed. This in turn lets AFA trade between lifetime and energy.

In the following, we have focused on a cellular LTE scenario in which users (and their traffic) vary with a granularity of 1 hour over a 15-day period. We refer the reader to [12] for a detailed description of this scenario. In brief, we consider a hierarchical network in which one macro BS provides coverage over an area, while 10 micro BSs are placed in hotspots. In this area, users are

placed preferentially close to the micro BSs. Each user is associated with a BS assuming a best serving BS allocation policy (we adopt the Walfish-Ikegami model for urban zones to model the path loss). Users then request different types of services (i.e., voice, web browsing, or data service). The set of users is varied across the time slots (according to a typical day-night trend). The network is dimensioned to satisfy the peak traffic with all the micro BSs powered on. Table 5d reports the lifetime variation of LIFE against LLA [10] for different HW parameters K and W and for each BS in the scenario. As expected, LIFE outperforms LLA in terms of lifetime for all the BSs, being able to limit the lifetime decrease. In particular, the normalized lifetime for BS_1 with LIFE is equal to 0.83, while with LLA it is only 0.09. Thus, there is a lifetime decrease of more than 90 percent with LLA, while only 17 percent with LIFE.

Finally, Table 1 reports a comparison of the main features of the proposed lifetime-aware solutions compared to the energy-aware algorithms (LF, MP, and LLA). More in depth, the LIFETEL solutions are focused on lifetime-aware management of linecards and third/fourth generation (3G/4G) BSs. Clearly, the lifetime decrease is very high for the solutions that target solely the maximization of energy consumption. Our lifetime-aware algorithms, on the contrary, are able to manage the energy-lifetime trade-off. Focusing on the implementation aspects, all the proposed solutions are centralized (i.e., a central node computing the power

	Features/algorithms	OPT-B [11]	AFA [13]	LF/MP [15]	OPT-C [12]	LIFE [12]	LLA [10]	
Scope	Type	Lifetime-aware	Lifetime-aware	Energy-saving	Lifetime-aware	Lifetime-aware	Energy-saving	Scope
	Targeted devices	Linecards	Linecards	Routers, linecards	3G/4G BSs	3G/4G BSs	3G/4G BSs	
Lifetime-energy	Impact on lifetime decrease	Low	Low	High	Low	Low	High	Lifetime-Energy
	Energy savings	Medium	Medium	High	Medium	Medium	High	
Implementation aspects	Operation mode	Centralized	Centralized	Centralized/distributed	Centralized	Centralized	Centralized	Implementation Aspects
	Time slot	Multiple	Single	Single	Multiple	Single	Single	
	Future traffic	Known	Unknown	Unknown	Known	Unknown	Unknown	
Performance metrics	QoS evaluation	Max. link utilization	Max. link utilization	Max. link utilization	Coverage, user data rate	Coverage, user data rate	Coverage, user data rate	Performance Metrics
	Complexity	NP	$\mathcal{O}(LS + L \log L + (2L + 1)N^2)$	$\mathcal{O}(L \log L + LN^2)$	NP	$\mathcal{O}(NS + N \log N + NU(N + 1))$	$\mathcal{O}(N \log N + NU(N + 1))$	

Table #. Comparison of lifetime-aware and energy-saving solutions (N = number of BSs for LIFE,LLA and number of nodes for AFA,LF,MP, L = number of linecards for AFA,LF,MP, S = number of past time slots, U = number of users for LIFE,LLA).

state for the devices in a network is required). Moreover, the optimal formulations consider the set of time slots jointly at the same time, while LIFE and AFA are run for each time slot. Additionally, both LIFE and AFA do not assume knowledge of future traffic variations, being able to set the power states based only on past decisions and the traffic in the current time slot.

Finally, the last two rows of the table detail the main performance metrics. Focusing on backbone networks, the QoS is guaranteed by ensuring that the link utilization is lower than a maximum one (eventually scaled by an overprovisioning factor). The QoS for cellular networks is instead ensured by always guaranteeing connectivity over the service area, as well as guaranteeing the data rate requested by users. The second metric for algorithm comparison is the time complexity. Clearly, the complexity of lifetime-aware formulations OPT-B and OPT-C is non-polynomial due to the fact that these problems are in general NP-hard. Focusing on AFA and LIFE solutions, their complexity is similar to LF, MP, and LLA, except for the fact that the iteration over past time slots S is required in order to compute the lifetime for each device. However, in practice, this increase of complexity can be reduced, for example, by setting a limited amount of past time slots to compute the lifetime.

CHALLENGES FOR OPERATION NETWORKS

The results presented in LIFETEL pose different challenges that should be investigated for full exploitation of lifetime-aware approaches. In particular, topics that have to be examined more in depth are the following ones.

Measurement analysis of the HW parameters. LIFETEL demonstrated that the lifetime is a metric jointly depending on the adopted SM strategy and the HW parameters. To this end, a deep experimental measurement analysis of the

HW parameters on different network devices subject to SM is a mandatory next step. More in depth, this activity would allow a better estimation of the impact of SM-AM power transitions on a device's lifetime.

Validation of classical analytical models.

The models adopted in LIFETEL are derived from classical physical theories. Additionally, the validation of these models applied to real devices employed by a network operator has to be performed. In particular, we call for a characterization of failures in backbone and cellular networks, with an emphasis on the ones triggered by the application of SM.

Development of specific models.

The experimental activity described in the previous items should lead to the definition of new and more accurate analytical lifetime models of real devices. These models may be used to perform a fine tuning of SM policies in order to find a trade-off between energy saving and device failure rates.

Summarizing, LIFETEL is a step toward a more comprehensive approach aimed at investigating the problem of energy saving in a network from an economical point of view. In this scenario, the monetary saving derived from the application of SM should be compared to the increase of costs incurred for repairing or replacing the network devices [3] (due to the fact that devices fail more frequently compared to the case in which they are always in AM). Moreover, the costs for users due to the QoS decrease should also be considered.

We think that all these issues should be investigated by the research community.

CONCLUSIONS AND FUTURE WORKS

We have proposed a novel framework to manage the device lifetime in telecommunication networks subject to SM. We have shown that the transitions between AM and SM tend to increase the device

failure rate. As a consequence, energy-aware approaches solely targeting energy efficiency may negatively impact the device lifetime. We have therefore optimally formulated the problem of managing the lifetime for a backbone network and a cellular one. The optimal formulations take into account multiple time slots. Additionally, we have proposed the AFA and LIFE heuristics to efficiently solve the problem by considering the current time slot and past decisions. Results show that the energy-lifetime trade-off can be efficiently managed by the LIFETEL approach.

As a next step, we plan to consider the joint design and management of lifetime-aware networks (i.e., to decide where to place a device in a network and how to manage it). Additionally, we plan to perform measurements of lifetime on real devices to validate the proposed models.

ACKNOWLEDGMENTS

The research leading to these results has received funding from Sapienza Awards 2014–2015, Increasing the Lifetime of Telecommunication networks (LIFETEL). We would like to thank Paolo Monti, Lena Wosinska, Filip Idzikowski, William Liu, Jairo Gutierrez, Josip Lorincz, and Esther Le Rouzic for fruitful discussions and suggestions. We would like to thank the reviewers for their comments, which have helped us to improve the quality of this work.

REFERENCES

- [1] M. N. Dharmaweera, R. Parthiban, and Y. A. Şekercioğlu, "Towards A Power-Efficient Backbone Network: The State of Research," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 1, 2015, pp. 198–227.
- [2] J. Wu et al., "Energy-Efficient Base Stations Sleep Mode Techniques in Green Cellular Networks: A Survey," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 2, 2015, pp. 803–26.
- [3] P. Wiatr, P. Monti, and L. Wosinska, "Energy Efficiency Versus Reliability Performance in Optical Backbone Networks," *J. Optical Commun. and Net.*, vol. 7, no. 3, 2015, pp. A482–A491.
- [4] L. Chiaraviglio et al., "Is Green Networking Beneficial in Terms of Device Lifetime?," *IEEE Commun. Mag.*, vol. 53, no. 5, 2011, pp. 232–405.
- [5] S. Arrhenius, *Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren*, Wilhelm Engelmann, 1889.
- [6] L. F. Coffin Jr. and U.S. Atomic Energy Commission and General Electric Company, *A Study of the Effects of Cyclic Thermal Stresses on a Ductile Metal*, Knolls Atomic Power Laboratory, 1953.
- [7] S. S. Manson, "Behavior of Materials Under Conditions of Thermal Stress," *NACA Report 1170*, 1954.
- [8] N. El-Sayed et al., "Temperature Management in Data Centers: Why Some (Might) Like It Hot," *ACM SIGMETRICS Perf. Eval. Rev.*, vol. 40, no. 1, 2012, pp. 163–74.
- [9] A. Baiocchi et al., "A Simple Analytical Model for the BS Lifetime," *IEEE Commun. Letters*, vol. 19, no. 2, 2015, pp. 2206–09.
- [10] L. Chiaraviglio et al., "Energy-Efficient Planning and Management of Cellular Networks," *9th Annual Conf. Wireless On-Demand Network Systems & Services*, Courmayeur, Italy, Jan. 2012, pp. 159–66.
- [11] L. Amorosi et al., "Sleep to Stay Alive: Optimizing Reliability in Energy-efficient Backbone Networks," *11th Wksp. Reliability Issues in Next Generation Optical Networks*, Budapest, Hungary, July 2015, pp. 1–4.
- [12] L. Chiaraviglio et al., "Sleep to Stay Healthy: Managing the Lifetime of Energy-Efficient Cellular Networks," *Proc. IEEE GLOBECOM*, San Diego, CA, Dec. 2015, pp. 1–6.
- [13] L. Chiaraviglio et al., "Lifetime Awareness in Backbone Networks with Sleep Modes," *Proc. 3rd Int'l. Wksp. Understanding the Inter-Play between Sustainability, Resilience, and Robustness in Networks*, Munich, Germany, Oct. 2015, pp. 1–8.
- [14] F. Idzikowski, "TREND D3.3 Final Report for the IRA Energy-Efficient Use of Network Core Resources," <http://www.fp7-trend.eu/system/files/content-public/309-d33-final-report-ira-31/trend-d33.zip>," June 2012.
- [15] L. Chiaraviglio, M. Mellia, and F. Neri, "Minimizing ISP Network Energy Cost: Formulation and Solutions," *IEEE/ACM Trans. Net.*, vol. 20, no. 2, 2012, pp. 463–76.

BIOGRAPHIES

LUCA CHIARAVIGLIO [SM] (luca.chiaraviglio@gmail.com) is a tenure track assistant professor at the University of Rome Tor Vergata, Italy, and an adjunct assistant professor with the University of Cassino and Southern Lazio, Italy.

He has held different positions with the University of Rome Sapienza, Italy, CNIT, Italy, INRIA Sophia Antipolis, France, and the Polytechnic of Turin, Italy. He has been a visiting researcher with ETECSA SA, Cuba, Boston University, Massachusetts, and the Auckland University of Technology, New Zealand. He has co-authored over 85 papers published in top international journals and conferences. His main research interests are in the fields of 5G network architectures, 5G technologies for rural and low-income areas, sustainability in the ICT sector, and mobile network measurement and analysis. He was the Publication Chair of IEEE LANMAN in 2016. He was a recipient of the Best Paper Award at IEEE VTC-Spring 2016. He serves on the Editorial Boards of *IEEE Communications Magazine* and *IEEE Transactions on Green Communications and Networking*.

LAVINIA AMOROSI is a Ph.D student in operational research at the Department of Statistical Sciences at University of Rome Sapienza. She received her Master's degree in statistical and decision sciences from the University of Rome Sapienza in 2014. Currently she is a visiting Ph.D student at Lancaster University Management School, United Kingdom, under the supervision of Prof. Matthias Ehrgott. She is a member of the DIAMETER Awards Project. Her research area is combinatorial optimization, with particular interest in network optimization and multiobjective programming.

ANDREA BAIOCCHI received his Laurea degree in electronics engineering in 1987 and his Dottorato di Ricerca, (Ph.D. degree) in information and communications engineering in 1992, both from the University of Roma Sapienza. Since January 2005 he has been a full professor in the Department of Information Engineering, Electronics and Telecommunications of the University of Roma Sapienza. His main scientific contributions are in traffic modeling and traffic control in ATM and TCP/IP networks, queueing theory, wireless networks, and radio resource management. His current research interests are focused on congestion control for TCP/IP networks, wireless access interface protocols, specifically for VANET applications, and traffic analysis and monitoring. His research activities have also been carried out in the framework of many national (CNR, MIUR) and international (European Union, ESA) projects, also taking coordination and responsibility roles. He has published more than 140 papers in international journals and conference proceedings. He has participated in the Technical Program Committees of over 60 international conferences. He has also served on the Editorial Board of the *Telecommunications Technical Journal* published by Telecom Italia for 10 years.

ANTONIO CIANFRANI received his Master's degree in telecommunications engineering in 2004 and his Ph.D in information and communication engineering in 2008, both from the University of Rome Sapienza. He is an assistant professor at the DIET of the University of Rome Sapienza. His fields of interest include routing algorithms, network protocols, performance evaluation of software routers, and optical networks. His current research interests are focused on green networks and cloud networking.

FRANCESCA CUOMO [SM] received her Laurea degree in electrical and electronic engineering in 1993, magna cum laude, from the University of Rome Sapienza. She earned her Ph.D. degree in information and communications engineering in 1998 from the same university. Since 2005 she has been an associate professor at the University of Rome Sapienza and teaches courses in telecommunication networks. She has advised numerous Master's students in computer engineering, and has been the advisor of nine Ph.D. students in networking. Her current research interests focus on vehicular networks and sensor networks, cognitive radio networks, 4G and 5G systems, and energy saving in the Internet and wireless systems. She has participated in several national and European projects on wireless network systems. She has been on the Editorial Board of Elsevier's *Computer Networks* and is now a member of the Editorial Board of *Ad-Hoc Networks* (Elsevier). She has served on several Technical Program Committees. She has authored over 100 peer-reviewed papers published in prominent international journals and conferences.

PAOLO DELL'OLMO is a full professor in operations research at the Department of Statistical Sciences of the University of Rome Sapienza. He has been department head (2005-2011), coordinator of the Ph.D. program in operations research (2005-2011), and President of the Italian Inter-University Center for Operations Research (2010-2015); he is currently director of the Master's program in data intelligence and strategic decisions. His research interests are mainly in combinatorial optimization also applied to real-life network problems, in particular, computational complexity, design and analysis of exact and approximated algorithms, mathematical programming applied to traffic management, logistics, coordination of traffic flows on networks, scheduling, and routing. He has been Scientific Coordinator of a number of national research projects and is an author of several books and papers published on international journals.

MARCO LISTANTI is full professor in telecommunication networks at the DIET of University of Rome Sapienza. Since November 2013 he has been Dean of the Faculty of Information Engineering, Informatics and Statistics. In about 35 years of activity, his research interests have mainly been focused on the area of architectures and performance evaluation of telecommunication and computer networks. He has participated in national and international standardization commissions and working groups in the field of telecommunication networks (CCITT, ETSI, ITU, IETF, etc.). He has coordinated the activity of many national and international study and research projects in the framework of programs RACE, ACTS, ITS, COST, ESA, and MIUR. He is an author of more than 300 papers, published in the most important scientific magazines and journals and in the proceedings of the main international conferences in the area of telecommunications and computer networks.

Our results show that the energy-lifetime trade-off can be efficiently managed by the LIFETEL approach.. As next steps, we envision to perform the joint design and management of lifetime-aware networks (i.e., to decide where to place a device in a network and how to manage it) as well as a measurement campaign to validate the proposed lifetime models.

Dynamic Energy Trading for Wireless Powered Communication Networks

Yong Xiao, Dusit Niyato, Ping Wang, and Zhu Han

The authors provide an overview of the possible architecture and functional components that enable DET in communication networks. Various design issues on how to implement DET in practice are discussed. An optimal policy is proposed for delay-tolerant wireless powered communication networks.

ABSTRACT

Wireless powered communication systems have recently attracted significant interest because of their potential to provide a ubiquitous and sustainable energy supply for communication networks. However, the energy that can be harvested from external energy sources is generally uncontrollable and intermittent. By allowing multiple devices to exchange their harvested energy, dynamic energy trading (DET) is introduced to improve the energy supply reliability and performance of wireless powered communication networks. This article provides an overview of the possible architecture and functional components that enable DET in communication networks. Various design issues on how to implement DET in practice are discussed. An optimal policy is proposed for delay-tolerant wireless powered communication networks in which each wireless powered device can schedule its data transmission and energy trading operations according to current and future energy availability. Finally, some potential topics and challenges for future research are highlighted.

INTRODUCTION

By allowing electric devices to harvest energy from the natural environment (e.g., solar, wind, radio wave, and vibration), energy harvesting is a promising technology to provide ubiquitously available and green alternative energy sources for communication devices. However, the uncontrollability, uncertainty, and unpredictability of external energy sources in the natural environment make it difficult to provide a reliable power supply for communication networks. Recent works on wireless power transfer suggest that it is possible to allow a certain amount of energy to be transferred from dedicated energy sources to each device to support high-energy-consuming services. This significantly widens the possible applications for wireless powered communication systems. For example, Stanford University's Global Climate and Energy Project has recently reported that it is possible to deliver up to 10 kW of power for a distance of 6.5 ft with transfer efficiency of around 40 percent, which has the potential to be applied in parking lots or highways to wirelessly charge electric vehicles in the future [1].

Motivated by the observations that the combination of energy harvesting and wireless power

transfer can take advantage of both technologies to further improve system reliability and energy utilization efficiency, energy harvesting and wireless power transfer-enabled systems have attracted significant interest in the communication research community. For example, the concept of network-assisted energy harvesting has been studied in [2] for a wireless power supply network consisting of a set of dedicated energy stations that can wirelessly transfer electric power to mobile devices coexisting and telecommunication networks. Each mobile device can coordinate with each other by utilizing the telecommunication networks and obtain reliability guaranteed wireless power supply from its nearby energy stations.

In this article, we introduce a new concept, referred to as dynamic energy trading (DET), for wireless powered communication networks. DET allows multiple wireless powered devices (WPDs) with temporal and spatial variations in their energy harvesting processes to negotiate and exchange energy with each other. In DET, the WPDs that obtain more energy than they can consume can transfer their surplus energy to those who cannot receive sufficient energy to support the required services. Compared to the existing energy harvesting and wireless power transfer-enabled systems, DET possesses the following benefits:

- It is known that the energy harvested from the natural environment is intermittent and time-varying. DET allows WPDs with different harvestable energy to help each other, and hence can further improve the reliability of the power supply for wireless powered communication networks.
- Most existing works focus on the wireless power transferred from a dedicated energy source to a fixed energy receiver. DET allows different WPDs with different harvestable energy to dynamically exchange energy with each other. This can mitigate energy waste and increase the energy utilization efficiency for wireless powered systems without requiring investment in dedicated power supply network infrastructures.

ARCHITECTURE AND OPERATIONS FOR DET ARCHITECTURE OF DET

In this article, we focus on the energy trading among multiple WPDs. Each WPD can correspond to a hardware equipment that belongs to a part of permanently deployed infrastruc-

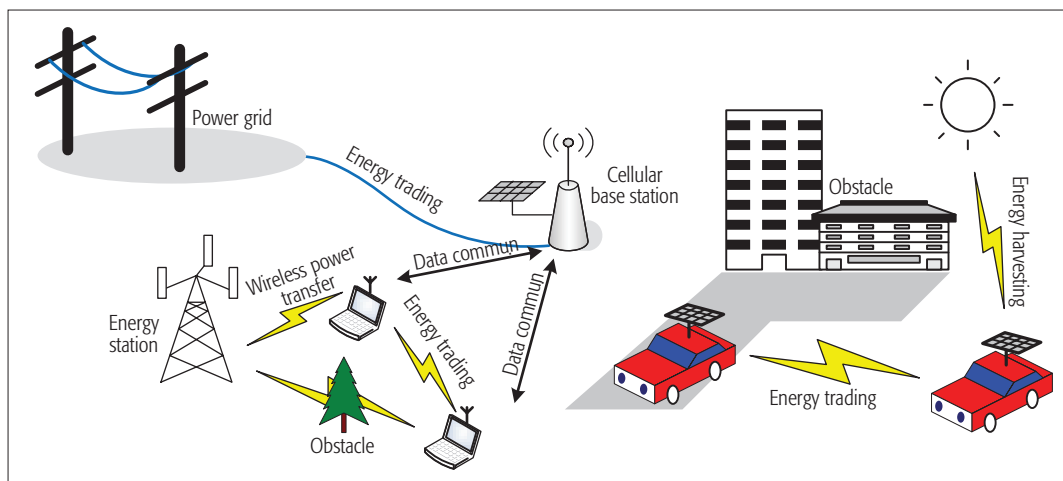


Figure 1. Dynamic energy trading in wireless powered communication networks: we illustrate three possible scenarios of energy trading: energy trading between a power grid and a cellular base station; energy trading between two wireless powered cellular mobile devices; and energy trading between two solar-powered electric vehicles.

DET has the potential to further improve the energy utilization efficiency by allowing the WPDs that can harvest more energy than they can use to sell their surplus energy to the WPDs that cannot harvest sufficient energy to support required services.

tures such as fixed energy stations (e.g., power beacons [3]). It can also be a portable device with energy transfer and receiving hardware. It is known that the energy that can be harvested by each WPD depends on various factors such as the location and orientation of energy harvesting equipments (e.g., antennas and solar panels), energy conversion efficiency, and distance to the external energy sources. Therefore, even closely located WPDs may harvest significantly different amounts of energy at the same time. DET has the potential to further improve energy utilization efficiency by allowing WPDs that can harvest more energy than they can use to sell their surplus energy to WPDs that cannot harvest sufficient energy to support required services. Depending on the energy delivery facilities, energy trading can be divided into two types.

Wired Energy Trading: WPDs are connected with the wired two-way power delivery infrastructure such as power grid and power transmission line, and can exchange and trade energy with each other [4, 5].

Wireless Energy Trading: In the case in which WPDs are not connected with each other through the wired power delivery infrastructure, they can exchange and trade their obtained energy through wireless power transfer.

In this article, we mainly focus on wireless energy trading. To simplify our description, we assume that the data communication, energy harvesting, and trading processes are time-slotted. In each time slot, the number of data packets received by each WPD, and energy that can be harvested and traded among WPDs are assumed to be fixed. To simplify our description, we normalize the duration of each time slot into unity and can therefore use the terms “energy” and “power” interchangeably. In each time slot, WPDs are divided into two types:

- **Energy suppliers** are WPDs that can provide controllable amounts of energy supply to other WPDs. Energy suppliers can be WPDs within a power delivery infrastructure such as the electrical generators

and power grid or dedicated energy sources deployed by network operators or utility companies. They can also be mobile WPDs with surplus energy that can be transferred to nearby WPDs.

- **Energy consumers** are WPDs that cannot obtain enough energy to support the required service without requesting a certain amount of energy to be transferred from the energy suppliers.

In DET, the sets of suppliers and consumers can change dynamically. Figure 1 illustrates several possible applications of DET in a wireless powered communication system.

The block diagrams of two WPDs corresponding to an energy supplier and an energy consumer in a wireless energy trading system are illustrated in Fig. 2. A WPD can have an *energy harvesting module* to convert external energy in the natural environment into electric current. The converted electric current will then charge an *energy storage module*, which can correspond to a (super) capacitor or rechargeable battery. If the WPDs also need to fulfill the required data communication service, they will include a *communication service module*, which consists of a data arriving queue to store arrived data packets. The *data transceiver module* provides data transmission and receiving functions for the communication service module to send the required data signals as well as the two-way negotiation and communication between suppliers and consumers during energy trading. Both the *energy transfer* and *receiving modules* include matching circuits, which can adjust the energy transfer and receiving parameters such as the transmit energy level, and transmission and receiving frequency. The *central processor module* plays a vital role in the energy trading process between suppliers and consumers, that is, it will decide which consumers or suppliers and how much energy to trade according to the energy level of its battery, harvestable energy, energy requested by the consumers, arrival rate of data packets, required quality of service (QoS), and other information such as

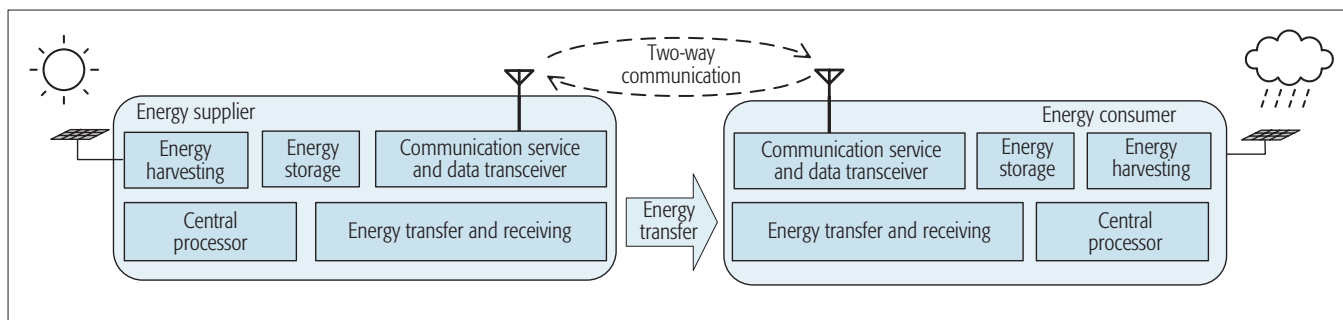


Figure 2. Diagram of WPDs in an energy-trading-enabled system.

knowledge about future changes in energy harvesting processes. Note that in some systems, the energy receiving module and energy harvesting module can be the same. For example, if WPDs have installed RF-based energy harvesting equipments, they can receive RF energy transferred from other WPDs using the energy harvesting module.

We use term *mode* to describe the decision made by each WPD about operating as the supplier or consumer in each time slot, and refer to the process for each WPD to decide its mode as *mode selection*.

ENERGY TRADING OPERATIONS

A DET process includes the following operations: Each WPD first decides its mode as supplier or consumer. Each supplier (or consumer) then tries to discover the identities and information about energy availability of its neighboring consumers (or suppliers). A two-way communication link can be established between each supplier and its neighboring consumers to negotiate details of energy trading. Once an agreement has been reached, the agreed amount of energy is transferred from the suppliers to the consumers. The possible energy trading operations are illustrated in Fig. 3. Let us give a more detailed discussion on each of these operations.

Mode Selection: As mentioned previously, the mode of each WPD can change dynamically. Therefore, it is important for each WPD to first decide its mode by evaluating the energy it can obtain as well as what should be spent on supporting the required service. Each WPD can also take into consideration the existing and future energy consumption of its installed modules. In [6], the authors studied a simple mode selection rule in which each WPD first decides its mode by comparing the current harvestable energy to the data transmission requirements. The WPD can then operate as a supplier if there is surplus energy after the required service has been fulfilled or as a consumer otherwise. In [7], each WPD selects its mode by also taking into consideration the evolution of the possible harvestable energy and energy that can be traded with other WPDs in the future.

Peer-Discovery and Coordination: Once each WPD has selected its mode, it then discovers the identity of suppliers or consumers in its surrounding area. The peer discovery approaches for the WPDs can be classified into *distributed discovery* and *network-assisted discovery*. In distributed discovery, each consumer (or supplier)

autonomously discovers the identity of its neighboring suppliers (or consumers). A simple peer discovery protocol was proposed in [7] in which each supplier broadcasts its available energy and unit price for energy transfer at the beginning of each time slot. Each consumer can then send its bid for the required amount of energy to its preferred suppliers. If a supplier accepts the request of the consumer, it starts transferring the requested amount of energy at the agreed time and frequency. Otherwise, the supplier simply ignores the energy requests of the consumers. The consumer updates its belief about whether the suppliers will accept its requests and only sends request to those suppliers that are highly likely to accept its energy request in the future. In the network-assisted discovery approach, each WPD discovers the nearby consumers or suppliers using the information provided by the network operator or the central controller.

Negotiation and Information Exchange:

Once an agreement has been reached between an energy supplier and a consumer, they form an energy trading pair. Consumers decide how much energy to request from suppliers based on the energy required to support their service and the cost for trading energy with each supplier. Each supplier can impose a price for each unit of energy transferred to consumers [8]. This will not only incentivize the suppliers to sell their surplus energy but could also avoid each consumer requesting unnecessarily large amounts of energy from the suppliers.

Energy Transfer: Once an agreement has been reached between suppliers and consumers, the suppliers start sending the energy with the agreed amount to the consumers. Since each energy supplier or consumer has unique properties with specific hardware requirements, the energy that can be transferred and successfully received depends on the specification of the installed power transfer and receiving hardware. For example, if the energy transfer between the suppliers and consumers has been achieved by wireless power transfer technologies, such as inductive coupling, RF energy transfer, or (strongly) coupled magnetic resonance, the energy loss during the wireless power transfer will be affected by the distance between suppliers and consumers, the energy transfer frequency, circuit design, antenna orientation, and so on.

DESIGN ISSUES FOR DET

In this section, we discuss the possible issues in designing an energy-trading-enabled system.

ENERGY TRANSFER AND USAGE SCHEDULING

Energy scheduling means that each WPD should “prepare for the future” by taking advantage of the knowledge about the future evolution of the natural environment. For example, a WPD can save some of its currently harvested energy for future use if there is a high chance that, during the next period of time, the harvested energy will be inadequate to support the required communication services.

The energy usage scheduling scheme and the resulting performance gain depend on knowledge about the future energy harvesting process at the WPDs. An optimal scheduling policy has been derived in [9] by assuming that the future change of the harvested energy has the Markov property, and the statistical feature of the transition between different levels of harvested energy can be perfectly known by each WPD. If the WPDs cannot know the probability distributions of the future evolution of the energy harvesting process, they can learn this information from past experience. In this case, there is a fundamental trade-off between how to take advantage of the knowledge that has already been learned by WPDs to maximize performance (exploitation) and how to explore new knowledge to further improve the energy scheduling gain (exploration). It has been shown in [10] that by applying the Bayesian reinforcement learning approach for each WPD to learn the statistical features of the energy harvesting process, the above trade-off can be solved by allowing each WPD to sequentially optimize its energy scheduling scheme to maximize its long-term performance.

INTERFERENCE MANAGEMENT

Interference has been regarded as one of the main factors that deteriorate the QoS for wireless communication services. However, interference can also be regarded as a potential energy source, which is especially beneficial for WPDs with RF energy harvesters [11]. A communication system powered by energy harvested from the ambient backscattered RF signals has been developed in [12] in which a prototype has been built to achieve 1 kb/s transmission rate over distances of 2.5 and 1.5 ft in outdoor and indoor environments, respectively.

SIMULTANEOUS WIRELESS INFORMATION AND POWER TRANSFER

In simultaneous wireless information and power transfer (SWIPT), the data communication signal can piggyback the energy signal sent to the WPDs [13]. SWIPT opens new opportunities to jointly analyze and optimize the wireless data communication and power transfer problems. Initial studies assume that both energy and information can be transferred using the same signal. Recent observations suggest that simply transferring energy and data signal simultaneously over the same frequency may result in intolerable interference to the data signal in most practical systems. How to efficiently split the energy and data signal during communication and achieve the optimal trade-off during signal transmission and wireless power transfer is one of the main challenges in SWIPT [3].

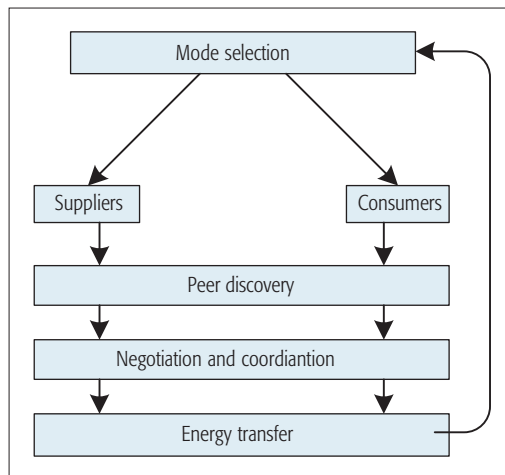


Figure 3. Energy trading operations.

ENERGY BEAMFORMING

It has been observed that if an energy supplier is deployed with multiple antennas, it can steer the energy transfer signal toward a specific direction. Energy beamforming can be further categorized into multiple-input multiple-output (MIMO) beamforming and distributed beamforming. In MIMO beamforming, the transmitter is installed with multiple antennas, and hence can change the angle of power transfer by adjusting the energy signals and power levels at each antenna [3]. In distributed beamforming, two or more energy stations can coordinate with each other to emulate an antenna array by transmitting a common energy signal in the direction of intended energy consumers. Distributed beamforming requires communication and coordination among multiple energy suppliers, which may result in energy transfer delay. It does, however, allow energy beamforming to be achieved for single-antenna WPDs.

ENERGY COOPERATION

It is known that both data and energy transfer signals deteriorate significantly with increase of transmission distance. To alleviate this problem, WPDs can cooperate with each other to relay data and/or energy signals for each other. Existing multihop relaying protocols such as amplify-and-forward and decode-and-forward have already been extended into energy relaying in [14]. Motivated by the fact that different relay nodes can result in different energy and data transfer efficiency, the relay selection problems were studied in [8].

AN OPTIMAL POLICY FOR A WIRELESS POWERED DELAY-TOLERANT COMMUNICATION NETWORK WITH DET

SYSTEM MODELS

Consider a communication network consisting of multiple WPDs, each of which is equipped with both energy transfer and receiving modules to exchange energy with others. At the beginning of each time slot t , each WPD i receives $\hat{u}_{i,t}$ data packets and knows the amount of energy $\hat{e}_{i,t}$ that can be harvested during the rest of time slot t .

Interference has been regarded as one of the main factors that deteriorate the QoS for wireless communication services. However, the interference can also be regarded as one of the potential energy sources, which is beneficial especially for the WPDs with RF energy harvesters.

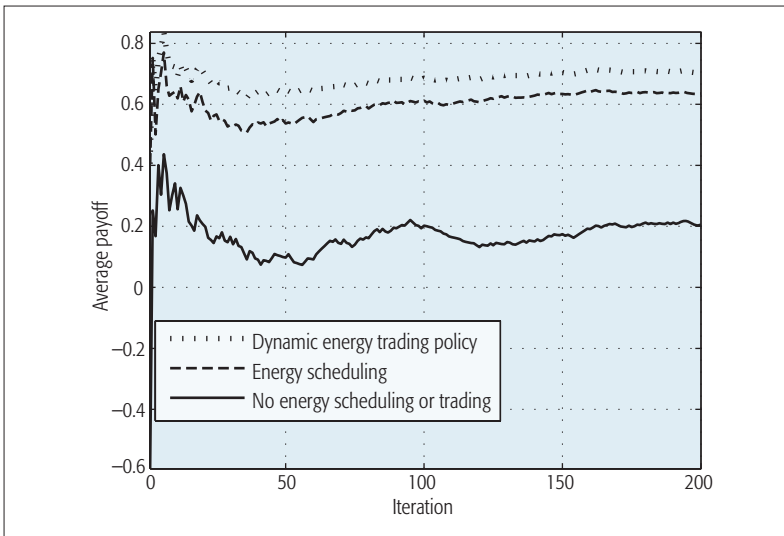


Figure 4. Comparison of the average payoff of a WPD with DET policy, energy scheduling, and no energy scheduling and trading under different iterations.

Each WPD i has a data buffer and a battery that can store no more than \bar{u}_i data packets and \bar{e}_i units of energy, respectively. We assume that the data transmission is delay-tolerant, and each WPD can intentionally delay the transmission of some data packets if it believes that it will obtain more energy and/or will have fewer data packets to transmit in the future. Due to the limit of the buffer size, if the number of newly arrived and buffered data packets exceeds the size of buffer, some of the data packets are dropped. In particular, the number of data packets that WPD i has to drop at the beginning of time slot t is given by $l_{i,t} = \max\{u_{i,t} + \hat{u}_{i,t} - \bar{u}_i, 0\}$ where $u_{i,t}$ is the buffer level of WPD i at the beginning of time slot t given by $u_{i,t} = \min\{u_{i,t-1} + \hat{u}_{i,t-1}, \bar{u}_i\} - v_{i,t-1}$, and $v_{i,t-1}$ is the number of data packets sent by WPD i during time slot $t-1$. The battery level of WPD i is given by $e_{i,t} = \max\{e_{i,t-1} + \hat{e}_{i,t-1} - w_{i,t-1}, \bar{e}_i\}$ where $w_{i,t-1}$ is the energy spent on transmitting data packets in time slot $t-1$.

Each WPD needs to send the received data packets to its corresponding destination with the required QoS. We assume that there is a one-to-one mapping function $f(\cdot)$ from the number of transmit data packets $v_{i,t}$ to the amount of energy $w_{i,t}$ that should be spent in sending $v_{i,t}$ data packets with the required QoS. Each WPD can receive reward $\alpha_{i,t}v_{i,t}$ by successfully sending $v_{i,t}$ data packets and will incur cost $\beta_{i,t}l_{i,t}$ for losing $l_{i,t}$ data packets in each time slot t where $\alpha_{i,t}$ and $\beta_{i,t}$ are the reward and cost of successfully sending and dropping each data packet, respectively. We assume that each WPD can always discover its nearby consumers and suppliers, and each consumer (or supplier) has already been assigned a supplier (or consumer).

Each WPD i will decide the following parameters at the beginning of each time slot:

- If WPD i chooses to operate as the supplier ($m_{i,t} = 0$), it decides how much energy $\Delta q_{i,t}$ is to be sent to consumers. We assume that WPD i can receive $\lambda_{i,t}\Delta q_{i,t}$ reward for selling $\Delta q_{i,t}$ energy units where $\lambda_{i,t}$ is the price for selling each unit of energy.

- If WPD i chooses to operate as a consumer ($m_{i,t} = 1$), it decides how much energy $\Delta q_{-i,t}$ to request from the supplier. In this case, WPD i pays $\rho_{i,t}\Delta q_{-i,t}$ to the supplier for transferring $\Delta q_{-i,t}$ energy units where $\rho_{i,t}$ is the price per unit of energy sent by the suppliers.

We can write the payoff of WPD i in time slot t as $\bar{\omega}_{i,t}(v_{i,t}, l_{i,t}, m_{i,t}, \Delta q_{i,t}, \Delta q_{-i,t}) = \alpha_{i,t}v_{i,t} - \beta_{i,t}l_{i,t} + (1 - m_{i,t})\lambda_{i,t}\Delta q_{i,t} - m_{i,t}\rho_{i,t}\Delta q_{-i,t}$.

AN OPTIMAL POLICY

We can formulate the decision making process for each WPD i in a DET system as a Markov decision process (MDP) with infinite horizon consisting of the following elements:

- *State space* \mathcal{S} is a finite set of the possible energy levels that can be harvested and the data packets required to be sent.
- *Action space* \mathcal{A} is a finite set of the possible number of transmit data packets, the mode selected by WPD i , and the possible energy that can be sent to the consumer if WPD i operates in supplier mode or possible energy that can be requested from the suppliers if WPD i operates in consumer mode.
- *State transition function* $\mathcal{T}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ specifies the probability distribution that, starting at state s_i using action a_i , the state ends in s'_i . This transition function can be estimated from the system model or obtained from past experience. In this article, we follow a commonly adopted assumption that the state transition function can be known by each WPD.

The main objective for each WPD is to find a decision policy π that maps the current state to action. We can therefore write the objective function of the joint optimization problem as follows:

$$\max_{\pi} \lim_{t \rightarrow \infty} \frac{1}{t} \mathbb{E}_{\pi} \left[\sum_{l=t}^{\infty} \bar{\omega}_{i,l}(a_{i,l}, s_{i,l}) \right]. \quad (1)$$

This optimal policy can be calculated numerically using the standard value iteration or policy iteration algorithms.

NUMERICAL RESULTS

We evaluate the performance of our proposed energy trading policy by considering a WPD that can harvest up to 10 mW of energy from the natural environment and send or request up to 10 mW to or from its nearby WPDs during each time slot. The battery of the WPD can store up to 20 mW. Energy transfer efficiency can be different with different wireless power transfer technologies. In this section, we assume that the energy transfer efficiency is 0.5. For example, it has been shown that a WPD is equipped with a self-resonant coil and can transfer energy with others using strongly coupled magnetic resonances at a distance up to 180 cm [15]. The WPD can transmit up to 20 mW of power during each time slot. We assume that the minimum amount of energy harvested and requested as well as transferred is 1 mW, and the minimum amount of energy required to send each data packet is 0.5 mW.

We compare the average payoffs achieved by our policy and the optimal energy schedul-

ing approach [9] without energy trading in Fig. 4. If the WPD uses neither energy scheduling nor trading, it will simply transmit signals with the harvested energy. We can observe that our proposed policy jointly optimizes the mode selection, energy scheduling, and trading, and hence achieves significant performance improvement compared to the traditional energy harvesting system without DET. In Figs. 5 and 6, we compare the probability for each WPD to decide to operate as the supplier and the average data packet loss under different buffer and battery sizes. We can observe that DET significantly improves the reliability of the data transmission, especially when each WPD can have the large battery and data buffer sizes.

CHALLENGES AND FUTURE RESEARCH TOPICS

DET opens many promising new applications in the future development of wireless powered communication networks.

POTENTIAL RESEARCH TOPICS FOR ENERGY SOURCE DISCOVERY

Future networks will consist of high densities of WPDs coexisting with various types of network infrastructures including cellular base stations (BSs) and Wi-Fi access points. It is important to develop an efficient protocol for each WPD to quickly discover the identities of neighboring suppliers or consumers with the assistance of the network infrastructure. For example, each WPD can report its identity and mode to its closest cellular BS. Each BS can then assign each consumer with the appropriate supplier and inform the pairing results to other WPDs within its coverage area. In addition, the network infrastructure can also help to regulate the energy trading among WPDs. For example, cellular BSs can broadcast a warning signal in the frequency bands before transmitting data signals. Each supplier should stop sending energy signals whenever it receives warning signals sent by BSs.

POTENTIAL RESEARCH TOPICS FOR ENERGY SCHEDULING

It is known that if WPDs can have perfect knowledge about the future evolution of the energy harvesting process, it can further improve its performance by optimally scheduling its energy usage. However, in practical systems, it is generally impossible to always accurately predict the statistical features of the future energy harvesting and transfer processes. It is therefore important to develop a unified framework that can characterize the relationship between the accuracy of the prediction and the performance gain achieved by the energy scheduling.

POTENTIAL RESEARCH TOPICS FOR ENERGY TRANSFER AND OFFLOADING

The total amount of harvestable energy within a given time duration is always limited. It is possible that the total amount of energy requested by consumers exceeds the limit. A fairness criteria and mechanism should be designed to properly and fairly divide the energy among consumers according to the amount of energy requested by each consumer, energy trading costs, and hardware specification of each supplier.

In addition, some suppliers will be overloaded

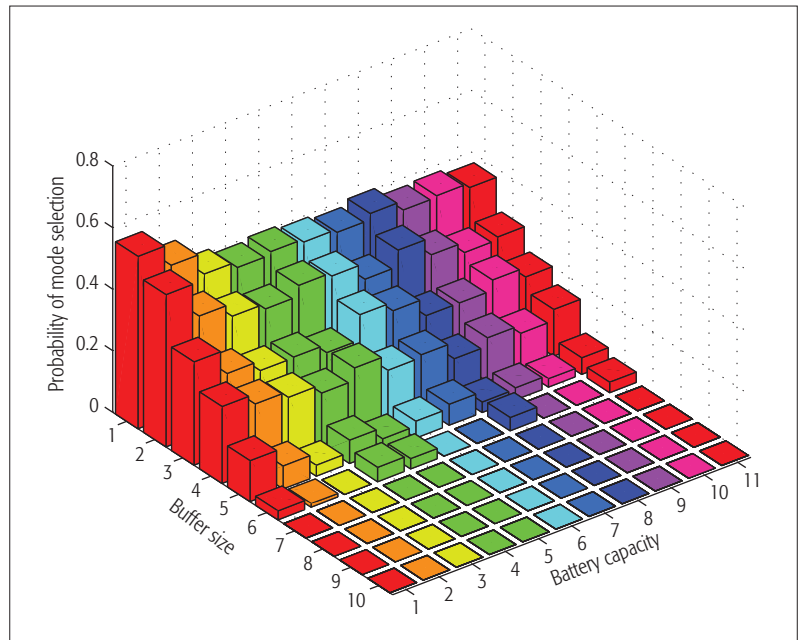


Figure 5. Probability of WPD i to operate in supplier mode.

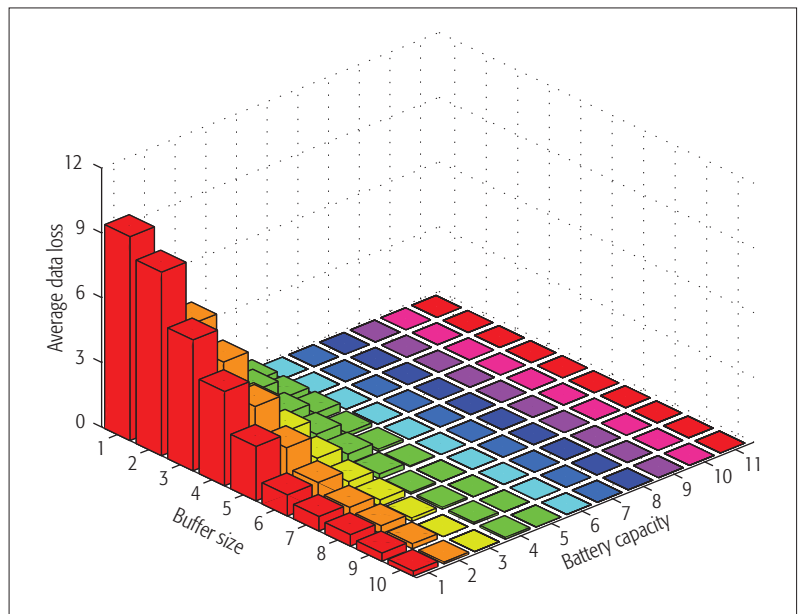


Figure 6. Average number of data packets lost.

if the number of consumers requesting energy transfer exceeds their maximum limits. Therefore, how to propose a simple and distributed mechanism to offload the energy requests of some consumers to other nearby suppliers is an interesting topic worth further investigation.

POTENTIAL RESEARCH TOPICS FOR COST EVALUATION AND PRICING MECHANISM

The data transmission requirement and harvestable energy of each WPD can change dynamically. Therefore, different WPDs will have different requirements for energy trading during different time slots. It is important for each WPD to properly evaluate its benefits and costs before trading energy with other WPDs. One possible solution is to introduce a virtual currency among energy

In this case, each supplier should decide a proper price for its transferable energy and broadcast the price to the potential consumers at the beginning of each time slot. Each consumer should then evaluate the price broadcasted by the suppliers and choose the suppliers with the most affordable price to purchase energy.

trading WPDs. In this case, each supplier should decide a proper price for its transferable energy and broadcast the price to the potential consumers at the beginning of each time slot. Each consumer should then evaluate the price broadcasted by the suppliers and choose the suppliers with the most affordable price from which to purchase energy. How to design an efficient pricing mechanism that can incentivize the energy trading among WPDs and avoid some WPDs to benefit from cheating their prices is still an open problem.

CONCLUSION

This article has presented an overview of DET and its possible implementations into the paradigm of wireless powered communication systems. We have introduced a general architecture and described the potential functional modules that enable energy trading in network systems. The design issues in implementing DET in practical systems have also been discussed. We have studied a delay-tolerant wireless powered communication system as an example to demonstrate how to optimize the energy trading in communication networks. An optimal policy has been developed for each WPD to sequentially decide its mode, when to transmit data packets, and the amount of energy traded with others. We have presented numerical results to justify the performance improvement that can be brought by DET and discussed future research topics. Currently, both energy harvesting and wireless power transfer are still in the early stages of development. This article can serve as a stepping stone for a wider range of research on the future generation of environmental-friendly wireless powered communication networking systems.

REFERENCES

- [1] X. Yu *et al.*, "Wireless Power Transfer in the Presence of Metallic Plates: Experimental Results," *AIP Advances*, vol. 3, no. 6, 2013, p. 062102.
- [2] K. Huang and V. Lau, "Enabling Wireless Power Transfer in Cellular Networks: Architecture, Modeling and Deployment," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, Feb. 2014, pp. 902–12.
- [3] R. Zhang and C. K. Ho, "MIMO Broadcasting for Simultaneous Wireless Information and Power Transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, May 2013, pp. 1989–2001.
- [4] J. Xu and R. Zhang, "Cooperative Energy Trading in CoMP Systems Powered by Smart Grids," *IEEE GLOBECOM*, Austin, TX, Dec. 2014, pp. 2697–2702.
- [5] J. Xu, L. Duan, and R. Zhang, "Cost-Aware Green Cellular Networks with Energy and Communication Cooperation," *IEEE Commun. Mag.*, vol. 53, no. 5, May 2015, pp. 257–63.
- [6] T. Jiang, G. V. Merrett, and N. R. Harris, "Opportunistic Energy Trading between Co-Located Energy-Harvesting Wireless Sensor Networks," *Proc. 1st Int'l. Wksp. Energy Neutral Sensing Systems*, Rome, Italy, Nov. 2013.
- [7] Y. Xiao *et al.*, "Dynamic Energy Trading for Energy Harvesting Communication Networks: A Stochastic Energy Trading Game," *IEEE JSAC*, Special Issue on Green Communications and Networking, vol. 33, no. 12, Dec. 2015, pp. 2718–34.
- [8] Y. Xiao, Z. Han, and L. A. DaSilva, "Opportunistic Relay Selection for Cooperative Energy Harvesting Communication Networks," *IEEE GLOBECOM*, Austin, TX, Dec. 2014.

- [9] A. Aprem, C. Murthy, and N. Mehta, "Transmit Power Control Policies for Energy Harvesting Sensors with Retransmissions," *IEEE J. Selected Topics in Signal Processing*, vol. 7, no. 5, Oct. 2013, pp. 895–906.
- [10] Y. Xiao, Z. Han, D. Niyato, and C. Yuen, "Opportunistic Relay Selection for Cooperative Energy Harvesting Communication Networks," *IEEE ICC*, London, U.K., June 2015.
- [11] V. Talla *et al.*, "Powering the Next Billion Devices with Wi-Fi," *11th ACM Int'l. Conf. Emerging Networking Experiments and Technologies*, Heidelberg, Germany, Dec. 2015, pp. 1–13.
- [12] V. Liu *et al.*, "Ambient Backscatter: Wireless Communication Out of Thin Air," *ACM Sigcomm*, Hong Kong, China, Aug. 2015.
- [13] I. Krikidis *et al.*, "Simultaneous Wireless Information and Power Transfer in Modern Communication Systems," *IEEE Commun. Mag.*, vol. 52, no. 11, Nov. 2014, pp. 104–10.
- [14] M. Tacca, P. Monti, and A. Fumagalli, "Cooperative and Reliable ARQ Protocols for Energy Harvesting Wireless Sensor Nodes," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, July 2007, pp. 2519–29.
- [15] A. Kurs *et al.*, "Wireless Power Transfer Via Strongly Coupled Magnetic Resonances," *Science*, vol. 317, no. 5834, July 2007, pp. 83–86.

BIOGRAPHIES

YONG XIAO [S'09, M'13, SM'15] is currently a research assistant professor in the Department of Electrical and Computer Engineering at the University of Arizona. He is also the center manager of the NSF BWAC Center at the University of Arizona. He received his B.S. degree in electrical engineering from China University of Geosciences, Wuhan, in 2002, his M.Sc. degree in telecommunication from Hong Kong University of Science and Technology in 2006, and his Ph. D degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2012. His research interests include machine learning, game theory, and their applications in wireless networks.

DUSIT NIYATO [M'09, SM'15] is currently an assistant professor in the School of Computer Engineering, Nanyang Technological University. He obtained his Bachelor of Engineering in computer engineering from King Mongkut's Institute of Technology Ladkrabang (KMUTL), Bangkok, Thailand. He received his Ph.D. in electrical and computer engineering from the University of Manitoba, Canada. His research interests are in the areas of radio resource management in cognitive radio networks and broadband wireless access networks. He has several research awards to his credit, which include the 7th IEEE Communications Society Asia Pacific Young Researcher Award, the IEEE WCNC 2012 Best Paper Award, the IEEE ICC 2011 Best Paper Award, and the 2011 IEEE ComSoc Fred W. Ellersick Prize paper award. Currently he serves as an Editor for *IEEE Transactions on Wireless Communications* and *IEEE Wireless Communications Letters*.

PING WANG [M'08, SM'15] received a Ph.D. degree in electrical engineering from the University of Waterloo, Canada, in 2008. She is currently an associate professor in the School of Computer Engineering, Nanyang Technological University. Her current research interests mainly focus on resource allocation in multimedia wireless networks. She was a coreipient of the Best Paper Award from IEEE WCNC 2012 and IEEE ICC 2007. She is an Editor of *IEEE Transactions on Wireless Communications*, the *EURASIP Journal on Wireless Communications and Networking*, and the *International Journal of Ultra Wideband Communications and Systems*.

ZHU HAN (S'01-M'04-SM'09-F'14) received his B.S. degree in electronic engineering from Tsinghua University in 1997, and his M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1999 and 2003, respectively. From 2000 to 2002, he was an R&D engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a research associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a professor in the Electrical and Computer Engineering Department as well as the Computer Science Department at the University of Houston, Texas. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, wireless multimedia, security, and smart grid communication. He received an NSF Career Award in 2010, the Fred W. Ellersick Prize of IEEE ComSoc in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, and several best paper awards at IEEE conferences, and is currently an IEEE ComSoc Distinguished Lecturer.

CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE
EDUCATION AND TRAINING: TELECOMMUNICATION STANDARDS EDUCATION

BACKGROUND

Technical standards are formal documents that establish uniform criteria, methods, and practices through accredited and consensus processes in numerous areas of engineering, science, and technology. They are also catalysts for technological innovation and global market competition. Standards and standardization processes are not traditionally incorporated into STEM university curricula. In most cases, where attempts are made to teach standards, traditional instruction methods are used. Since these are not effective with standards content, they result in limited impact, and do not help elevate interest in standards' education. This situation results in a knowledge gap among STEM graduates and professionals. And this gap is impacting the ability of our workforce to face emerging global challenges.

Telecommunication is a field of substantial standardization activities. Although standards constitute cornerstone of the Telecom Industry and business, university students in telecom-related disciplines usually experience only little knowledge, if any, about standards. Several efforts are underway today to deal with this shortcoming. In academic STEM education, which is undergoing re-evaluation in the US, there are efforts to integrate standards education into university curricula. There are also efforts to explore innovative instruction methods which can improve STEM learning, and standards education can benefit from these efforts. On the other hand, the IEEE (including IEEE-SA, IEEE-EA, and societies), ITU, and other professional societies and Standards Development Organizations (SDOs) are introducing education programs and resources to help professionals and academicians to overcome the lack of knowledge of standards.

This feature topic on "Telecommunication Standards Education" will discuss the challenges facing standards education in this discipline, analyze the reasons why university curricula lack sufficient coverage of standards and standardization topics, propose methods to promote and integrate knowledge of standards into engineering and STEM programs, propose innovative and effective instruction methods, and explore avenues for industry and academia to work together in this regard. This special topic is intended to provide an opportunity for educators and standards professionals to share their experience, best practices, and case studies.

SCOPE OF SUBMISSIONS

Authors from industry, academia, and government are invited to submit papers for this FT (Feature Topic) of IEEE Communications Magazine on Telecommunication Standards Education. The FT scope includes, but is not limited to, the following:

- Case studies of the incorporation of standards into telecommunications engineering curricula.
- Best practices for incorporating standards into telecommunication curricula.
- Case studies of the incorporation of knowledge of telecommunication standards into professional training.
- Best practices for incorporating telecommunication standards education into professional training.
- Development of innovative instruction strategies and tools for use in standards education
- Impact of standards education on engineering, technology, society, and economy

SUBMISSIONS

Articles should be tutorial in nature, with the intended audience being all members of the communications technology communities. They should be written in a style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words. Figures and tables should be limited to a combined total of six. The number of archival references is not to exceed 15. Complete guidelines for manuscript preparation can be found via the link <http://www.comsoc.org/commag/paper-submission-guidelines>. Please send a PDF (preferred) or MS-Word formatted paper via Manuscript Central (<http://mc.manuscriptcentral.com/commag-ieee>). Register or log in, and go to the Author Center. Follow the instructions there. Select "May 2017/Telecommunication Standards Education."

IMPORTANT DATES

- Manuscript Submission Deadline: December 1, 2016
- Decision Notification: January 15, 2017
- Final Manuscript Due Date: February 15, 2017
- FT Publication Date: May 2017

GUEST EDITORS

Tarek El-Bawab (Lead)
Jackson State University, USA
telbawab@ieee.org

David G. Michelson
University of British Columbia, Canada
davem@ece.ubc.ca

Periklis Chatzimisios
Alexander Technological Educational Institute
of Thessaloniki, Greece
peris@it.teithe.gr

Power-Saving Methods for Internet of Things over Converged Fiber-Wireless Access Networks

Dung Pham Van, Bhaskar Prasad Rimal, Jijia Chen, Paolo Monti, Lena Wosinska, and Martin Maier

The authors leverage converged fiber-wireless (FiWi) access networks to design a shared communication infrastructure for supporting both IoT applications and traditional services. Given the paramount importance of energy efficiency in both IoT and access networks, they discuss the possibilities and potential challenges of designing and implementing power-saving mechanisms to prolong battery life of IoT devices while reducing energy consumption of the optical backhaul network.

ABSTRACT

The IoT has been emerging as the next big leap in the information and communication technology sector. Providing a unified communication platform to support billions of smart connected devices seamlessly alongside existing voice and Internet services is vitally challenging. This article leverages converged fiber-wireless (FiWi) access networks to design a shared communication infrastructure for supporting both IoT applications and traditional services. Given the paramount importance of energy efficiency in both IoT and access networks, the article discusses the possibilities and potential challenges of designing and implementing power-saving mechanisms to prolong battery life of IoT devices while reducing energy consumption of the optical backhaul network. In-depth technical guidelines are provided through end-to-end power-saving solutions proposed for typical IoT deployment scenarios.

INTRODUCTION

While the *wired and mobile Internet* revolutionized the telecommunication paradigm by connecting people “anywhere” at “any time,” the emerging Internet of Things (IoT) is creating another paradigm in which “anything” can be remotely accessed and/or controlled, allowing for more direct integration between the physical world and machine-based systems. Unlike the traditional Internet for delivering human-centric services (e.g., file sharing, voice telephony, video streams), the IoT relies on machine-to-machine (M2M) communications with a focus on smart devices such as sensors, actuators, wearables, and metering devices. The IoT has its application in numerous and diversified fields, from connected vehicles to smart grids, spanning industries of utilities, healthcare, and transportation, just to name a few [1]. Therefore, the IoT is expected to become a new driving force of the information, communications, and technology (ICT) industry with about 50 billion devices connected to the Internet in 5 years [2]. Recent research has revealed that the IoT will potentially bring an economic impact of around 11 percent of the world economy in 2025 [3].

Despite enormous opportunities, the advent of the IoT alongside its integration with existing wired and mobile Internet creates huge challenges in designing a unified communication platform to fully support a wide range of IoT applications with diverse requirements and human-centric services at the same time [4]. Figure 1 presents key challenges of the paradigm of converged Internet technologies including network integration, energy efficiency, coexistence, diversity, and scalability. Network integration deals with the efficient merging of various types of networks, for example, converged wired and wireless access networks and collaboration among different network operators. Another important challenge is to handle the coexistence of conventional human-to-human (H2H) traffic, such as triple-play (voice, video, and data), and emerging machine-to-machine (M2M) traffic in order to ensure that high-priority traffic is not jeopardized. Note that coexistence is an inherent issue due to the integration of IoT devices into existing access network infrastructure.

The tremendous growth of mobile data traffic together with the increasing integration of radio access technologies (RATs) and the cell densification paving the way to 5G networks gradually shifts the bottleneck from the radio interface toward the backhaul segment. With the emerging IoT, the *backhaul bottleneck* is expected to become even more critical. However, until recently, existing studies on IoT connectivity largely focused on enhancements of RATs without looking into the backhaul segment [5]. Meanwhile, economic considerations may play an important role in the successful rollout of IoT, as experienced in smart grids [6]. To address the backhaul bottleneck in a cost-efficient way, a prominent solution is to share the already widely deployed high-capacity and reliable optical access network (OAN) infrastructure, which was originally for fixed broadband access [7]. This can be facilitated by converged fiber-wireless (FiWi) access networks that are able to seamlessly integrate an OAN and a multi-RAT front-end network in support of both human-centric and IoT applications on the same infrastructure. The integration of IoT and FiWi networks gives rise to so-called *IoT over FiWi networks*.

The work of D. Pham Van was partially supported by the EIT Digital project EXAM (Energy-Efficient Xhaul and M2M) and the Göran Gustafssons Stiftelse Foundation. The work of B. P. Rimal was supported by the Fonds de recherche du Québec – Nature et Technologies (FRQNT) MERIT Doctoral Research Scholarship Program.

Digital Object Identifier:
10.1109/MCOM.2016.1500635CM

Dung Pham Van, Jijia Chen, Paolo Monti, and Lena Wosinska are with KTH Royal Institute of Technology; Bhaskar Prasad Rimal and Martin Maier are with INRS.

Although the envisioned IoT over FiWi networks hold great promise to address some of the aforementioned challenges faced by the convergence of Internet technologies, energy efficiency is still a significant problem open for investigation, particularly for battery-constrained devices. While one could recharge his/her mobile phone on a daily basis, recharging or replacing batteries for billions of ubiquitous IoT devices is impractical or even infeasible in many use cases such as power grid transmission line monitoring. This leads to more stringent requirements on battery life (e.g., several years for temperature sensors). Meanwhile, optical backhaul and wireless/cellular networks are not energy-efficient since their low average utilization results in a huge waste of energy [8]. Therefore, end-to-end power-saving mechanisms considering both fiber and wireless segments are highly needed for IoT over FiWi networks.

This article investigates power-saving solutions to prolong the battery life of IoT devices while improving the energy efficiency of the backhaul in IoT over FiWi networks, thus reducing network operational expenditures (OPEX) as well as mitigating carbon emissions. The article first describes the envisioned network architecture and overviews existing power-saving modes that are defined in relevant industry standards. Note that power-saving methods aim to reduce energy consumption by turning off network devices when they are idle or operate at low loads through control protocols and scheduling algorithms without requiring major hardware modifications (i.e., targeting the data link layer). While most existing solutions consider either the optical backhaul or the wireless/cellular front-end segment [9], it is desirable to have end-to-end power-saving solutions that jointly consider both segments for truly energy-efficient IoT over FiWi networks. The challenges for designing and implementing such end-to-end power-saving mechanisms are then identified followed by in-depth design guidelines. Given that wireless local area networks (WLANs) and fourth generation (4G) Long Term Evolution (LTE) are widely considered for front-end IoT connectivity, while Ethernet-based passive optical networks (i.e., EPON and 10G-EPON) are cost-effective and simple OAN technologies, two case studies, IoT over EPON-WiFi and IoT over 10G-EPON-LTE, are investigated in greater detail. To the best of the authors' knowledge, the power-saving solutions presented in this work are the first that address the energy efficiency challenge faced by IoT networks taking network integration, scalability, and H2H/M2M coexistence into account.

IoT over FiWi Access Networks

This section overviews key wireless technologies widely considered for IoT connectivity and describes the introduced IoT over FiWi network architecture.

Wireless Front-End Technologies for IoT Connectivity

A plethora of RATs has been emerging for IoT connectivity in the wireless front-end segment including IEEE 802.15.4x, IPv6 over low-power wireless personal area networks (6LoWPAN),

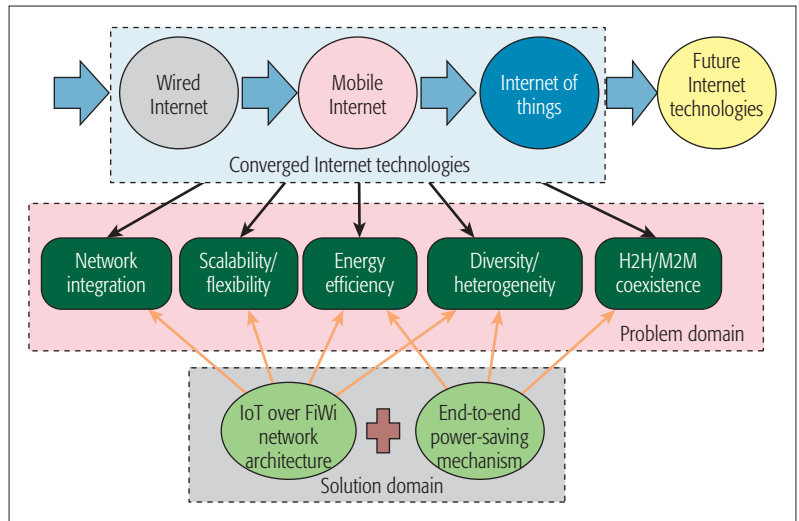


Figure 1. Overview of key challenges in integrated IoT over FiWi access networks.

low-power wide area (LPWA), Bluetooth Smart, WiFi, and cellular networks. Among those, WiFi and 4G LTE with various enhancements are considered the most promising options [1].

WiFi (IEEE 802.11) technology is usually suitable for short-range IoT applications such as smart home and industrial automation. However, the IEEE 802.11 community recently started to apply duty cycling and hardware optimization, resulting in the establishment of the IEEE 802.11ah project Low-Power WiFi, which targets simple, low-cost, low-power, and long-range (up to 1 km) M2M connectivity. Another advantage of IEEE 802.11ah is its easy integration into existing infrastructures with built-in IP compatibility [10]. Nevertheless, WiFi technologies suffer from some fundamental limitations including the lack of efficient backhaul links, which limits network scalability and coverage [11].

Due to its wide coverage, relatively low deployment costs, and access to dedicated spectrum, 4G LTE is considered another attractive solution for IoT connectivity, especially for large-scale IoT applications such as connected vehicles and remote health monitoring. However, cellular networks have been neither historically designed with link budget requirements of IoT devices nor optimized for machine-based IoT traffic. Therefore, recently, the Third Generation Partnership Project (3GPP) has initiated activities to augment LTE for supporting IoT applications with various enhancements [11].

Network Architecture

The IoT over FiWi network architecture aims to cover not only the front-end but also the backhaul in an attempt to provide a comprehensive solution. FiWi networks are realized by integrating wireless/cellular access technologies with OANs. There are two major types of FiWi networks, based on either radio-and-fiber (R&F) or radio-over-fiber (RoF) technologies. The R&F and RoF technologies may be used separately from each other or jointly. RoF networks use optical fiber as an analog transmission medium between remote antenna units and a central station that is in charge of controlling access to

In IoT over FiWi networks, since optical fiber infrastructures are shared for fixed access, mobile backhaul, and IoT connectivity, the number of involved network equipments is minimized. As a result, the network architecture helps reduce the total deployment costs and energy consumption.

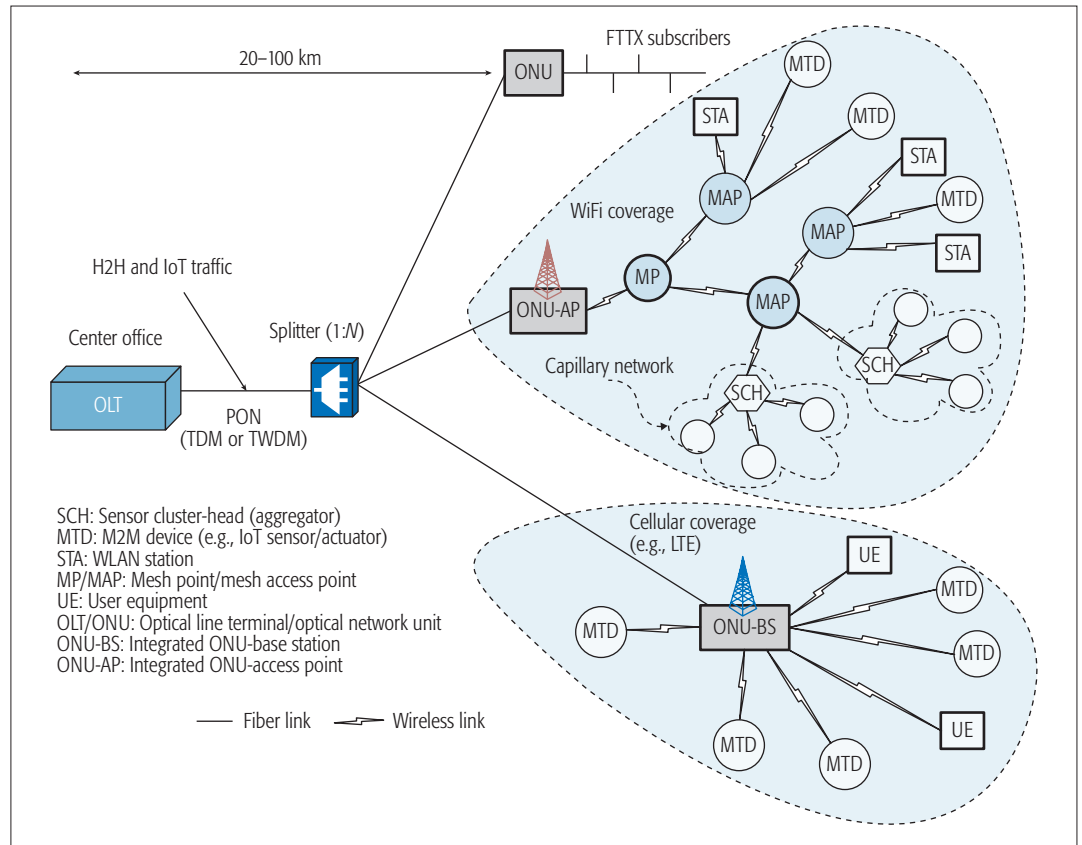


Figure 2. IoT over FiWi network architecture.

both optical and wireless media in a centralized manner. Conversely, in R&F networks, access to the optical and wireless media is controlled in a decentralized manner by using two different medium access control (MAC) protocols in the optical and wireless media with protocol translation taking place at their interfaces. Given that decentralization is an important aspect of the 5G vision (e.g., via local content caching or mobile-edge computing), R&F-based FiWi networks are likely to become the future FiWi type of choice in support of IoT connectivity in the 5G era [12].

Figure 2 depicts the IoT over FiWi network architecture as a shared communication platform for both H2H services and IoT connectivity. The OAN backhaul consists of an optical line terminal (OLT) located at the central office that serves optical network units (ONUs) located at the customer premises. In Fig. 2, a time-division multiplexed/multiple access (TDM/TDMA) PON-based backhaul is employed with its typical tree-and-branch topology. A subset of ONUs may be located at the premises of residential or business subscribers, providing FTTx services (e.g., fiber-to-the-home/curb/building) to a single or multiple wired subscribers. The second subset of ONUs is equipped with a mesh portal point (MPP) to interface with the WiFi mesh network consisting of decentralized mesh points (MPs) and mesh access points (MAPs), each serving mobile users within their limited coverage area. The collocated ONU-APs are realized by using R&F technologies. Conversely, an ONU of the third subset connects to a cellular base station (BS) with a given coverage area. The BS might be a conventional macrocell BS, for example,

an eNB in LTE or a small cell in LTE-Advanced heterogeneous networks (HetNets), or even future 5G BS. The collocated ONU-BS may rely on either the decentralized R&F or centralized RoF technologies. A typical example for the latter case is the cloud radio access network (C-RAN) [12].

IoT devices connect to the FiWi infrastructure the same way as conventional front-end user devices to exchange their sensing and monitoring data. Typically, there are two types of communication for IoT smart devices in the envisioned network. The IoT devices can be directly connected to wireless APs/cellular BSs (i.e., the direct access method). Furthermore, they can communicate with an aggregator/gateway node to form a local cluster of IoT nodes (i.e., the capillary network in Fig. 2), where a cluster head is responsible for communicating with the integrated ONU-AP and scheduling transmissions for cluster members. This is called the indirect access method. Note that the heterogeneous multi-RAT front-end in the introduced architecture provides IoT devices with more possibilities to be connected as well as a larger coverage area. However, depending on the capabilities of a device and its compatibility with the RAT(s) to which it can be connected, an appropriate access method is chosen. For example, a detailed review of random access mechanisms for IoT connectivity is provided in [13].

In IoT over FiWi networks, since optical fiber infrastructures are shared for fixed access, mobile backhaul, and IoT connectivity, the number of involved network equipments is minimized. As a result, the network architecture helps reduce

the total deployment costs and energy consumption. Moreover, to address the scalability issue imposed by not only the tremendous growth in traffic volume but also a massive number of devices in both H2H and M2M applications, future IoT over FiWi networks may rely on next-generation OAN technologies, such as time and wavelength-division multiplexing (TWDM) PON, which is capable of providing up to 40 Gb/s.

ENERGY EFFICIENCY IN IoT OVER FiWi NETWORKS

This section first reviews existing power-saving modes defined in relevant industry standards. It then identifies technical challenges for implementing end-to-end power-saving mechanisms in IoT over FiWi networks. Design guidelines and solutions for two typical deployment scenarios are then discussed in greater detail.

EXISTING POWER-SAVING METHODS

A taxonomy of relevant power-saving methods for IoT over FiWi networks is depicted in Fig. 3. The implementation of any individual or combinations of those power-saving methods can help improve the overall energy efficiency.

The power save mode (PSM) was defined in WLAN (IEEE 802.11a/b), where a wireless station (STA) can stay awake or enter sleep mode to reduce its energy consumption. An AP broadcasts a traffic indication map in the beacon frame indicating the STAs to receive queued packets. A STA must stay awake in every beacon interval to receive the beacon frame. To receive data packets from the AP, it must contend for the channel by means of a PS-Poll frame. The PSM scheme was improved in follow-up enhancements. In particular, IEEE 802.11e defined automatic power save delivery to avoid the PS-Poll procedure, whereas IEEE 802.11s enhanced the PSM scheme for mesh STAs. To address the contention when many STAs simultaneously send PS-Poll frames, the power save multi-poll scheme was defined in IEEE 802.11n. Besides, WLAN also supports the doze mode (turning off only the transmitter) of an AP, in which a dozing AP specifies its doze time in broadcast beacon frames.

The discontinuous reception (DRX) mechanism defined in 4G LTE is a typical timeout-driven power-saving mechanism, where the behavior of a DRX-enabled user equipment (UE) is driven based on a set of timers that collectively specify when it monitors the physical downlink (DL) control channel (PDCCH) for its scheduled DL and uplink (UL) resources and when it enters sleep mode. They include DRX cycle time, on-duration timer, and inactivity timer. The UE wakes up and monitors the PDCCH for the entire on-duration period. If no scheduling assignment is received, the UE falls asleep again. Otherwise, it monitors the PDCCH until the timer expires [9]. Since IoT traffic is known to be different from H2H traffic with infrequent and small transmissions, and IoT devices are typically low-power and resource-constrained sensors with limited functionality, enhancements are needed to render DRX more suitable for IoT applica-

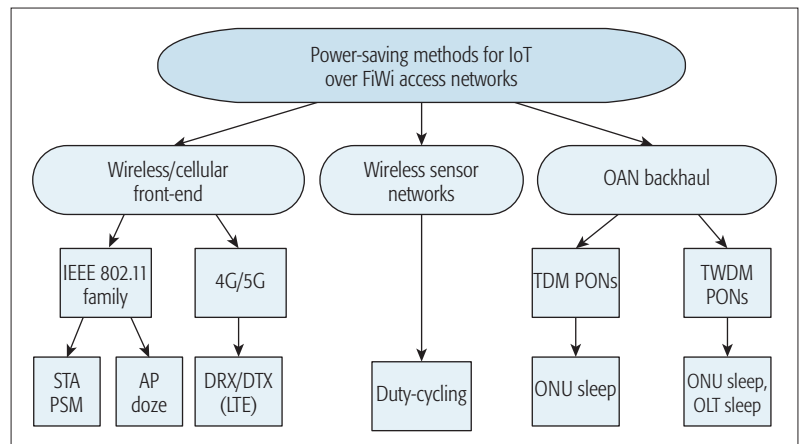


Figure 3. Taxonomy of power-saving methods for IoT over FiWi networks.

tions [9].

IEEE 802.15.4 defines duty-cycling techniques to reduce power consumption of battery-powered wireless sensors. A duty cycle is defined as the fraction of time a node is active during its life, and thus determines the trade-off between battery life and packet latency. To coordinate sleep/wakeup times among sensor nodes, a scheduling algorithm is required. However, in the context of IoT over FiWi networks, power-saving mechanisms for IoT sensors can be based on the duty-cycling or power-saving methods defined for wireless/cellular devices with relevant modifications depending on deployment scenarios (e.g., direct or indirect access).

Backhaul power-saving methods for OANs usually target ONU energy reduction due to its larger number of devices compared to the OLT. The most common method is polling-based ONU sleep, in which the behavior of an ONU strictly follows the TDMA-based polling system used for UL dynamic bandwidth allocation (DBA). However, timeout-driven ONU sleep mechanisms (e.g., sleep and periodic wake-up) are also studied. An ONU sleep mechanism aims to save energy for ONUs, while preserving the UL transmission and minimizing incurred DL packet delay [14]. There are also power-saving methods to improve OLT energy efficiency, especially in TWDM PONs. The combination of the ONU sleep and OLT sleep helps improve the overall energy efficiency.

DESIGN CHALLENGES AND GUIDELINES

Given that different OAN technologies can be employed at the backhaul, and various types of RATs can be employed at the front-end segment of IoT over FiWi networks, power-saving solutions vary depending on specific deployment details, such as technologies of choice, channel access mechanisms, network scale, and spatial coverage. Achievable energy efficiency gains and network performance greatly depend on how the backhaul and front-end power-saving modes are implemented and combined in addition to their device profiles, that is, wake-up capability (sleep-to-active overhead time) and power consumption in different states.

Although power-saving methods are defined in standards, implementation details are left open. Among those, it is challenging to deter-

In a nutshell, the overall energy efficiency is improved through scheduling operation modes for network devices according to their actual traffic loads and incorporate them into the underlying dynamic bandwidth allocation process to maintain the network performance.

mine *what* to turn off, *when* to enter power-saving mode, *how long* to sleep/doze, and, most importantly, *how* to incorporate power-saving modes into the underlying resource allocation process in order to save energy without compromising network performance. Key performance parameters include power ratio (i.e., the ratio of power levels in sleep and active states) and time ratio (i.e., the ratio of sleep-to-active time and cycle time). A common strategy is to reduce either or both ratios [14]. Reducing the power ratio means turning off more components during sleep state (i.e., answering the question *what*), whereas reducing the time ratio means either devising a device with a faster wake-up capability or configuring a longer sleep time in a cycle (i.e., answering the question *how long*). Triggering criteria (i.e., *when* to enter power-saving modes) can be based on configured timers, polling mechanism, or traffic condition. The question *how* to relates to the design of new power-saving-aware DBA algorithms with extensions of standard control messages.

In converged FiWi networks, an important factor when designing end-to-end power-saving mechanisms is how to coordinate and synchronize power-saving modes of different network elements of both backhaul and front-end segments. In principle, the optimal overall performance is only achieved when such harmonization is ensured in conjunction with the incorporation of power-saving modes in the resource management mechanism. For example, without such a unified solution, an ONU may be “OFF,” due to scheduling of the OLT that employs a timeout-driven sleep mechanism, when IoT mission-critical traffic arrives from the front-end segment. As a consequence, IoT applications will experience unsatisfactory network performance. Moreover, signaling mechanisms used for exchanging power-saving messages among network elements may incur high protocol overhead. To this end, designing a sleep/doze mechanism based on extensions of existing control frames, such as multi-point control protocol (MPCP) messages (GATE and REPORT) in EPON, is a preferable choice to minimize the overhead.

As discussed earlier, under the paradigm of converged Internet technologies, the coexistence of H2H and M2M traffic further diversifies communication characteristics and requirements. This poses a unique challenge to resource management in general and power-saving mechanisms in particular. For instance, it is more difficult to design power-saving solutions with quality of service (QoS) guarantees for IoT over FiWi networks than in conventional access networks. The hurdle is to allocate a shared but limited amount of resources for all human-centric devices and IoT smart devices, and at the same time schedule their power-saving modes in order to meet battery life requirements of some applications while satisfying QoS levels imposed by others. Furthermore, the power-saving mechanisms need to take the scalability issue into account. In particular, it is challenging to answer the question of to what extent a power-saving solution is efficient and scalable given that large-scale IoT is a common deployment scenario. In such a case, synchronization and interference

among wireless network elements becomes critical.

Since research in the area of IoT, as it is today, is just at the beginning of its long journey, building energy-efficient IoT networks has much to accomplish in the future. With a foreseen highly heterogeneous multi-RAT front-end in future networks, power-saving methods will need to handle the heterogeneity among different access technologies to ensure that the energy consumption is minimized while IoT service and conventional human-centric applications run smoothly over the envisioned network architecture. In short, it is challenging to address the question of how to optimize the energy efficiency with power-saving modes taking into account all the key aspects summarized in Fig. 1 at the same time with QoS in mind.

TYPICAL USE CASES

Given the heterogeneity of both multi-RAT front-end and backhaul technologies in diverse IoT deployment scenarios, there is no one-size-fits-all power-saving solution for the envisioned network. Based on the earlier discussion about the two enabling FiWi technologies (i.e., R&F and RoF), this work proposes end-to-end power-saving solutions for two typical deployment scenarios of IoT over R&F-based FiWi networks: IoT over EPON-WiFi and IoT over 10G-EPON-LTE. The main goal is to provide *complementary insights* into potential energy efficiency gains while considering the characteristics of network integration, scalability, and coexistence of IoT over FiWi networks. Note, however, that by following the aforementioned design guidelines, solutions for other deployment scenarios could be reached. In a nutshell, the overall energy efficiency is improved through scheduling operation modes for network devices according to their actual traffic loads and incorporating them into the underlying dynamic bandwidth allocation process to maintain the network performance. While the proposed solutions address the design issues of *what*, *when*, *how long*, and *how to* for the two deployment scenarios, other optimization methods can be applied to further improve the overall performance.

IoT over EPON-WiFi Scenario: WLANs are considered for medium- and small-scale IoT applications, where each ONU-AP zone covers a small number of devices. A typical example of this scenario can be found in smart grid communications, where static wireless sensor nodes are used to provide geographically and temporally coordinated monitoring and control actions for a wide variety of distributed grid elements. TDMA scheduling is employed, which helps improve energy efficiency, avoid collisions, and reduce interference among wireless and sensor nodes, thereby improving the overall network performance compared to contention-based protocols. This is particularly suitable for periodic timeout-driven smart grid sensors. By organizing the whole network in multiple TDMA layers, this approach is applicable to both direct and indirect access methods.

Figure 4a illustrates a unified bandwidth allocation and sleep/doze mode scheduling mechanism for the direct access scenario. The overall

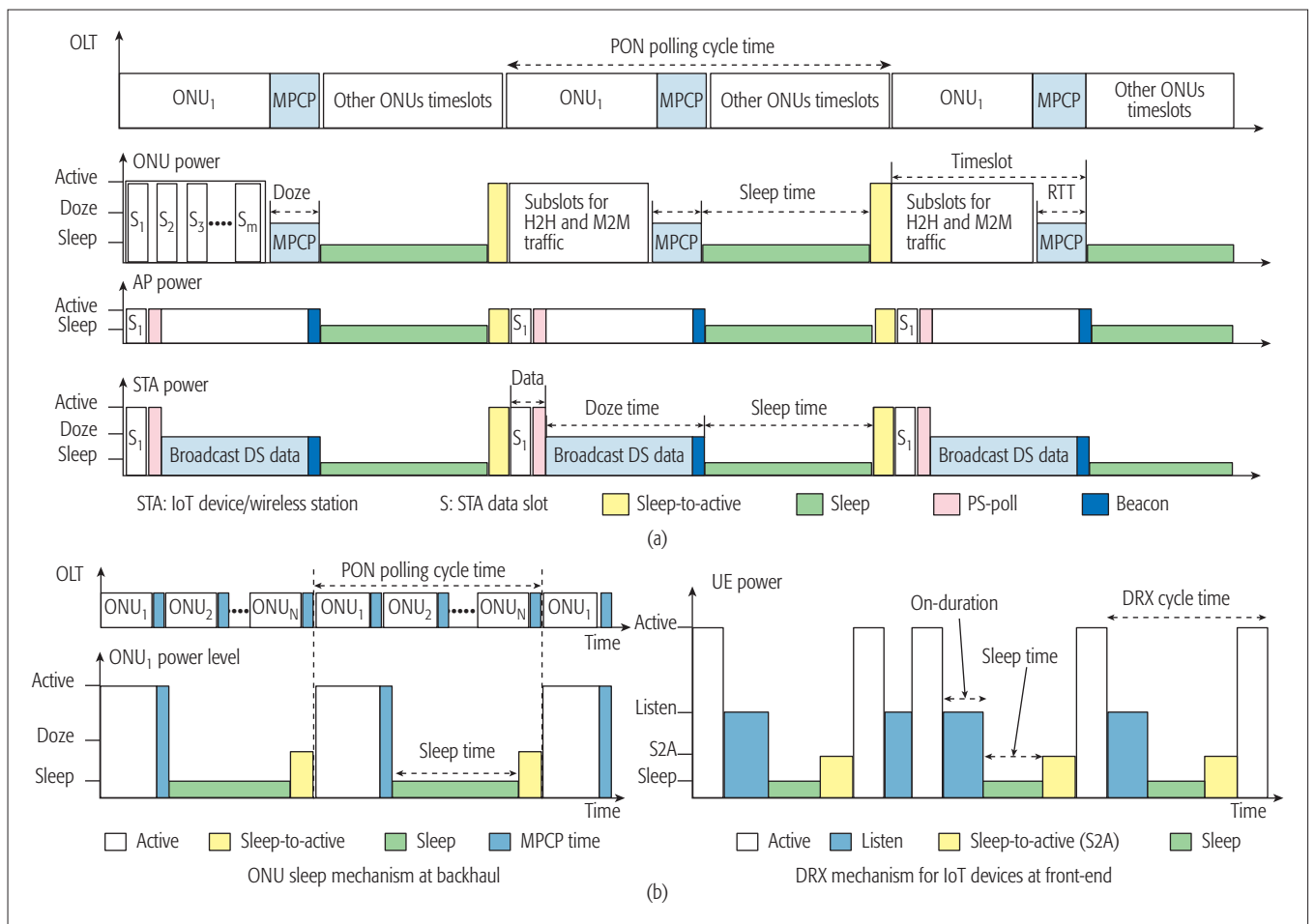


Figure 4. Illustration of power-saving solutions for IoT over FiWi networks: a) TDMA-based power-saving mechanism for IoT over PON-WLAN in a direct access scenario (RTT: round-trip time between OLT and ONU); b) power-saving mechanism for IoT over PON-LTE in a fragmented scenario.

objective is to make the active time of a network device proportional to its actual traffic load to minimize its energy consumption. To realize this, all ONU-APs and their associated STAs including IoT smart devices are assigned time slots for their data transmissions and exchanging control messages. Outside assigned time slots, they are scheduled to enter sleep mode for saving energy. Note that the OLT is kept always active to serve multiple ONU-APs with high aggregated traffic intensity. Time slot duration is determined based on actual traffic loads, which are reported to the scheduler (the OLT and APs) by means of control messages. The mechanism is polling-based, whereby operations of the OLT and ONU-APs are based on the energy-aware DBA (EDBA) algorithm [15], while each STA strictly follows its associated ONU-AP. Since network nodes are static (no mobility), network synchronization is maintained by adopting the timestamp mechanism specified in the EPON standard, where all network devices assign their local clocks to the OLT global clock embedded in DL signaling messages (i.e., MPCP GATE and WLAN Beacon frames).

IoT over 10G-EPON-LTE Scenario: This scenario applies to large-scale IoT deployments with 4G LTE employed at the front-end. As discussed earlier, large-scale IoT applications usu-

ally have relaxed QoS but more stringent battery life requirements. The solution is to implement a modified DRX mechanism with an extended DRX cycle at the front-end and the ONU sleep at the 10G-EPON-based backhaul, as illustrated in Fig. 4b. The left side of Fig. 4b details the operation sequence and corresponding power levels of an ONU with respect to the PON polling cycle (upper part). Given a massive number of IoT devices connecting to each ONU-eNB, the eNB module is kept always active to exchange data with IoT devices.

It is worth noting that most existing studies on DRX mainly focus on conventional UEs with H2H traffic without considering the sleep-to-active time. Indeed, such an overhead time can be relatively long with respect to the DRX cycle, resulting in significant differences between the assessed and actual performance in terms of battery life and incurred packet delay. Therefore, a modified DRX model with a separate sleep-to-active (S2A) state and an extended DRX cycle is considered in this scenario, as illustrated on the right side of Fig. 4b. Regarding end-to-end performance, the total packet delay is computed as the sum of delay components incurred by both ONU sleep and DRX mechanisms. Meanwhile, the overall energy efficiency gain is computed based on the gain obtained by each network seg-

The evaluation of the TDMA-based mechanisms is based on M/G/1 queueing model analysis, where the OLT/AP is viewed as the server that polls and serves multiple ONUs/STAs as clients and each timeslot is composed of a data interval, a reservation interval, and a vacation interval.

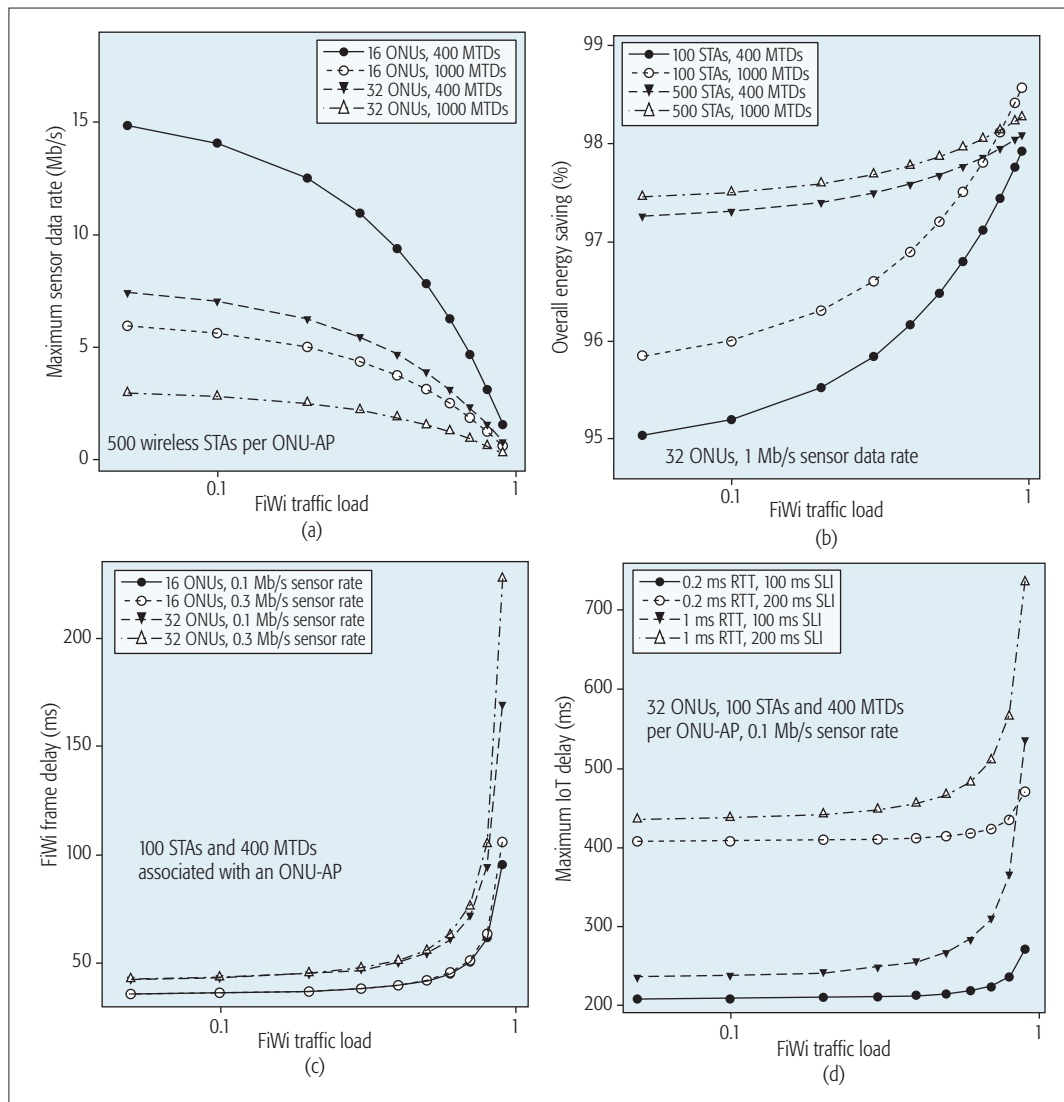


Figure 5. Performance of TDMA-based mechanisms for H2H/M2M coexistence scenario with sensor duty cycle of 1 percent: a) maximum IoT sensor data rate in a H2H/M2M coexistence system (MTD: IoT devices); b) overall energy saving; c) mean FiWi packet delay (200 ms sensor listening interval); d) maximum IoT packet delay (SLI: sensor listening interval; RTT: round-trip time between OLT and ONUs).

ment and their weights in the total energy consumption of the network [9].

PERFORMANCE EVALUATION

This section discusses results and findings obtained from an analytical evaluation of the two considered scenarios. Energy saving is defined as the relative energy consumption decrease with respect to the energy consumption without power-saving mode(s) enabled. Packet delay is the time a packet waits in a data buffer during the inactive states of a device. Battery life of IoT devices is computed based on the capacity of the battery under consideration and the average power consumption. The evaluation of the TDMA-based mechanisms is based on M/G/1 queueing model analysis, where the OLT/AP is viewed as the server that polls and serves multiple ONUs/STAs as clients, and each time slot is composed of a data interval, a reservation interval, and a vacation interval. Meanwhile, the performance of the DRX mechanism is analyzed

by using a semi-Markov process following the state diagram depicted in Fig. 4b with configured DRX parameters [9].

IoT OVER EPON-WIFI SCENARIO

The values of considered parameters include 1 Gb/s transmission capacity of both DL and UL, FiWi US traffic load (intensity) varied from 0.05 to 0.9 with Poisson distribution, and average packet transmission time of 5.09 μ s, IoT sensors with constant bit rate data, and extended 100 km network reach between the OLT and ONUs. ONU power profile includes 5052 mW, 3850 mW, and 750 mW, in active, doze, and sleep states, respectively. STA power levels in active, doze, and sleep states are 3500 mW, 1500 mW, and 20 mW, respectively. IoT sensor power levels in active and sleep states are 38.25 mW and 0.015 mW. Both AP and OLT consume 13,000 mW in active state. Sleep-to-active times of ONU, STA, and IoT sensor are 2 ms, 0.25 ms, and 2 ms, respectively.

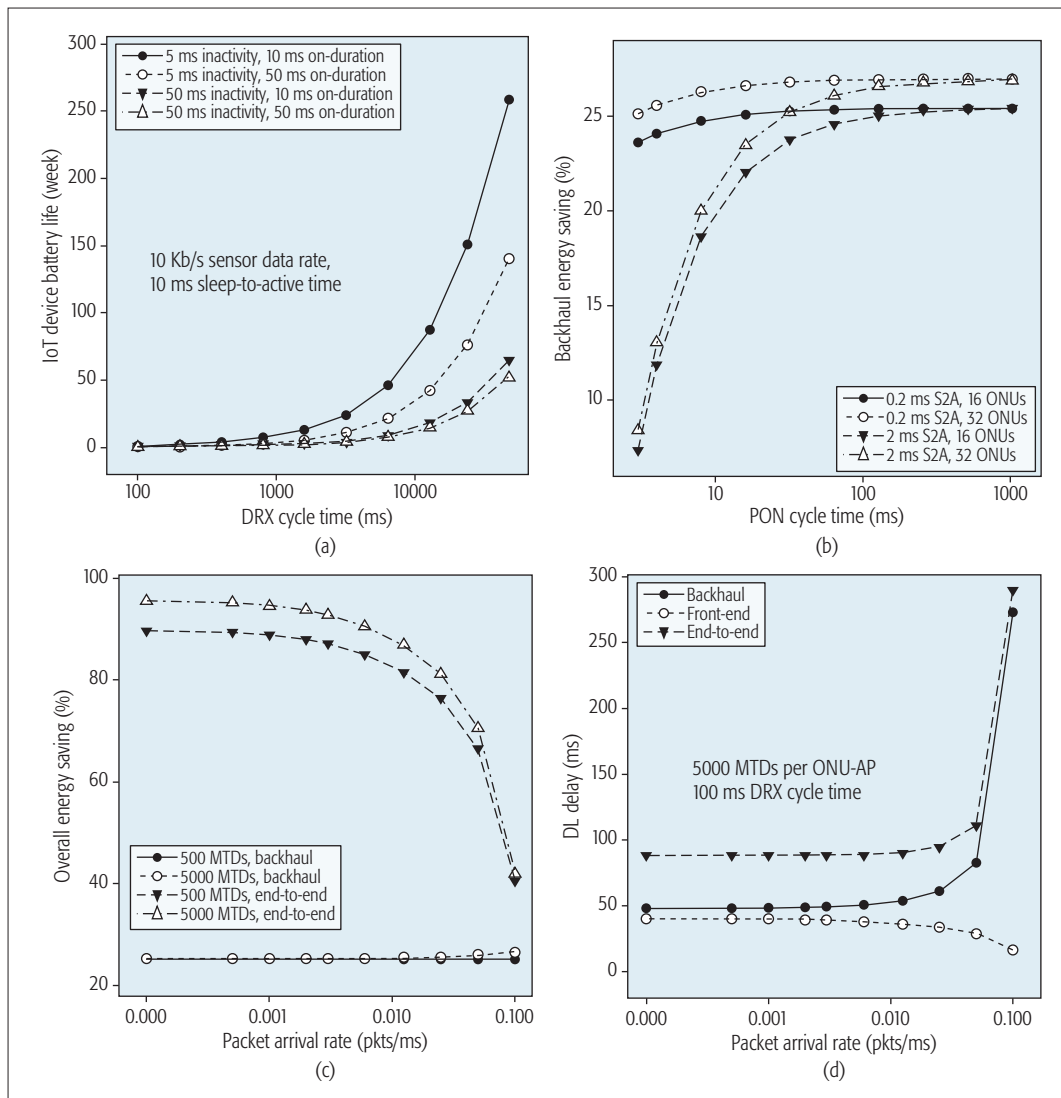


Figure 6. Performance of dense IoT deployment over PON-LTE with ONU sleep enabled backhaul and DRX enabled LTE front-end: a) IoT device battery life vs. DRX cycle time; b) backhaul energy saving vs. PON cycle time; c) end-to-end energy saving performance; d) end-to-end packet delay performance.

As an important guideline for this H2H/M2M coexistence scenario, Fig. 5a shows the maximum allowable IoT sensor data rate for a given FiWi traffic load for the configuration, in which IoT sensors have a duty cycle of 1 percent. As the FiWi traffic load increases, a lower sensor data rate is required. Noticeably, Fig. 5b shows that the integrated network saves more than 95 percent of total energy consumption of a typical system with 32 ONUUs for considered values of the parameters. The number of STAs significantly contributes to the energy saving gains, whereas the number of IoT sensors slightly affects it. It is important to highlight that the higher the traffic load, the more the energy saving. This is because of the TDMA nature of the mechanism (Fig. 4a).

Figure 5c clearly shows the dependence of FiWi packet delay on traffic load. It remains rather low when the load is low, but rapidly increases when the load increases. In addition, a larger number of ONU-APs also makes the delay increase because of the sharing nature in TDMA scheduling. On the other hand, the IoT

sensor data rate has marginal impact on overall FiWi delay since it slightly affects the polling cycle time that determines incurred delay. Figure 5d shows that FiWi parameters have marginal impact on the IoT sensor delay when traffic is light. But a higher value of FiWi traffic load results in less bandwidth available for IoT sensors and thus incurs a higher sensor delay. Instead, the sensor listening interval, the interval between two consecutive active periods, has significant impact on the delay.

IoT OVER 10G-EPON-LTE SCENARIO

This evaluation scenario considers 10 Gb/s transmission capacity of both UL and DL, 64 bytes packet size of sensor data, 10 ms sleep-to-active time, and IoT sensor arrival rate varied from 0.1 to 100 packets/s. Power consumption of the LTE eNB module is 13,000 mW. Considered LTE-enabled IoT sensors consume 500 mW, 255 mW, 255 mW, and 0.03 mW in active, listen, sleep-to-active, and sleep states, respectively. IoT device battery life is considered for a 1.5 V AA battery.

The number of STAs significantly contributes to the energy saving gains, whereas the number of IoT sensors slightly affects it. It is important to highlight that the higher the traffic load, the more the energy saving. This is because of TDMA nature of the mechanism.

The obtained results showed that in the considered small-scale WiFi-based IoT scenario, more than 95 percent of energy can be saved by employing TDMA-based scheduling, whereas up to 5 years of battery life can be achieved in the large-scale LTE-based IoT scenario by employing the proposed DRX mechanism.

Figure 6a reports the battery life of an IoT device as a function of DRX cycle time. Besides the cycle time, a low value of either/or Inactivity and On-duration timers helps extend the battery life. Remarkably, a 258-week (≈ 5 years) battery life can be achieved by employing the proposed DRX mechanism. Figure 6b shows the backhaul energy saving vs. the 10G-EPON cycle time. A long cycle time yields high energy saving. As the OLT and eNB modules are never put into sleep mode, backhaul energy saving is at most 27 percent for the considered parameters. The ONU wake-up time significantly affects its energy saving. This matches the general guideline provided earlier that besides a smart sleep mechanism, the transceiver architecture should be designed to have fast wake-up capability to improve the overall energy efficiency.

Figures 6c and 6d summarize the end-to-end performance as a function of an IoT sensor's packet arrival rate. Figure 6c reveals that the wireless front-end segment saves much more power than the backhaul. This is mainly due to the massive number of IoT devices (5000 sensors per ONU-AP). The total energy saving is therefore dominated by the front-end. Remarkably, more than 80 percent of overall energy can be saved by employing the proposed solution when traffic is light. As observed in Fig. 6c, the overall energy saving decreases with increasing arrival rate because the IoT device must stay longer in active and listen states when traffic is heavy. Figure 6d shows that packet delays have similar behavior with their respective energy saving curves. Backhaul TDMA scheduling offers a lower DL delay compared to the DRX delay. However, for high traffic loads, the backhaul delay increases rapidly. These results confirm the importance of a high-capacity backhaul link to alleviate the backhaul bottleneck in the envisioned network. Importantly, similar to the first scenario (Figs. 5b–5d), the energy-delay trade-off is clearly portrayed by Figs. 6c and 6d, in which the maximum achievable energy saving can be specified for any given delay constraints.

CONCLUSIONS

This work advocates the use of converged FiWi networks that integrate a capacity-centric OAN backhaul and a coverage-centric multi-RAT front-end network to underpin the IoT ecosystem. As energy efficiency is one of the primary hurdles for widespread adoption of IoT, this article provides deep insights into design and implementation of end-to-end power-saving mechanisms covering both backhaul and front-end network segments. The article then proposes power-saving solutions to jointly deal with the energy efficiency, network integration, scalability, and H2H/M2M coexistence for two typical deployment scenarios. The obtained results show that in the considered small-scale WiFi-based IoT scenario, more than 95 percent of energy can be saved by employing TDMA-based scheduling, whereas up to 5 years of battery life can be achieved in the large-scale LTE-based IoT scenario by employing the proposed DRX mechanism. Furthermore, desirable energy efficiency can be traded with acceptable packet delays in the presented solutions.

REFERENCES

- [1] S. Andreev *et al.*, "Understanding the IoT Connectivity Landscape: A Contemporary M2M Radio Technology Roadmap," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 32–40.
- [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] M. G. Institute, "The Internet of Things: Mapping the Value Beyond the Hype," research report, June 2015.
- [4] J. Thompson *et al.*, "5G Wireless Communication Systems: Prospects and Challenges," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 62–64.
- [5] M. Fiorani *et al.*, "On the Design of 5G Transport Networks," *Photonic Net Commun.*, vol. 30, no. 3, 2015, pp. 403–15.
- [6] D. Pham Van, B. P. Rimal, and M. Maier, "Fiber Optic vs. Wireless Sensors in Energy-Efficient Integrated FiWi Smart Grid Networks: An Energy-Delay and TCO Comparison," *Proc. IEEE INFOCOM*, Apr. 2016.
- [7] T. Orphanoudakis *et al.*, "Exploiting PONs for Mobile Backhaul," *IEEE Commun. Mag.*, vol. 51, no. 2, Feb. 2013, pp. S27–S34.
- [8] D. Kilper *et al.*, "Power Trends in Communication Networks," *IEEE J. Sel. Topics in Quantum Electron.*, vol. 17, no. 2, Mar. 2011, pp. 275–84.
- [9] D. Pham Van *et al.*, "Machine-to-Machine Communications over FiWi Enhanced LTE Networks: A Power-Saving Framework and End-to-End Performance," *IEEE/OSA J. Lightwave Tech.*, vol. 34, no. 4, Feb. 2016, pp. 1062–71.
- [10] S. Tozlu *et al.*, "Wi-Fi Enabled Sensors for Internet of Things: A Practical Approach," *IEEE Commun. Mag.*, vol. 50, no. 6, June 2012, pp. 134–43.
- [11] 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.
- [12] M. Maier and B. P. Rimal, "Invited Paper: The Audacity of Fiber-Wireless (FiWi) Networks: Revisited for Clouds and Cloudlets," *China Commun.*, vol. 12, no. 8, Aug. 2015, pp. 33–45.
- [13] M. Hasan, E. Hossain, and D. Niyato, "Random Access for Machine-to-Machine Communication in LTE-Advanced Networks: Issues and Approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, June 2013, pp. 86–93.
- [14] L. Valcarengi *et al.*, "Energy Efficiency in Passive Optical Networks: Where, When, and How?," *IEEE Network*, vol. 26, no. 6, Nov. 2012, pp. 61–68.
- [15] D. Pham Van *et al.*, "Energy-Saving Framework for Passive Optical Networks with ONU Sleep/Doze Mode," *Opt. Express*, vol. 23, no. 3, Feb. 2015, pp. A1–A14.

BIOGRAPHIES

DUNG PHAM VAN is a postdoctoral researcher at the Optical Networks Lab (ONLab), KTH Royal Institute of Technology, Sweden. He received his M.Sc. in ICT from Waseda University, Japan, in 2009 and Ph.D. (cum laude) in telecommunications from Scuola Superiore Sant'Anna, Italy, in 2014. From January to August 2015, he was a postdoctoral researcher at the Optical Zeitgeist Laboratory, Institut National de la Recherche Scientifique (INRS), Montréal, Québec, Canada. He was a visiting researcher at the University of Melbourne, Australia, in the first half of 2014. He has published about 40 papers in international journals and conference proceedings. He is the recipient of the Distinguished Student Paper Award presented at the OptoElectronics and Communication Conference and Australian Conference on Optical Fibre Technology 2014 (OECF/ACOFT) and the Best Student Paper Award (first class) presented at the Asia Communications and Photonics Conference 2013 (ACP). His current research interests include converged fiber-wireless networks, 5G backhaul, the Internet of Things, mobile-edge computing, and data center networks.

BHASKAR PRASAD RIMAL [S] received his M.Sc. degree in information systems from Kookmin University, Seoul, Korea. He is currently pursuing his Ph.D. degree in telecommunications at the Optical Zeitgeist Laboratory, INRS. His research interests include mobile-edge computing (MEC), fiber-wireless enhanced networks, tactile Internet, the Internet of Things, and game theory. He was the recipient of the Doctoral Research Scholarship from the Québec Merit Scholarship Program for foreign students of Fonds de Recherche du Québec -Nature et Technologies (FRQNT), the Korean Government Information Technology (IT) Fellowship, the Kookmin University IT Scholarship, and the Kookmin Excellence Award as an Excellent Role Model Fellow. He is a student member of ACM and OSA.

JIAJIA CHEN received a B.S. degree (2004) from Zhejiang University, China, and a Ph.D. degree (2009) from KTH Royal Institute of Technology. She is working as an associate professor in the Optical Networks Lab (ONLab) at KTH. She is a co-author of over 100 publications in international journals and conferences in the area of optical networking. Her main research interests are optical transport and interconnect technology supporting future 5G and cloud environments. She has been involved in various European research projects like European FP7 projects IP-OASE and IP-DISCUS, and EIT-ICT projects. Moreover, she is principal investigator/co-principal investigator of several national research projects funded by the Swedish Foundation of Strategic Research and the Swedish Research Council.

PAOLO MONTI [SM] is an associate professor at KTH Royal Institute of Technology. He serves on the Editorial Boards of *IEEE Transactions on Green Communications and Networking* and the *Springer Photonic Network Com-*

munications Journal. He has co-authored more than 100 technical papers, with three best paper awards. He regularly participates in TPCs including IEEE GLOBECOM and IEEE ICC. He is currently a TPC Co-Chair of IEEE Online-GreenComm 2016, HPSR 2017, the ONS Symposium at IEEE GLOBECOM 2017, and the OGN Symposium at ICNC 2017. His main research interests are within the networking and sustainability aspects of all-optical networks, with a special focus on optical transport solutions for 5G networks.

LENA WOSINSKA received her Ph.D. degree in photonics and doctorate degree in optical networking from KTH Royal Institute of Technology, where she is currently a full professor in telecommunication in the School of Information and Communication Technology. She is founder and leader of the Optical Networks Lab (ONLab). She has worked in several EU projects and coordinated a number of national and international research projects. Her research interests include fiber access and 5G transport networks, energy-efficient optical networks, photonics in switching, optical network control, reliability and survivability, and optical data center networks. She has been involved in many professional activities including Guest Editorship of IEEE, OSA, Elsevier, and Springer journals, serving as General Chair and Co-Chair of several IEEE, OSA, and SPIE conferences and workshops, serving on the TPCs of many conferences, as well as being a reviewer for scientific journals and project proposals. She has

been an Associate Editor of the *OSA Journal of Optical Networking* and *IEEE/OSA Journal of Optical Communications and Networking*. Currently she is serving on the Editorial Board of the *Springer Photonic Networks Communication Journal* and *Wiley Transactions on Emerging Telecommunications Technologies*.

MARTIN MAIER is a full professor with INRS. He was educated at the Technical University of Berlin, Germany, and received M.Sc. and Ph.D. degrees (both with distinction) in 1998 and 2003, respectively. In the summer of 2003 he was a postdoctoral fellow at the Massachusetts Institute of Technology (MIT), Cambridge. He was a visiting professor at Stanford University, California, from October 2006 through March 2007. Further, he was a Marie Curie IIF Fellow of the European Commission from March 2014 through February 2015. He is a co-recipient of the 2009 IEEE Communications Society Best Tutorial Paper Award and Best Paper Award presented at the International Society of Optical Engineers (SPIE) Photonics East 2000-Terabit Optical Networking Conference. He is the founder and creative director of the Optical Zeitgeist Laboratory (www.zeitgeistlab.ca). He currently serves as the Vice Chair of the IEEE Technical Subcommittee on Fiber-Wireless Integration. He is the author of the book *Optical Switching Networks* (Cambridge University Press, 2008), which was translated into Japanese in 2009, and lead author of the book *Fiber Access Networks* (Cambridge University Press, 2012).

Sustainability Information Model for Energy Efficiency Policies

Ana Carolina Riekstin, Bruno Bastos Rodrigues, Viviane Tavares Nascimento, Claudia Bianchi Progetti, Tereza Cristina Melo de Brito Carvalho, and Catalin Meirosu

The authors present the existing ways of representing policies, focusing on information models. They then propose SLIM, the Sustainability Information Model for Energy Efficiency Policies, an extension to the IETF Policy Core Information Model extension that allows unification of the management of green capabilities and protocols throughout the network.

ABSTRACT

The need to manage the energy consumption of network infrastructure has been addressed by a significant body of work in recent years. In general, energy management capabilities were developed independently and optimized for particular network layers and node features. The interaction between multiple such green capabilities when deployed simultaneously, as well as potential interactions with other existing functionality such as quality of service functions, need to be managed transparently by the operators. We developed SLIM, the Sustainability Information Model for Energy Efficiency Policies, as an add-on to the IETF Policy Core Information Model Extension to allow unifying the management of green capabilities throughout the network. We illustrate the flexibility of our approach by presenting a use case and describing an energy management system where SLIM was used.

INTRODUCTION

Driven by the widespread adoption of smartphones and other devices such as sensors and gadgets connected to the Internet, the amount of electrical power consumed by networks is growing. In response, energy efficiency features are being deployed in the network to make the operation more sustainable. However, due to the diversity of vendors and legacy equipment, the use of energy efficiency features and their coordination in an automated fashion is a complex task for network operators.

Network operators aim to automate the deployment and operations of new services by increasingly relying on software abstractions and the use of programmatic control methods. Abstractions and virtualization simplify the design and provisioning of services by moving dedicated appliances to generic servers. The programmatic control supports management and automation, helping administrators in delivering new services faster. In this context, “a generic policy-based management model that can be used to express policies on top of arbitrary configuration data models is essential” [1].

Policy-based network management (PBNM) and energy efficiency capabilities have been studied in recent years for sustainability-oriented

ed policies. Putting together the requirements for modeling policies and saving energy, Riekstin *et al.* [2] determined that policy refinement advances are needed for sustainability-oriented policies. Analyzing the existing solutions, the same authors proposed in [3] a method able to refine policies considering multiple abstraction levels and orchestrate different energy efficiency capabilities. One important component of the method is the information model, which is able to represent the different abstraction levels of sustainability-oriented policies. PBNM demands information models comprising business and system entities that can be implemented in an easy manner [4].

The Internet Engineering Task Force (IETF) defined an information model in RFC 3198 as an “abstraction and representation of the entities in a managed environment, their properties, attributes and operations, and the way that they relate to each other. It is independent of any specific repository, software usage, protocol, or platform” [5].¹ According to Agrawal [6], most policy systems do not have explicit information models defined, but they probably were somehow implicitly defined in the developers’ minds.

The Policy Core Information Model (PCIM), later extended to the Policy Core Information Model Extensions (PCIME) [7], defined policy rules in a vendor-independent way, supporting the definition of different levels of abstraction. It was based on the IETF/Distributed Management Task Force (DMTF) Core Information Model (CIM), a conceptual framework for the schema of the managed environment.

To represent sustainability-oriented policies, besides supporting traditional policies for authorization and obligation (e.g., access control and quality of service), an information model must be flexible enough to accept green metrics, time representation, or scenario conditions to enforce energy efficiency rules. Furthermore, it should comprise the modeling of actions that are specific to energy management, such as the *sleeping* and *performance adaptation* actions. It should also support modeling new types of metrics, such as the Watts per bits ratio along with the traditional performance metrics, and, in this way, support the study of the trade-offs between saving energy and maintaining performance.

¹ According to Agrawal *et al.* [6], “an information model is a structure for organizing information or data. For this reason, it is often called a data model.” Many discussions exist on information vs. data model, as can be seen in RFC 3444. We stick with the traditional definition from the IETF.

While PCIM and PCIME support specification of policies, their existing classes do not have the specificity required for the green policies modeling. Although the existing models enable the definition and translation of business policies, the enforcement of their capabilities demands a different modeling approach. We are not aware of works in the literature that define a generic information model specifically aimed at energy management policies for telecommunications infrastructure.

In order to fill this gap, we define a Sustainability Information Model (SLIM)² specified using Unified Modeling Language (UML), allowing energy management policies to be expressed at different levels of abstraction. To model the policies at each level of abstraction, PCIME [7] was used as the basis, and extended to comprise energy management aspects. In this work we revisit the main concepts and related works, considering energy efficiency capabilities, PBNM, and policy representations. We present the SLIM and its different levels of abstraction followed by a use case in order to discuss its benefits, limitations, and future work directions. Finally, we summarize the work, presenting our final remarks.

BACKGROUND

To understand the main concepts and related works involved in the proposal of SLIM, we divide this section in two parts. The first part describes PBNM concepts, while the second part focuses on efforts toward policy representation using information models.

PBNM

PBNM supports dealing with the complexity and heterogeneity of systems, devices, applications, and networks. IETF RFC 3198 [5] defined policy according to two perspectives. The first defines policy as a goal, purpose, or method to guide decisions. The second view, taken from RFC 3060 [10], describes policy as a set of rules that enables the management, administration, and control of the network resources.

Strassner [9] established five levels of abstraction for policies: business, system, network, device, and instance levels. Altogether, they were referred to as the Policy Continuum. Carvalho *et al.* [10] applied the Policy Continuum to sustainability-oriented policies and proposed a methodological approach for refinement. The business level comprises service level agreements (SLAs), guides, and goals. The system level comprises sustainability and performance indicators. The network level includes metrics for network operations related to technologies. The device level comprises metrics for device operation. The instance level has variables applied to devices and their components.

PBNM and energy efficiency capabilities have been studied in recent years for sustainability-oriented policies, defined as policies “that manage energy efficiency features in the network” [3]. Seven requirements were identified for the refinement of sustainability-oriented policies [2]. The refinement method should ideally:

- Support the *translation* of high-level policies into network-level actions

- Take into account the *resources* in the network, including the energy efficiency capabilities available
- *Verify* that the refined policies meet the requirements of the original policies
- Detect and solve *conflicts* among policies
- Handle *dynamicity* by supporting different time slots or be able to determine what to do when the scenario changes
- *Orchestrate* different energy efficiency capabilities, that is, choose the best capability considering the network situation, combine capabilities in order to increase the energy efficiency, and avoid combining conflicting capabilities
- Be able to *represent policies* in order to keep the context, coherence, and integrity of the network

Recent advances in software defined networks (SDN) have led to extra features to address these requirements. It introduces more programmability and flexibility to the control plane through a centralized management point, aware of the whole network. This allows the development of more complex management techniques in an easier way compared to legacy networks.

POLICY REPRESENTATION

Several proposals have been developed to represent policies. The Policy Description Language (PDL) introduced semantics and an architecture using the event-condition-action form to define a policy as a function that maps a series of events into a set of actions [13]. Although not covering all classes of policies, the syntax of PDL was simple, and policies could be described as an implementation of the Extensible Markup Language (XML) syntax, which resulted in self-describing documents.

Ponder enforces obligation policies in a similar approach to PDL, and offered a complete toolkit to specify, analyze, and enforce policies. Its last version, Ponder2 [11], was an object-oriented runtime policy management framework for authorization and obligation policies. Policies in Ponder2 were defined using PonderTalk, a high-level language based on Smalltalk.

Using information models is another way to represent policies, as suggested by the IETF. It is typically represented using UML to represent managed objects at a conceptual level, modeling the structure, relationships, constraints, rules, and operations between managed objects and policies that affect their management [12]. Often, policies have the form *if Condition, then Action*, and could be represented in what is called a *condition-action policy information model*. One real-world extension of this model incorporates the event in the policy definition. The event can also be described in the condition part of the basic policy rule as *if Event and condition C is true, then Action* [6].

Extracting information from application-specific information models presents some benefits, such as easy modification when needed, policy reuse in different application domains by deploying a different information model, and standardized information representation. Further, Damianou [13] states that objects can be mapped to structure specifications, such as XML.

The Open Networking Foundation (ONF)³ released its Core Information Model (ONF-

Using information models is another way to represent policies, as suggested by the IETF. It is typically represented using UML to represent managed objects at a conceptual level, modeling the structure, relationships, constraints, rules, and operations between managed objects and policies that affect their management.

² Essentially, SLIM is just one information model with different levels of abstraction, but we can call the different levels *information models*.

³ The Open Networking Foundation (ONF) is an organization that promotes the development and adoption of software-defined networking.

To specialize PCIME to sustainability-oriented policies, new classes were created to describe sustainability rules, conditions, and actions. The information model should also model the elements the administrator manipulates to compose the business rules.

CIM), which provides a representation of data plane resources for management control. According to the document, “the controller expresses a view of the network, in terms of ONF-CIM artifacts, to client SDN controllers/applications to meet the needs of that client.” As a future work, they intend to develop a policy module.

PCIM and PCIME: PCIM is an “object-oriented information model for representing policy information” [8]. It was developed as a complementary document of the CIM, an object-oriented information model published by the DMTF. In order to bring quality of service (QoS) requirements to the network, the specific QoS Policy Information Model (QPIM) was standardized by [14]. It specializes the PCIM to describe QoS actions. PCIM is not bound to a particular implementation; therefore, it can be used to exchange information in a variety of ways.

The model structure comprises two types of objects:

- *Structural classes*, which define ways of representing and controlling policy information
- *Associative classes*, which indicate how the class instances are related

In PCIM, a policy (class *Policy*) is defined by a set of rules (class *Policy Rules*), grouped by the *Policy Group* class. Each rule is composed of a set of conditions (class *Policy Condition*) and a set of actions (class *Policy Action*). The rules can also comprise time conditions (class *Policy Time Period Condition*). The actions can be executed in a specific order using the attribute *PolicyRule.SequencedAction*. Variables and values are used to build conditions following the structure (*<variable> MATCH <value>*) [7].

PCIME was built on top of PCIM, and two main changes were introduced: the inclusion of new elements, extending PCIM to areas that it did not previously cover; and the update of deprecated elements, such as policy rule priorities, replaced by priorities tied to associations that refer to policy rules.

SUPA: Simplified Use of Policy Abstractions (SUPA) is under development at IETF [1]. Given that different industry actors embrace specific policy languages based on terminologies and concepts that are familiar to each specific technology domain, SUPA aims to define a model for expressing policies at different levels of abstraction, independent of languages, protocols, and data repositories. Rather than using multiple software for each policy language, it aims to provide a common information model addressing different representations.

SUPA is focused on addressing some PCIM shortcomings, such as class inheritance issues and the lack of design patterns that would help build hierarchies (SUPA uses the composite pattern), among others. We kept using PCIM and PCIME for SLIM because SUPA was work in progress, and its new constructs would not address all gaps regarding the energy management policies that we need to describe. In the future, SLIM could be evaluated as a specialization of SUPA.

SLIM

As detailed in RFC 3640 [7], extensions of the defined information models may be defined, as QPIM extended PCIM for describing QoS pol-

icies. It states that properties can be included in the existing classes, while new classes and sub-classes can be defined [7]. The policy description, enforced by the information model, may address anything starting from the business directives all the way to the detailed configuration of the nodes.

OVERVIEW

The PCIME base classes that are present in the different SLIM abstraction levels are:

- *PolicySet*, used to group policies.
- *PolicyRoleCollection*, an addition in relation to PCIM, used to aggregate resources that share a common role.
- *PolicyGroup*, derived from *PolicySet*, which is a “container for a set of related *PolicyRules* and *PolicyGroups*.”
- *PolicyRule*, inherited from *PolicySet*, which is the main class to represent the “If Condition then Action” semantics. It has two mandatory properties, *SequencedActions* and *ExecutionStrategy*, which are further defined by other classes aggregated by *PolicyRule*.
- *PolicyCondition*, an abstract class used to define policy conditions.
- *PolicyAction*, an abstract class used to define policy actions.
- *SimplePolicyCondition*, composed by the *<variable> MATCH <value>* structure.
- *PolicyTimePeriodCondition*, which represents the periods during which a *PolicyRule* is active.
- *CompoundPolicyCondition*, which aggregates simple policy conditions. This class has the property *ConditionListType*, used to specify if the associated policies are in disjunctive normal form (DNF), the default option, or conjunctive normal form (CNF).
- *SimplePolicyAction*, which models the structure *SET <variable>* to *<value>*.
- *CompoundPolicyAction*, which aggregates simple policy actions. This class has the properties *SequencedActions* and *ExecutionStrategy*, reported to the *PolicyRule* class.
- *PolicyValue*, an abstract class to define value objects. Examples of derived classes are *PolicyIPv4AddrValue* and *PolicyMACAddrValue*.
- *PolicyVariable*, an abstract class used to define variables, which are used in individual conditions.
- *PolicyExplicitVariable*, which “indicates the exact model property to be evaluated or manipulated.”
- *PolicyImplicitVariable*, which models “implicitly defined policy variables (that) are evaluated outside of the context of the CIM Schema.” Sub-classes must specify the data type and semantics of the variables [7].

We refer the reader to RFC 3460 [7] for more details on the classes and sub-classes.

To specialize PCIME to sustainability-oriented policies, new classes were created to describe sustainability rules, conditions, and actions. The information model should also model the elements the administrator manipulates to compose the business rules (user, SLA, class of service,

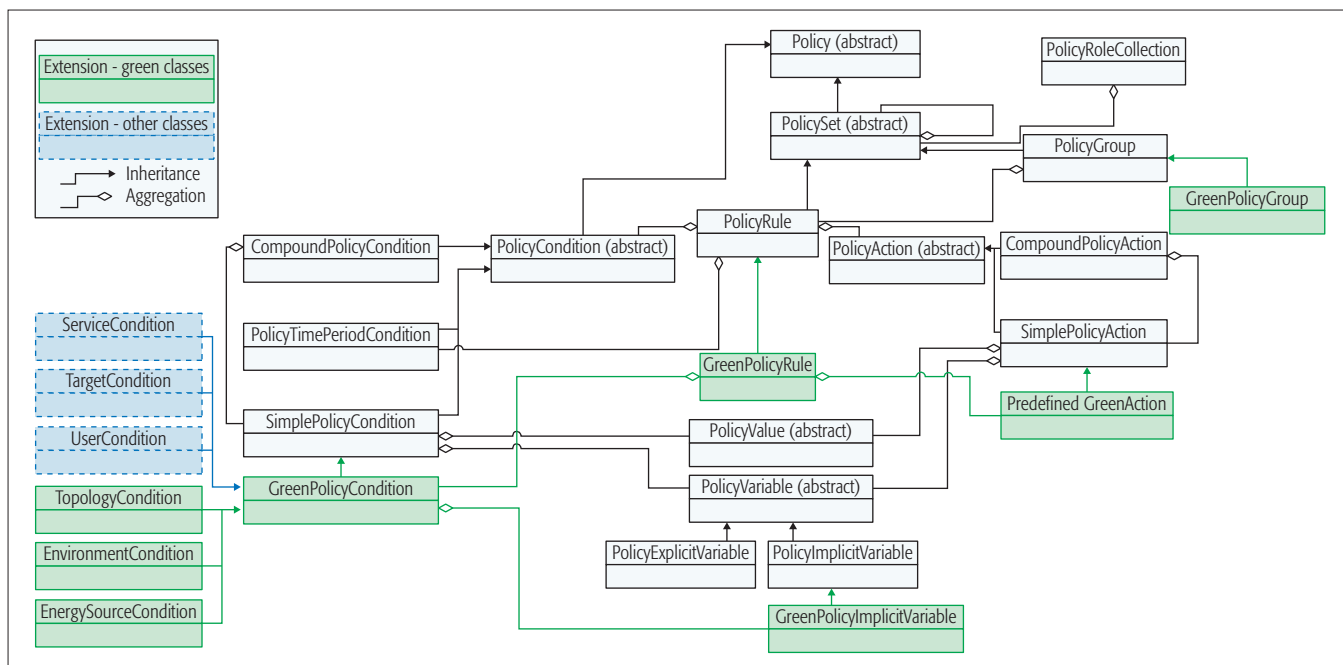


Figure 1. SLIM — business/system level of abstraction.

application, servers), and the information needed to configure the devices [15].

Considering the sustainability-oriented policies' requirements, the different Policy Continuum levels, the PCIM/PCIME characteristics, how QoS can be handled by QPIM and the extension proposed by Beller *et al.*, and how the SLA semantics was supported by them [15], three abstraction levels are proposed for SLIM: business/system, network, and device levels for sustainability-oriented policies representation. The Instance Level Information Model is particular for each technology or vendor, and therefore was not detailed.

The business and system information models are the same — they completely represent both levels with the same classes. They are still considered as two different layers to ensure the generality necessary for the near future systems. This first SLIM level is further specialized in each of the other levels. The Network Level Information Model is directly influenced by QPIM and includes technology-specific, device-independent information. The Device Level Information Model is the most detailed and considers device-specific information, including the variables that are expected to be managed to put the green capabilities to work.

BUSINESS AND SYSTEM LEVELS

Figure 1 details the business and system information model level. The *GreenPolicyRule* derives from the *PolicyRule* and is used to determine the conditions and actions for sustainability-oriented policies.

The *SimplePolicyCondition* is extended by the *GreenPolicyCondition*. The latter is extended by the *EnvironmentCondition*, which relates to the current situation in the network, such as the traffic being handled by the network, and by the *EnergySourceCondition*, which relates to the influence of different energy sources on the pol-

icies (e.g., for green SLAs that limit the amount of carbon to be emitted). The *GreenPolicyCondition* is also extended by the *TopologyCondition*, which represents the different topologies for the network, since specific topologies (e.g., SoHo, WAN, fat tree) can trigger different actions (e.g., the elastic tree capability works only for fat trees).

The other classes inheriting from the *GreenPolicyCondition* are supporting classes to represent users, services, and any associated condition (e.g., users in the HR department).

The *SimplePolicyAction* is extended by the *Predefined GreenAction* in the Business/System Levels Information Model, which determines different green plans for users under different contracts (Green SLAs). This is the part that changes more in the Network Level Information Model and the Device Level Information Model in relation to the Business and System Levels.

The Variables are extended by the *GreenPolicyImplicitVariable* class, which is further extended to comprise:

- At the business level, for instance, the operational expenditures (OPEX) accounting or total carbon emissions for Green SLAs
- At the system level, Watts per hour per energy source or the power usage effectiveness (PUE) facility, among others

NETWORK LEVEL

Figure 2 details the network information model level. This level is related to the capabilities that act on the whole network, such as those related to green traffic engineering. This level also needs to be aware of the capabilities that act in the device and its components to ultimately enforce the rules.

The network level information model details the business and system classes, giving policies a more technical format, with configuration details, represented in the class *Config GreenPolicyRule*.

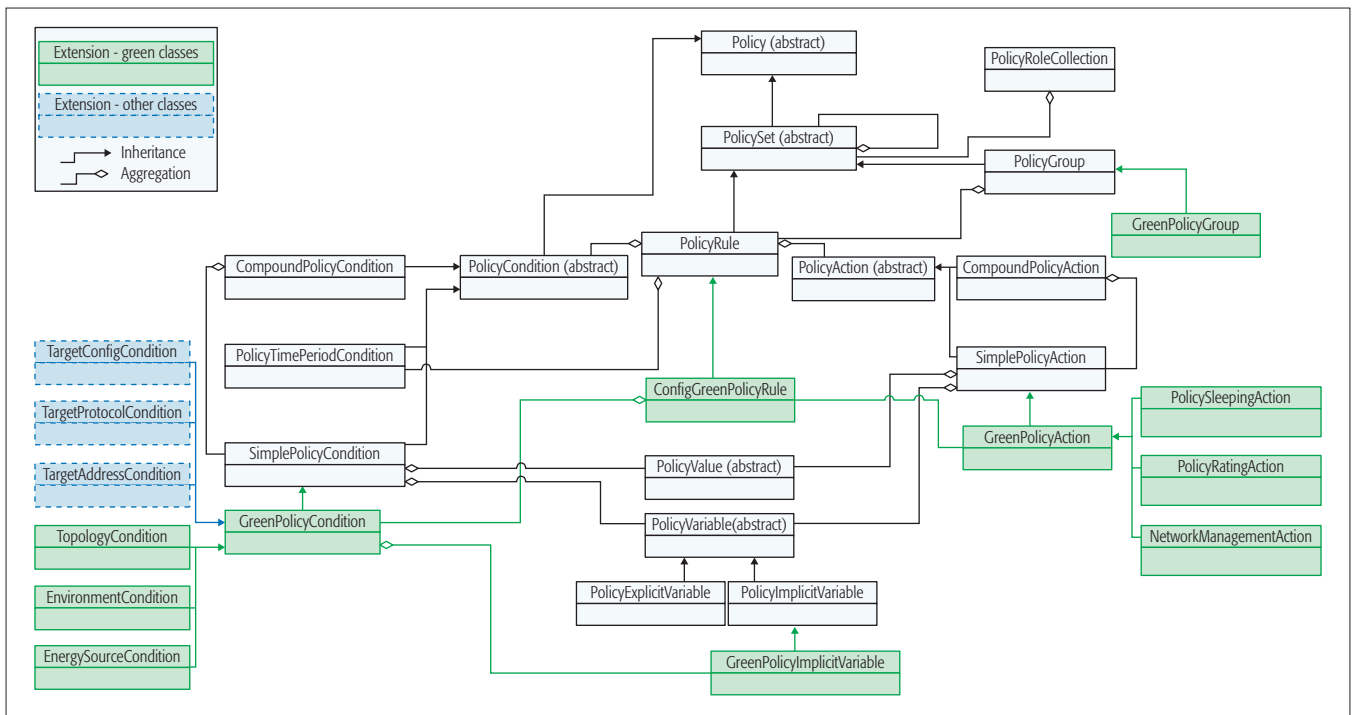


Figure 2. SLIM — network level of abstraction.

This abstraction level groups the conditions for the capabilities employment on the *GreenPolicyCondition* class. The defined conditions suggest the configuration of protocols and addresses components susceptible to the application of energy efficiency features. These classes are *TargetConfigCondition*, *TargetProtocolCondition* and *TargetAddressCondition*.

These supporting classes demand the definition of different *PolicyValues*, comprising, for instance, a value for the traffic load. The *GreenPolicyImplicitVariable* also gets a more technical aspect, representing, for instance, maximum loads for the whole network, or the Watts/bits ratio.

In the Action part, the business and system goals take the format of a more technical *GreenPolicyAction*. The Sleeping and Rating types of actions appear, detailing the approach each policy requires to put the higher-level policies into practice. *PolicyRatingAction* checks which devices are below or above the load levels defined by the user for each link rating (e.g., 10 Mb/s, 100 Mb/s, and 1 Gb/s), while the *PolicySleepAction* changes the device state when it reaches the specified load level of the system policies. The *NetworkManagementAction* class models the capabilities that are aware of the entire network to make decisions such as those related to green traffic engineering (followed by sleeping actions to save energy on underutilized nodes), or to virtual machines migration.

DEVICE LEVEL

Figure 3 details the device information model level. This level is related to the capabilities that act in the node, but also deals with the device components. The instance level policies will ultimately be applied on the node components and capabilities. The device level details the conditions and the device-specific actions, naming the variables that each green capability must handle.

It models, in a generic way, the possible energy efficiency actions in a device: sleeping and rating. The *SimplePolicyAction* is extended by the *GreenPolicyAction*, which determines different policies related to energy efficiency capabilities: *PolicySleepingAction* and *PolicyRatingAction*. These classes can then be extended to comprise information for the capabilities available in the network, using *DeviceSleepingAction* and *DeviceRatingAction*.

There are also supporting classes that describe the conditions of the device and its components, which are *TargetDeviceCondition*, *TargetConfigCondition*, *TargetProtocolCondition*, and *TargetAddressCondition*. These supporting classes demand the definition of different *PolicyValues*, comprising, for instance, the TargetDevice ID, its interfaces, power profile, and maximum and current load it can handle. The *GreenPolicyImplicitVariable* also goes down to the device level, representing, for instance, the node Energy Consumption Rating (ECR) or its Energy Proportionality Index (EPI), which can be used inside the conditions in this level. Table 1 summarizes the extensions.

This section has presented the different levels of abstraction proposed. As shown, all the enforcement aspects of the energy management capabilities were included in the SLIM, ranging from the conditions to the actions considered. SLIM enables an independent specification of the refinement functionality associated to policy translation between the levels.

USE CASE

To illustrate how SLIM works, we briefly describe the orchestration method proposed by Riekstin *et al.* [3], in which SLIM was used. Figure 4 depicts the method implementation architecture. The method uses table lookups for the high-level policies translation, with the support

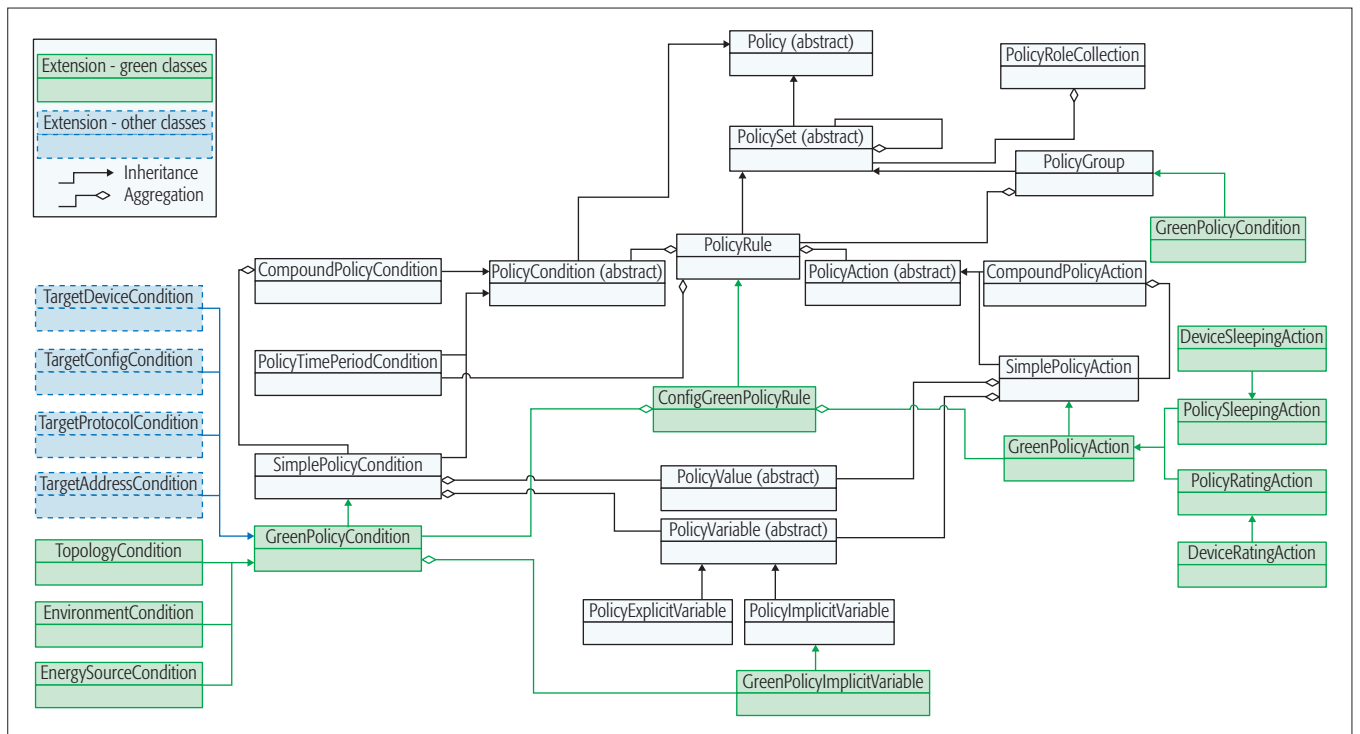


Figure 3. SLIM — device level of abstraction.

of SLIM to represent the policies. Translated policies are then used in conjunction with a utility function to deploy a decision tree able to select the best energy efficiency capability (or combination of capabilities) for a given scenario, ensuring conflict-free operation and a capability(-ties) selection that will not lead to congestion or high packet losses.

A set of tables needed to be predefined, related to Environment Condition, Time Condition, and Action. The *Environment Condition* was used to model the high or low use of the capacity of the network structure: the input will determine the traffic load conditions for the use of the energy capabilities. Table 2 illustrates the data that can be defined as environment conditions. Each row represents one possible environment condition, while the column represents the translated policy from one level to the other, using the information from SLIM.

The *Time Condition* is the input that will provide the data regarding the period of the day when the energy capabilities can be applied to the network infrastructure. Table 2 presents the time conditions. Each row represents one possible time condition, while the column represents the translated policy from one level to the other, using SLIM.

The last input is the *Action*. It determines the energy efficiency behavior that must be applied in the network infrastructure or if the performance of the network is more important than the reduction of the electrical expenditure at a given point of time. Table 2 shows the data that can be defined as actions. Each row represents one possible action, while the column represents the translated policy from one level to the other, using the information from SLIM.

Each one of the business policies received is compared to the content of the respective table.

Extension	Summary
Business/System level	
Green policy rule	Has the conditions and actions for sustainability-oriented policies
Green policy condition	Sustainability-oriented conditions: Topology, Environment, and EnergySource conditions representation
Green policy implicit variable	Represents the energy-related variables, such as Watts/bits ratio
Predefined green action	Represents different green plans for the users
Network level	
Config green policy rule	Conditions and actions in a more formal manner, such as event-condition-action policies
Green policy condition	Conditions get more technical format, such as representing the network total capacity and protocols
Green policy implicit variable	Supporting the conditions, also gets a more technical aspect, for instance, comprising the maximum load the network can handle
Green policy action	Represents management actions (whole network) and is aware of node actions (sleeping and rating)
Device level	
Green policy condition	Conditions go to the devices, detailing, for instance, IDs, their roles (core/edge), etc.
Green policy implicit variable	Variables related to the devices, such as device indexes
Green policy action	Represents device actions, grouped in rating and sleeping (can be extended to comprise others in the future)

Table 1. Examples of policy refinement using SLIM: business to system to network to device level.

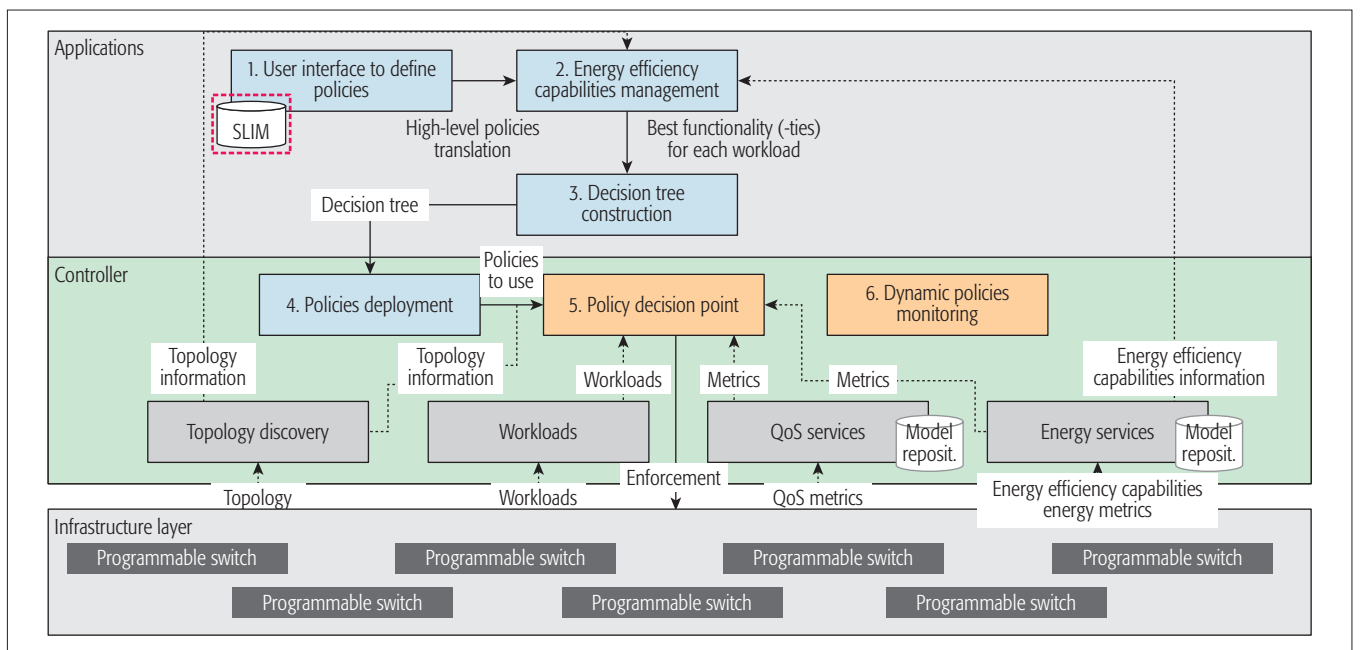


Figure 4. Orchestration method implementation modules with the SLIM highlighted [3].

The translated information is selected from the table, and is saved in a repository for future use.

To help understand how table lookup is employed by the method, an example of information defined in the Interface and its respective translation is provided below. An example of business policies defined in the interface is:

Environment Condition: *if usage is low*
 Time Condition: *during the night*
 Action: *save energy*

The module opens the *Environment Condition — Business to System Level* table, searches for the *if usage is low* input, and then selects the percentage of the load for the condition. The module then opens the *Environment Condition — System to Network Level* table and performs the same operation to translate the percentage to Mb/s information.

if usage is low → 20% → *NetworkCapacity = 1 Gb/s and Load < 200 Mb/s*

The module opens the *Time Period Condition* table, searches for the *during the night* input, and then selects the time (start and end time) that the method can be employed.

during the night → 22 h < hour < 6 h → *starting time: 22 h, ending time: 6 h*

The module opens the *Action — Business to System* table, searches for the *save energy* input, and then selects the action required for the action. The module performs the same to translate from the system to the network level of the Policy Continuum,

save energy → *use energy efficiency in the network* → *check in the decision tree the best capability for the given bandwidth utilization, all capabilities available are allowed*

The result is the metrics definition in a repository, already translated, and ready for the use by the orchestration part of the method:

[*'NetworkCapacity==1000 and load < 200.0', 'time > = 22 and time < = 6', check in the decision tree the best capability for the given bandwidth utilization, all capabilities available are allowed'*]

For instance, the available capabilities are related to green traffic engineering followed by putting the unused nodes to sleep. Therefore, on the network level, the policy will allow the green traffic engineering capability and enforce it, using the representation of the Network Management Action in SLIM. On the device level, the policy will tell the system to put unused nodes to sleep, using the representation of the Device Sleeping Action in SLIM.

DISCUSSIONS AND FUTURE WORK DIRECTIONS

Defining information models focused on sustainability takes a step toward a common representation of green policies. This work extends PCIME to support specific requirements and constraints that sustainability-oriented policies require. SLIM was defined with different levels of abstraction to be aligned to the Policy Continuum and thus operate at several levels of abstraction. SLIM was implemented and validated as part of a method that was previously published [3].

These new classes introduced by SLIM, however, do not restrict the policy definition for new capabilities that may be proposed: this is because SLIM was defined generically, not restricting the green classes to existing technologies. If required, new sustainability aspects could be included, such as life cycle assessment. Regarding the actions, currently, they can be related to sleeping or rating, so they can be modeled using SLIM proposed classes. If new types of energy management capabilities are

Business level	System level	Network level	Device level (1..N)
Environment condition			
"If usage is low"	20%	NetCapacity = 1 Gb/s and Load < 200 Mb/s	DevCapacity = 100 Mb/s and DevLoad 20 Mb/s
"If usage is high"	80%	NetCapacity = 1 Gb/s and Load < 800 Mb/s	DevCapacity = 100 Mb/s and DevLoad 80 Mb/s
"In any condition"	Always	NetCapacity = 1Gb/s and Load = ANY	DevCapacity = 100 Mb/s and DevLoad = ANY
Time condition			
During the night	22h hour 6h	Start: 22h/End: 6h	Start: 22h/End: 6h
During the morning	6h hour 12h	Start: 6h/End: 12h	Start: 6h/End: 12h
During the afternoon	12h hour 18h	Start: 12h/End: 18h	Start: 12h/End: 18h
During the evening	18h hour 22h	Start: 18h/End: 22h	Start: 18h/End: 22h
During the day	6h hour 22h	Start: 6h/End: 22h	Start: 6h/End: 22h
Action			
"Save energy"	"Use energy efficiency in the network"	"Check in the decision tree the best capability for the given bandwidth utilization, all capabilities available are allowed. E.g., use green traffic engineering to concentrate traffic"	"Enforce the selected green capability in each device. E.g., apply Synchronized Coalescing (sleeping capability) in all nodes or put unused nodes to sleep (after green traffic engineering)"
"Prioritize performance"	"Provide maximum QoS"	"Don't apply any energy effic. capability"	"Keep devices working at full performance"
"Save energy without reducing perform."	"Use energy effic. without reducing QoS"	"Apply only link rating capabilities"	"Enforce link rating capability(-ties) in each device"

Table 2. Examples of policy refinement using SLIM: business to system to network to device level.

developed, new classes could be added to support them, extending SLIM. The model presented in this work supports the refinement through the levels as shown with the use case presented.

SLIM addresses specifically the requirement "Be able to represent policies in order to keep context, coherence, and integrity of the network," one of the requirements a method should fulfill to be able to refine and orchestrate energy efficiency capabilities [2]. Despite being specifically tied to this requirement, it is totally aligned with the others, supporting the translation, resources representation, verification, conflicts detection and resolution, dynamicity, and orchestration. For instance, the time dynamicity is only possible with the information model class *TimePeriod-Condition*. Another example is the resources representation support, with the different levels of abstraction representing capabilities, nodes, and the whole network.

The information modeling exercise done for PCIME can also be done for the cloud computing context with the OpenStack Congress, for instance, so that the framework is able to model energy efficiency capabilities using specific constructs. The SLIM for cloud computing environments must take into account, besides business level policies, energy efficiency capabilities, the different resources in the infrastructure (compute, storage, network), and the management of virtual and physical resources. Another interesting future work can emerge from the ONF-CIM initiative of defining a Policy Module in the near future.

FINAL CONSIDERATIONS

Energy efficiency is presented as an environmental and economical advantage for network operators since it can reduce the operational costs and

the emission of pollutant gases. Many capabilities and protocols have been proposed to improve energy efficiency in networks. PBNM associated with information models help to manage the network infrastructure considering the availability of such capabilities and the traditional QoS and access control constraints, besides enabling a more automated operation.

Although different approaches for improving policy information models have been proposed, green policies were not considered previously. In this work we present SLIM, a sustainability-oriented information model with different levels of abstraction to fill this gap. We extend PCIME including new groups and rules of green policies, as well as new conditions and green actions. Through the information model proposed, we hope to bridge the gap in order to improve the control of network energy efficiency, addressing one important part of sustainability-oriented policies refinement.

REFERENCES

- [1] J. Strassner, J. M. Halpern, and J. Coleman, "Generic Policy Information Model for Simplified Use of Policy Abstractions (SUPA)," IETF, Internet-Draft draft-strassner-sup-a-generic-policy-info-model-03, Jan. 2016, work in progress, accessed 25 July 2016; <https://tools.ietf.org/html/draft-strassner-sup-a-generic-policy-info-model-03>.
- [2] A. Riekstin *et al.*, "A Survey of Policy Refinement Methods as A Support for Sustainable Networks," *IEEE COMST*, vol. 18, no. 1, 1st qtr. 2016, pp. 222–35.
- [3] A. C. Riekstin *et al.*, "Orchestration of Energy Efficiency Capabilities in networks," *J. Network and Computer Applications*, vol. 59, Jan. 2016, pp. 74–87.
- [4] S. Abeck, *Network Management: Know It All*, Morgan Kaufmann, 2009.
- [5] A. Westerinen *et al.*, "Terminology for Policy-Based Management," IETF RFC tech. rep., Nov. 2001, accessed 25 July 2016; <https://www.ietf.org/rfc/rfc3198.txt>.
- [6] D. Agrawal *et al.*, *Policy Technologies for Self-Managing Systems*, Pearson Education, 2008.
- [7] B. Moore, "RFC3640 – Policy Core Information Model (PCIM) Extensions," IETF RFC 3460, Jan. 2003, accessed 25 July 2016; <http://www.ietf.org/rfc/rfc3460.txt>.

- [8] B. Moore *et al.*, "Policy Core Information Model (PCIM) Extensios," ietf RFC 3060, Feb. 2001, accessed 25 July 2016; <http://www.ietf.org/rfc/rfc3060.txt>.
- [9] J. Strassner, *Policy-Based Network Management: Solutions for the Next Generation*, 1st ed., Morgan Kaufmann, Sept. 2003.
- [10] T. C. M. B. Carvalho *et al.*, "Towards Sustainable Networks – Energy Efficiency Policy from Business to Device Instance Levels," *ICEIS '12*, SciTePress, 2012, pp. 238–43.
- [11] K. P. Twidle *et al.*, "Ponder2: A Policy System for Autonomous Pervasive Environments," *5th Int'l. Conf. Autonomic and Autonomous Systems*, Valencia, Spain, 20–25 Apr. 2009, pp. 330–35.
- [12] S. Davy, B. Jennings, and J. Strassner, "The Policy Continuum-Policy Authoring and Conflict Analysis," *Comp. Commun.*, vol. 31, no. 13, Aug. 2008, pp. 2981–95.
- [13] N. Damianou, *A Policy Framework for Management of Distributed Systems*, Ph.D. dissertation, Imperial College London, 2002, accessed 25 July 2016; <http://pubs.doc.ic.ac.uk/policy-framework-distrib-systems/>
- [14] Y. Snir *et al.*, "Policy Quality of Service (QoS) Information Model," Nov. 2003, accessed 25 July 2016; <http://tools.ietf.org/html/rfc3644>
- [15] A. Beller, E. Jamhour, and M. Pellenz, "Defining Reusable Business-Level QoS Policies for DiffServ," *Utility Computing*, ser. Lecture Notes in Computer Science, A. Sahai and F. Wu, Eds. Springer Berlin Heidelberg, vol. 3278, 2004, pp. 40–51.

BIOGRAPHIES

ANA CAROLINA RIEKSTIN [M] (a.ca.riekstin@ieee.org) is a postdoctoral fellow at Synchromedia Laboratory, École de Technologie Supérieure, Montreal, Canada. She received her Ph.D. (2015) and M.Sc. (2012) in computer engineering from the Polytechnic School of the University of São Paulo (USP). She got her B.Sc. in computer science (2007) from the Institute of Mathematical and Computer Sciences, USP. She worked previously at the Laboratory of Sustainability in ICT (LASSU) at USP, Univesp, Microsoft Research (internship), Volkswagen do Brasil, and PromonLogicalis.

BRUNO BASTOS RODRIGUES (brodrigues@larc.usp.br) has a B.Sc. in computer science from the University of the State of Santa Catarina (UDESC) in 2013. At the moment he is an M.Sc. candidate and researcher at LASSU in the

Polytechnic School, USP. During his M.Sc., he has co-authored three patent applications in the area of network management, service virtualization, and energy efficiency.

VIVIANE TAVARES NASCIMENTO (vianetr@larc.usp.br) is an M.Sc. candidate and researcher at LASSU, Polytechnic School, USP, currently working on the Energy Efficiency Cloud (E2C) project. She graduated as an electrical engineer in 2008 from Universidade Estadual Paulista, Engineering Faculty of Ilha Solteira, and received her M.B.A. degree at the Laboratory of Network Architecture (LARC) of the Polytechnic School, USP. She worked from July 2008 to January 2013 at Itau-Unibanco Bank.

CLAUDIA BIANCHI PROGETTI (cprogetti@larc.usp.br) is a Ph.D. candidate at LASSU, Polytechnic School, USP. She received her M.Sc. in management and technology in production systems in 2014, her M.B.A. in business management with emphasis in project management in 2008, and her post graduate degree in systems analysis for applications and web solutions in 2005. Her career has developed in the area of information technology, with experience in large corporations in strategic and leadership positions.

TEREZA CRISTINA MELO DE BRITO CARVALHO (terezacarvalho@usp.br) is an associate professor and director of LASSU at the Polytechnic School, USP. She received her B.Sc. (1980), M.Sc. (1988), Ph.D. (1996), and free-docency (2012) in electronic engineering from USP. She concluded her Sloan Fellows Program (2002) as postdoctoral work at MIT. Her specialization is in local network projects in the Alfred Krupp von Bohlen und Halbach Foundation, Nürnberg Germany (1990). She has more than 90 scientific and technology publications.

CATALIN MEIROSU (catalin.meirosu@ericsson.com). Is a master researcher with Ericsson Research, Stockholm, Sweden, which he joined in 2007. He holds a Ph.D. in telecommunications (2005) from Politehnica University, Bucharest, Romania, and was a project associate of the ATLAS experiment at the Large Hadron Collider at CERN, Geneva, Switzerland during his Ph.D. He has 10 granted patents and has co-authored over 30 scientific papers. He has been involved in several research projects on SDN, cloud computing, and energy management.

Software Defined Networking, Caching, and Computing for Green Wireless Networks

Ru Huo, Fei Richard Yu, Tao Huang, Renchao Xie, Jiang Liu, Victor C.M. Leung, and Yunjie Liu

ABSTRACT

Recent advances in networking, caching, and computing will have a profound impact on the development of next generation green wireless networks. Nevertheless, these three important areas have traditionally been addressed separately in existing works. In this article, we propose a novel framework that jointly considers networking, caching, and computing techniques in a systematic way to naturally support energy-efficient information retrieval and computing services in green wireless networks. This integrated framework can enable dynamic orchestration of different resources to meet the requirements of next generation green wireless networks. Simulation results are presented to show the effectiveness of the proposed framework. In addition, we discuss a number of challenges in implementing the proposed framework in next generation green wireless networks.

INTRODUCTION

Increasingly rigid environmental standards and rising energy costs have led to great interest in improving the energy efficiency of green wireless networks [1, 2]. Recent advances in networking, caching, and computing have been extensively studied in the development of green wireless networks. In the area of networking, software-defined networking (SDN) has attracted great interest in both academia and industry. SDN introduces the ability to program the network via a logically software-defined controller, and separates the control plane from the data plane. It is beneficial to extend SDN to wireless networks [3, 4]. Software defined wireless networks enable direct programmability of wireless network controls and abstraction of the underlying infrastructure for wireless applications, with improved energy efficiency and great flexibility in green wireless network management [5].

Information-centric networking (ICN) has been extensively studied in recent years. To promote the content to first-class citizenship in the network, in-network caching is used in ICN to speed up content distribution and improve network resource utilization. In ICN, requests no longer need to travel to the content source, but are typically served by a closer ICN “content node” along the path. Information-centric wireless networks enable content caching in both the air and the mobile devices, which has been recognized as one of the promising techniques for next generation green wireless networks [6, 7].

In the area of computing, cloud computing has been widely adopted to enable convenient access to a shared pool of computing resources [8]. Cloud computing will have a profound impact on green wireless networks as well. The introduction of cloud computing to wireless mobile networks enables mobile cloud computing (MCC) systems, and cloud computing for radio access networks leads to green cloud radio access networks (C-RAN) [9]. Nevertheless, as the distance between the cloud and the edge device is usually large, cloud computing services may not provide guarantees to low-latency applications, and transmitting a large amount of data (e.g., in big data analytics) from the device to the cloud may not be feasible or economical. To address these issues, fog computing [10] and mobile edge computing (MEC) [11] have been proposed to deploy computing resources closer to end users.

Although some excellent work has been done on networking, caching, and computing in wireless networks, these three important areas have traditionally been addressed separately in existing studies. In this article, we propose to jointly consider networking, caching, and computing techniques in order to improve the performance of next generation green wireless networks. Specifically, based on the programmable control principle originated in SDN, we incorporate the ideas of information centricity originated in ICN. This integrated framework can enable dynamic orchestration of networking, caching, and computing resources to meet the requirements of next generation green wireless networks.

The rest of this article is organized as follows. We describe an overview of SDN, ICN, cloud/fog computing, and MEC. We present the proposed framework that integrates networking, caching, and computing for next generation green wireless networks. Simulation results are presented and discussed. Some open research issues are discussed. Finally, we conclude this study.

OVERVIEW OF SOFTWARE-DEFINED NETWORKING, INFORMATION-CENTRIC NETWORKING, CLOUD/FOG COMPUTING, AND MOBILE EDGE COMPUTING

In this section, we briefly introduce SDN, ICN, cloud/fog computing, and MEC.

The authors propose a novel framework that jointly considers networking, caching, and computing techniques in a systematic way to naturally support energy-efficient information retrieval and computing services in green wireless networks. This integrated framework can enable dynamic orchestration of different resources to meet the requirements of next generation green wireless networks.

This work is jointly supported by the National High Technology Research and Development Program (863) of China (No. 2015AA016101), the National Natural Science Foundation of China (No. 61501042), Beijing Nova Program (No.Z15110000315078), and the Natural Sciences and Engineering Research Council of Canada (NSERC).

Ru Huo, Tao Huang, Renchao Xie, Jiang Liu, and Yunjie Liu are with Beijing University of Posts and Telecommunications and Beijing Advanced Innovation Center for Future Internet Technology; Fei Richard Yu is with Carleton University; Victor C. M. Leung is with the the University of British Columbia.

Digital Object Identifier: 10.1109/MCOM.2016.1600485CM

SDN is a new type of network architecture, realizing the separation of the data plane and the control plane, and programming on the control plane directly. The separation of the data plane and the control plane is conducive to abstract and manage the infrastructure and resources of underlying networks.

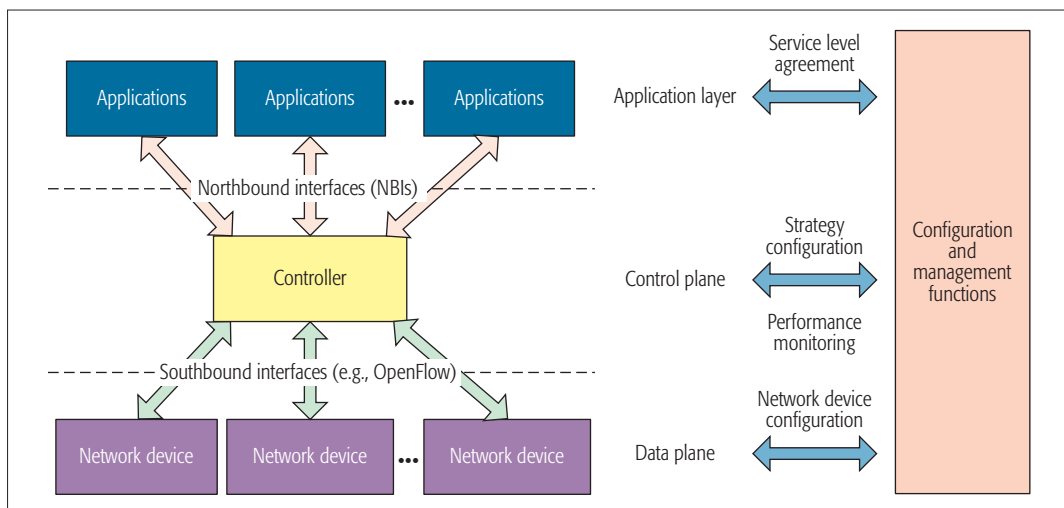


Figure 1. A typical architecture of SDN.

SOFTWARE-DEFINED NETWORKING

The idea of SDN, which originated in the Ethane project, was proposed by the Clean Slate Research Group of Stanford University. After that, first the Global Environment for Network Innovations (GENI) project conducted experimental development of OpenFlow-based SDN in campus networks and backbone-sized networks. Then some industrial companies established the Open Networking Foundation (ONF) organization, intending to promote the technology of SDN, which took OpenFlow as representative through the industry alliances. Furthermore, Google Inc. deployed the notable B4 network over its 12 global data centers based on SDN, for traffic engineering between the data centers. All these actions and practices have played important roles in a great boom in the global technology research and industrialization of SDN.

SDN is a new type of network architecture, realizing the separation of the data plane and the control plane, and programming on the control plane directly. The separation of the data and control planes is conducive to abstracting and managing the infrastructure and resources of underlying networks. Actually, SDN has a logically centralized and programmable control plane with open interfaces. Meanwhile, the control function is no longer confined to routers, or programmed and defined only by the manufacturers of equipments. Therefore, SDN achieves better flexibility and controllability.

A typical architecture of SDN is shown in Fig. 1. The control plane of SDN located in the middle of this architecture is implemented on the basis of software, and is responsible for monitoring global information and realizing network intelligence. Thus, the SDN network device acts as a logical switching device in terms of the upper applications and strategies, which greatly simplifies the control and operation of the network. Meanwhile, in terms of the underlying data forwarding plane, the network switching equipments do not have to support a large number of protocol standards, but only accept the instructions from the controller, which also fully simplifies the data forwarding plane. With OpenFlow, switching equipment and control equipment can

communicate with each other. That means the flow table, which is a core data structure in the equipment of the data forwarding plane used to implement forwarding, is programmed by SDN through the network interaction protocol. A controller will send, delete, and modify the flow table in switches of the data forwarding plane, to guide data or traffic forwarding of switches according to the regulations defined by OpenFlow. According to the items in the flow table, a packet matching the rules will be handled following the responding actions.

INFORMATION-CENTRIC NETWORKING

With a variety of applications appearing on the Internet, one of the main applications is to request massive contents. Popular contents are transmitted repeatedly on the Internet, wasting resources and reducing quality of service (QoS). In order to meet the essential behavioral patterns of the existing network, which requests and acquires information, a set of clean slate network architectures, ICN, has been proposed to change the traditional communication mode based on end-to-end communications [6]. Among these architectures, here we choose named data networking (NDN) [12] as a typical representative to introduce ICN, because it uses content name to route and retrieve, which is a direct method of content delivery. The hourglass model of the IP protocol stack is retained in NDN, but the waist layer is changed with a hierarchical content naming structure. This naming structure is similar to URLs. Meanwhile, all the routing nodes of NDN are equipped with caches or repositories. Therefore, routing, forwarding, and caching are all implemented based on flexible naming. Besides, NDN integrates security into data itself by cryptographically signing every data packet to ensure integrity and authenticity.

In the communication process of NDN, there are an interest packet and a data packet. The interest packet contains content name, and the data packet contains content name, encrypted information, and data. The router of NDN contains the content store (CS), pending interest table (PIT), and forwarding information base (FIB). The data packets are cached in the CS

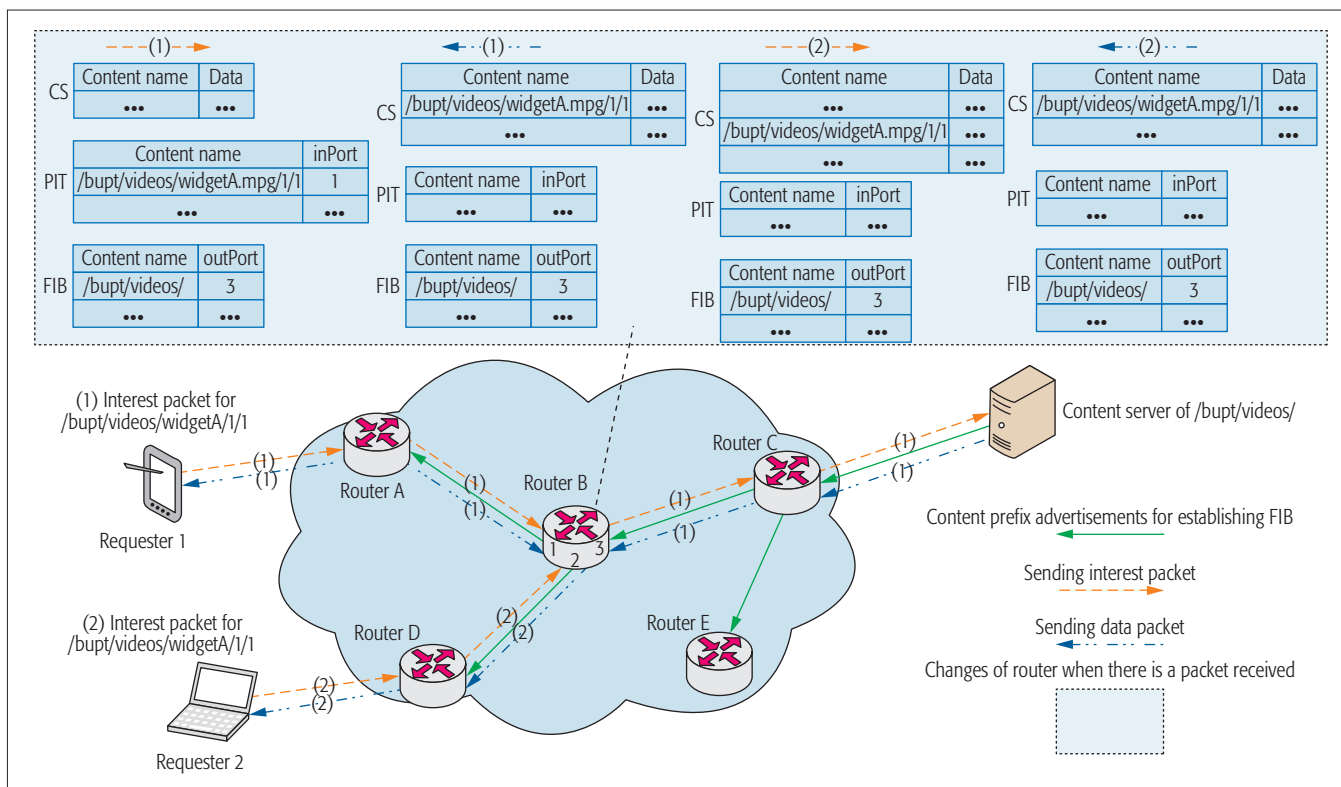


Figure 2. A typical architecture of ICN.

and reused by other relevant interest packets. The port where the interest packets originate is recorded in the PIT, and once the responding data packets arrive, the recording items are deleted. The same interest packets from different ports are recorded in one item to avoid repetitive transmitting of data, called aggregation of PIT. When content providers publish contents to the network, the routing table is established according to some rules in the FIB. A typical architecture of ICN is shown in Fig. 2. The content server first publishes advertisements, and then the routing information is established in routers. When an interest packet for */bupt/videos/widgetA/1/1* is sent from requester 1, it will be forwarded to the content server of */bupt/videos/*, and the data packet will be cached in routers A, B, and C. When the interest packet for */bupt/videos/widgetA/1/1* is sent from requester 2, it is forwarded to the content server of */bupt/videos/*. However, if there is a copy cached in router B, the data packet is sent directly from router B.

It is worth mentioning that the caching function or storage capacity in network devices is also an excellent method to avoid repetitive transmitting. Using caching or storage resource could reduce traffic redundancy and decrease delay and energy consumption, which is extremely beneficial to green applications of distributing and processing massive contents.

CLOUD COMPUTING, FOG COMPUTING, AND MOBILE EDGE COMPUTING

Cloud computing has been widely adopted to enable convenient, energy-efficient, on-demand network access to a shared pool of configurable computing resources [13]. Cloud computing

belongs to a new large-scale distributed computing paradigm and treats everything as a service (XaaS). It has different meanings to different people. For applications and users, it provides computing, storage, and applications over the Internet from centralized data centers (software as a service, SaaS). For developers, it is an Internet-scale software development platform and runtime environment (platform as a service, PaaS). For infrastructure providers and administrators, it is a massive and distributed data center infrastructure connected by IP networks (infrastructure as a service, IaaS).

Although cloud computing has been applied in diverse domains, most services need to be processed in data centers, which are usually far away from the end users. Therefore, cloud computing services may not provide guarantees to low-latency applications, and transmitting a large amount of data (e.g., in big data analytics) from the device to the cloud may not be feasible or economical. To address these issues, fog computing [10] (also called edge computing) has been proposed. The edge devices mainly refer to the local servers with weaker performance and different functions. These devices compose a distributed computing system, like personal cloud, private cloud, and enterprise cloud. Low latency and location awareness are two obvious characteristics of fog computing. Similar to the concept of edge/fog computing, MEC has recently been proposed, particularly for RANs [11], in close proximity to mobile subscribers. It is defined as running IT-based servers (called MEC servers) at the edge of a RAN, applying the concept of cloud computing, and combining computing and storage capacities with a mobile base station to accelerate contents, services, or applications.

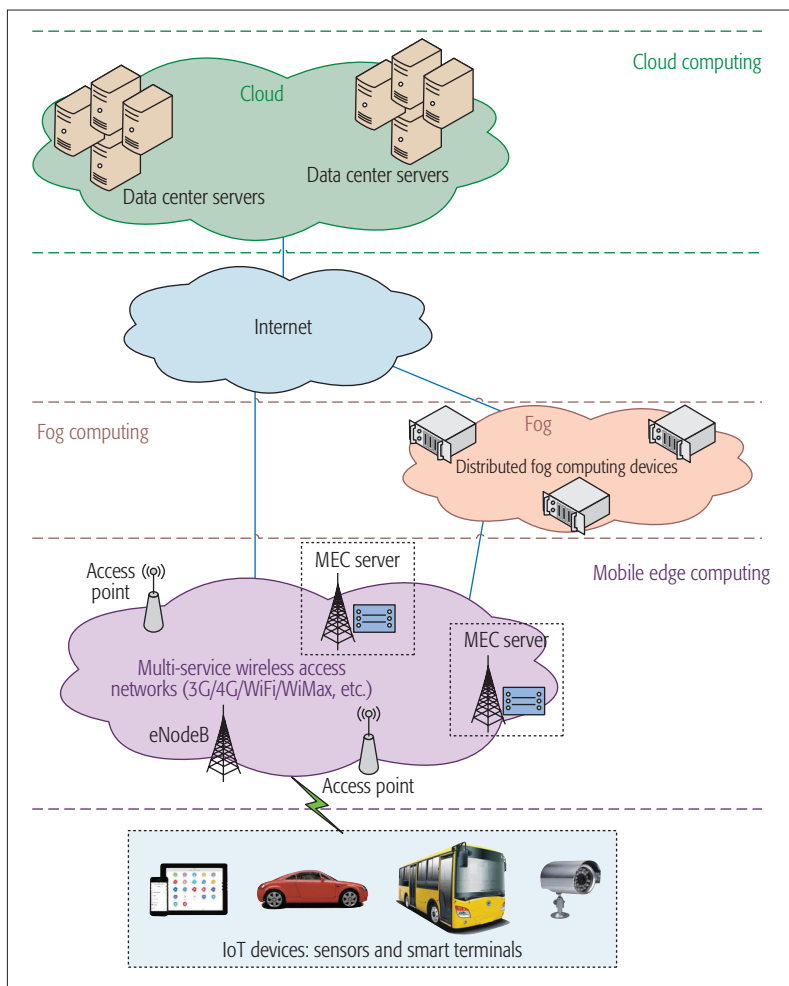


Figure 3. A typical architecture of combined cloud computing, fog computing, and mobile edge computing.

Cloud/fog computing and MEC can be widely used in the Internet of Things (IoT) [10]. A typical architecture of combined cloud/fog computing and MEC in the IoT scenario is shown in Fig. 3. Users can use cloud/fog computing and MEC depending on their situation. Energy-efficient sensors or data acquisition devices acquire and deliver the information to servers with computing and storage capacities to process.

SOFTWARE DEFINED NETWORKING, CACHING AND COMPUTING FOR NEXT GENERATION WIRELESS NETWORKS

In this section, we propose a framework that integrates networking, caching and computing for next generation wireless networks with heterogeneous access methods. We name this framework software defined networking, caching, and computing. The motivations, design, and implementation details of this framework are described as follows.

MOTIVATIONS

With the popularity of smart mobile devices, wireless networks have developed rapidly, and a variety of wireless access technologies (e.g., cellular networks and WLANs) have been

deployed. While these heterogeneous access methods are complementary to each other, there is a tendency to integrate these access technologies in next generation wireless networks. However, facing the heterogeneous access methods, network operators have difficulty managing and controlling them uniformly [14]. To tackle this issue, software defined wireless networks have been proposed to enable direct programmability of wireless network controls and abstraction of the underlying infrastructure for wireless applications, with improved energy efficiency and great flexibility in wireless network management and configuration [14].

Moreover, different from traditional wireless networks that only have networking resource, recent studies also consider developing caching and computing services in the underlying infrastructure, which will be beneficial for content retrieval and data processing for different applications. From the application's point of view, network, cache, and compute are underlying resources enabling upper layer applications. However, with a large number of underlying resources distributed in different heterogeneous wireless networks, how to manage and control these resources becomes a problem worthy of study to improve the performance of the upper layer applications.

Most existing works on wireless networks consider networking, caching, and computing separately, which could result in suboptimal performance. Therefore, it is desirable to have an integrated framework that enables the network, cache, and computing to be abstracted as a common resource pool for the upper applications. In addition, this framework could enable the resources to be scheduled based on the requirements of different applications, no matter what kind of radio access methods is used.

Therefore, in this article, we propose a novel framework that integrates networking, caching, and computing in a systematic way to naturally support data retrieval and computing services for next generation green wireless networks. Based on the programmable control principle originating in SDN, we incorporate the ideas of information centricity originating from ICN. This integrated framework can enable dynamic orchestration of networking, caching, and computing resources to meet the requirements of different applications.

FRAMEWORK DESIGN OF SOFTWARE DEFINED NETWORKING, CACHING, AND COMPUTING

Integrating networking, caching, and computing resources for next generation green wireless networks is not trivial. Such integration of heterogeneous wireless networks needs modifications of several aspects. The first one is from the aspect of communication protocols. As there are different kinds of access methods, different access methods have different communication protocols. If these three kinds of resources are integrated in heterogeneous wireless networks, the communication protocols for adapting the integration into every type of wireless network could be reconsidered. The second one is from the aspect of communication devices. The reconsideration of communication protocols leads to

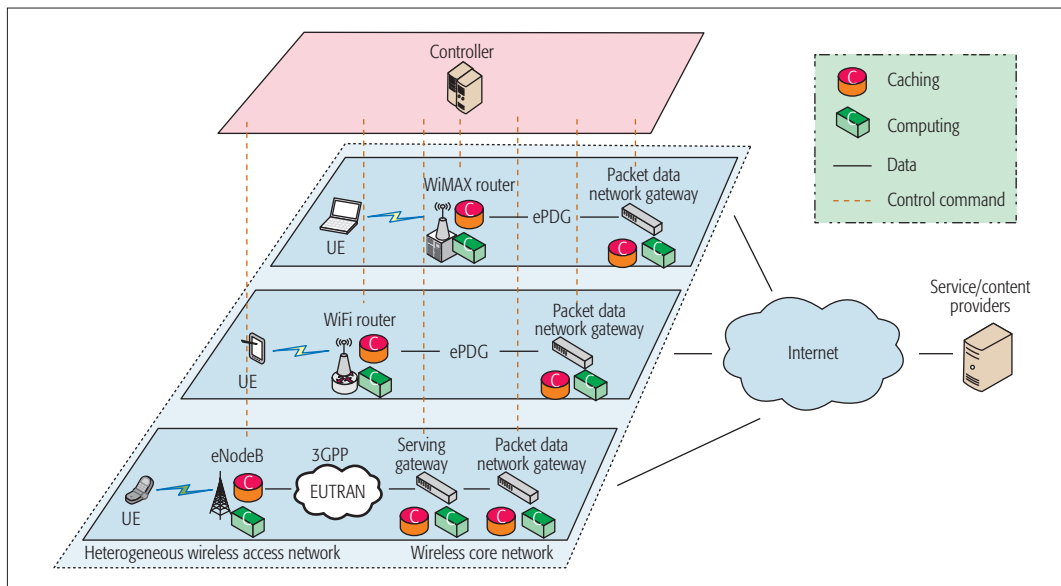


Figure 4. A framework of software defined networking, caching, and computing for next generation green wireless networks.

change of communication interfaces. The interfaces should support the interaction among the integrated resources. It means that there is a redesign of the device. The third one is from the aspect of incentives. More modifications will cause more expenses, and take a long time to complete the development. To make our integration more flexible and achievable, we adopt a software-defined approach to integrate networking, caching, and computing resources for next generation green wireless networks with heterogeneous access methods.

In this article, we choose three types of wireless access methods as examples: cellular networks, WLANs, and WiMAX networks. We equip every network node with caching and computing capabilities, and place a controller in the green wireless networks. This controller can manage the topology of the whole wireless network and conduct the packet forwarding strategies. The architecture of the proposed framework is shown in Fig. 4. Multiple user equipments (UEs) respectively access the wireless access points (APs) (e.g., eNodeB, WiFi router and WiMAX router). The controller could manage all the nodes equipped with caching and computing capacities. The caching resource is used for caching massive contents including the data collected by terminals and popular contents requested by users. Therefore, once users want to access contents or some applications want to analyze and process the data, the contents maybe hit from the in-network caching of wireless access networks or wireless core networks. This could be easy for users or applications to access contents, alleviating traffic redundancy and in-network burden. In addition, integrating computing resources is very desirable for processing the massive content or data, reducing transmission delay greatly and guaranteeing the timeliness of applications. Thus, wireless networks provide not only essential communication capability, but also caching capability and computing capability.

IMPLEMENTATION DETAILS OF SOFTWARE DEFINED NETWORKING, CACHING, AND COMPUTING

As we equip every in-network node with caching and computing capability, how to use the software defined approach to implement our system is urgent to be solved. Here we give an example implementation of our proposed framework, as shown in Fig. 5. There are some network managers who mainly administrate the controller through the program in the network applications layer, to implement tasks or deploy policies ordered from managers, such as traffic engineering or in-network computing and caching. In other words, these network applications send requests to the logical controller, and the specific behaviors and rules are designated by the control plane for the data plane. The goals of these applications could be reducing the energy consumption, enhancing QoS, improving the utilization of network resources, and achieving other optimization objectives. Therefore, these network applications could include routing algorithms, intrusion detection systems, load balancing, traffic engineering, developing ICN and MEC, and so on. The northbound interface is used for implementing network control and operation logic provided by the set of applications. Particularly, the policies defined by the applications will be translated to another kind of instructions presented below, to program the behavior of the underlying infrastructure.

The control plane can provide some basic services (e.g., monitoring network state, generating network configuration according to the policies of network applications, discovering devices, managing network topology) and important abstraction functions for the underlying infrastructure, including terminal hypervisor, switches hypervisor, network hypervisor, topology hypervisor, and so on. All the infrastructure could be abstracted as isolated resource slices. These hypervisors could manage the resource slices dynamically, realizing on-demand resource allocation. The terminal hypervisor is responsible for managing

Integrating computing resources is very desirable for processing the massive content or data, reducing transmission delay greatly and guaranteeing the timeliness of applications. Thus, wireless networks not only provide essential communication capability, but also caching capability and computing capability.

The northbound interface is used for implementing network control and operation logic provided by the set of applications. Particularly, the policies defined by the applications will be translated to another kind of instructions presented below, to program the behavior of the underlying infrastructure.

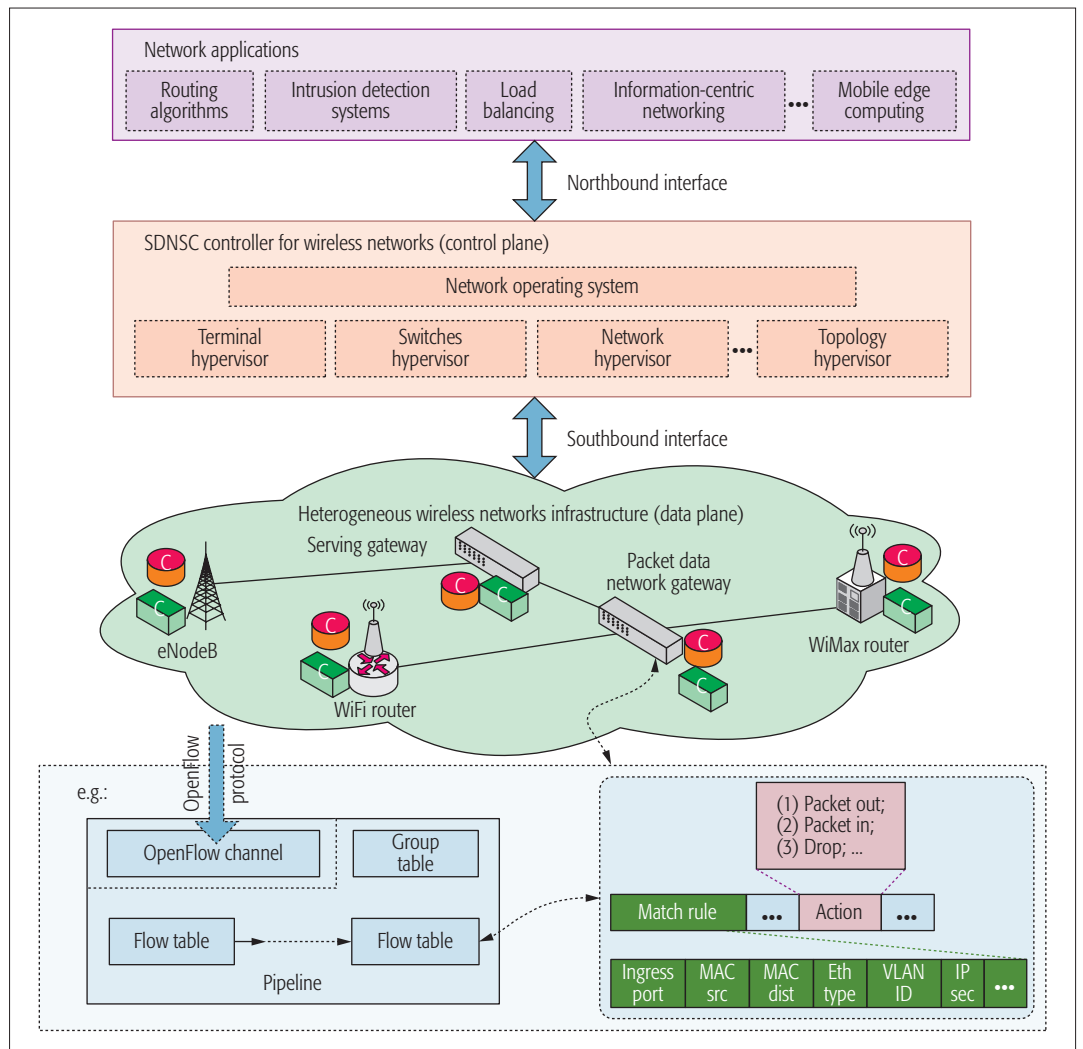


Figure 5. Implementation of software defined networking, caching and computing.

edge nodes. The switch hypervisor mainly implements and administrates the communication between switches and controller. The network hypervisor is used to monitor the networking status, such as congestion. The topology hypervisor masters all the physical nodes, links, and ports through regular monitoring. These hypervisors will map the abstracted resource slices to the physical infrastructure. Based on the information mastered by these hypervisors, the controller could implement some operations or strategies from the network applications layer, and ensure the isolation. Furthermore, the controller could guide packet forwarding of the devices in data plane, as well as perform the commands of communicating, computing, and accessing according to these information lists.

The intelligence shared between the control plane and the data plane is through the southbound interface, which enables the heterogeneous wireless communication devices to be programmed by the controller dynamically. In this article, we adopt the widely used OpenFlow protocol as the representative communication protocol of the southbound interface, which provides event-based messages, flow statistics, and packet-in messages to the controller. According to the topology information and network applica-

tions, the controller defines how packets should be handled and sends them to the flow tables. If the devices receive packets, they look up in the first table until executing the corresponding packet out action (forwarding packets to the next switch or destination) or packet in action (encapsulating the packet and sending to the controller to handle) or dropping (no rule is found), as illustrated in Fig. 5. There could be some other optional actions and detailed matching rules, and we will not describe them in detail here. With the management of the controller and the execution of devices, the packets are transmitted, cached, and computed in the green wireless networks.

SIMULATION RESULTS AND DISCUSSIONS

In this section, we present some simulation results to show the effectiveness of our proposed software defined networking, caching, and computing for next generation green wireless networks. The proposed framework can decrease latency and save energy by jointly considering all the three resources including network, cache, and computing together in the heterogeneous wireless access networks and core networks.

In the simulations, we consider one eNodeB, one WiFi router, and one WiMAX router, each of which has 10 active users randomly distrib-

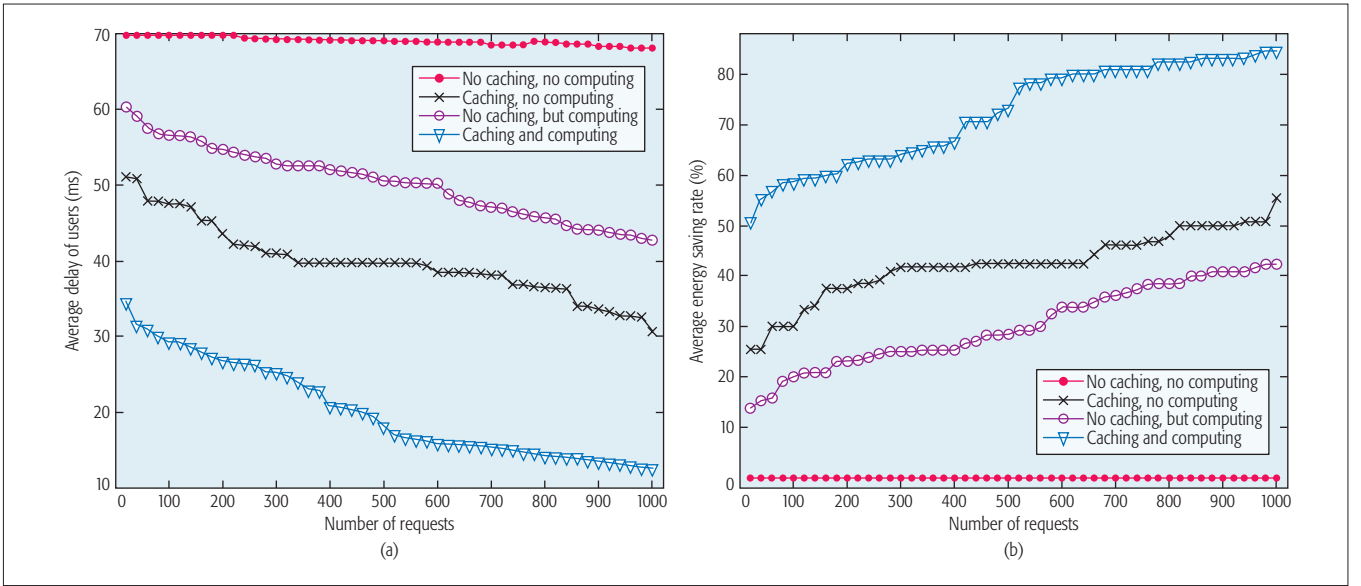


Figure 6. Performance evaluation of software defined networking, caching, and computing for green wireless networks: a) average delay of users' requests response; b) average energy saving rate (AESR).

uted. There is a server connected through the Internet with five routers, which provides caching and computing capacity on the cloud. There are 100 popular contents in the server that users request. Assume that there are totally 100 popular contents requested by all 30 active users from heterogeneous access networks, and there are totally 1000 requests for the popular contents, caching, and computing from all 30 users randomly. The link latency is assumed to be random, ranging from 10 to 20 ms per transmission hop, and the average hops count is 3 in the wired network. Here, we deploy the same cache and compute in every node of wireless networks, every cache size is 20 percent of the total content size (the caching delay is neglected), and the computational rate of the CPU F_T is set as 10^{10} cycles/s. We carry out the trace-driven simulation to evaluate the performance with the number of requests increasing, and consider the average response delay of users' requests and the average energy saving rate (AESR) as our performance metrics. Here, AESR is defined as the average ratio between the saved energy by using different resources and the consumed energy without using caching and computing resource.

The energy consumption model includes transmitting energy E_{tr} , caching energy E_{ca} , and computing energy E_{co} . The transmitting energy of user i for request k is expressed as $E_{tr}^{i,k} = (h_{i,k}P_l + 2h_{i,k}P_n)s_k$, the caching energy of user i for request k is expressed as $E_{ca}^{i,k} = W_{ca}s_kX_k^{ca}t_{ca}$, and the computing energy of user i for request k is expressed as $E_{co}^{i,k} = W_{co}s_kX_k^{co}t_{co}$. Here $h_{i,k}$ represents the hops from user i to the destination where request k can be served, P_l represents the energy density of a link (Joules per bit), P_n represents the energy density of a node to process and forward a request in wireless networks or wired networks (Joules per bit), s_k represents the size of the content or service request k wants (bit), W_{ca} represents the caching power efficiency (Watts per bit), X_k^{ca} represents whether request k is served with cache or not, t_{ca} represents the

caching time for request k , W_{co} represents the computing power efficiency (Watts per bit), X_k^{co} represents whether request k is served with compute or not, and t_{co} represents the computing time for request k . Therefore, the energy consumption can be expressed as $\sum_{AllUsers} \sum_{AllRequests} (E_{tr}^{i,k} + E_{ca}^{i,k} + E_{co}^{i,k})$. Referring to [15], we set our parameters as $P_l = 0.15 \times 10^{-8}$ J/bit, $P_n = 2 \times 10^{-8}$ J/bit, $W_{ca} = 10^{-9}$ W/bit, and $W_{co} = 2.5 \times 10^{-9}$ W/bit.

Figure 6 shows the performance of different schemes. We can observe that the traditional wireless networks that only have networking resources but no caching and computing resources have higher average delay and lower AESR compared to other schemes. Enhancing the traditional wireless networks with caching resources or computing resources can improve the system performance metrics. As there is a portion of requests for popular contents, the popular content requested before is cached in all the nodes of wireless networks, and the same request for the popular content can be accessed even at the first hop. Thus, this avoids redundant transmission traffic to reduce the delay and save the energy of using networking resource. Meanwhile, computing capability depends on F_T , and all the nodes in wireless networks can compute the tasks collaboratively to decrease the computation in the remote cloud. Our proposed framework integrates all three resources, and exhibits the best and most stable performance, especially when the distribution of all the requests comes to the steady state. Due to the integration of both caching service and computing service in wireless networks, our proposed scheme can enable joint optimization to meet the demands of users and the requesting services. With the advances in caching techniques and computing techniques, these resources can achieve better performance at a reasonable price. Since our approach can provide users a good delay experience and AESR, it can enable next generation green wireless networks.

As we jointly consider networking, caching and computing techniques in our proposed framework, it is not trivial to develop this framework in practice. It is possible that Internet service providers (ISPs) will take the responsibility to develop this framework due to the improved users experience and energy efficiency. Nevertheless, it is a significant challenge for ISPs to develop this framework.

OPEN RESEARCH CHALLENGES

Despite the potential vision of software defined networking, caching, and computing, there are some research challenges that need to be overcome. In this section, we discuss some of these challenges.

SCALABILITY

In the proposed framework, we use the software defined approach to centrally manage and control networking, caching, and computing resources. This means the controller in the control plane should supervise all the devices and resources in the data plane. Since there are various access devices, gateway devices, and network nodes in heterogeneous wireless networks, the controller has to maintain a large central database. Therefore, the performance of the controller could be degraded due to the frequent and rapid flow table update requests as well as caching and computing. In order to promote large-scale development of our scheme in real wireless networks, it is desirable to design a controller that can handle larger flow tables. Further research is needed on the design of a scalable controller in the proposed framework.

NETWORKING/CACHING/COMPUTING RESOURCE ALLOCATION STRATEGIES

In traditional SDN, the optimization problem mainly focuses on the networking resources. In contrast, in the proposed framework, there are networking, caching, and computing resources. These resources should be carefully managed to maximize the resource utilization and enhance the user experience. Therefore, it is important to design the networking/caching/computing resource allocation strategies to make a trade-off between the deployment and operation costs (e.g., energy consumption) and performance benefits (e.g., decreasing latency). Research to study the interrelationship and mutual impacts among different factors are needed, such as the relationship among the number/locations of resource nodes, energy consumption, traffic, latency, and so on. In addition, there will be different strategies corresponding to different optimization objectives.

SECURITY

The controller in the proposed framework could be attacked, resulting in a single point of failure in the system. Once an attacker acquires the permissions of the controller, the entire wireless network will be exposed to danger. Therefore, it is desirable to build an attack-tolerant system, which is designed by a fault-tolerant design approach and can operate correctly despite the existence of attacks. For instance, an attack-tolerant system may provide services meeting a service level agreement (SLA) even under an attack by triggering automatic mechanisms to regain and recover the compromised services and resources. Other descriptions used for similar themed research include survivability, resilience, trustworthy systems, and autonomic self-healing systems. How to use the software-defined characteristics to realize a tolerant system is a new direction that needs to be addressed by future research efforts.

COOPERATION INCENTIVES AMONG STAKEHOLDERS

As we jointly consider networking, caching, and computing techniques in our proposed framework, it is nontrivial to develop this framework in practice. It is possible that Internet service providers (ISPs) will take the responsibility to develop this framework due to the improved user experience and energy efficiency. Nevertheless, it is a significant challenge for ISPs to develop this framework. Therefore, it is desirable to design some cooperative incentives among different stakeholders, including ISPs and content providers. It is important to provide scalable and flexible interfaces for these stakeholders to interact and cooperate to achieve attractive benefits.

CONCLUSIONS AND FUTURE WORK

In this article, we review recent advances in networking, caching, and computing. We propose to integrate networking, caching, and computing in a systematic framework for next generation green wireless networks. We develop the architecture of the proposed framework for software defined networking, caching, and computing. We detail its key components of data, control, and management planes. Simulation results have been presented to show that our proposed framework can improve users' experience and energy efficiency. In addition, we discuss some open research challenges, including scalable controller design, networking/caching/computing resources allocation strategies, and security issues. Future work is in progress to address these research challenges.

ACKNOWLEDGMENT

We thank the reviewers for their detailed reviews and constructive comments, which have helped to improve the quality of this article.

REFERENCES

- [1] J. Wu *et al.*, "Green Communications and Computing Networks [Series Editorial]," *IEEE Commun. Mag.*, vol. 54, no. 5, May 2016, pp. 106–07.
- [2] S. Bu *et al.*, "When the Smart Grid Meets Energy-Efficient Communications: Green Wireless Cellular Networks Powered by the Smart Grid," *IEEE Trans. Wireless Commun.*, vol. 11, Aug. 2012, pp. 3014–24.
- [3] H. Hu *et al.*, "Software Defined Wireless Networks: Part 1 [Guest Editorial]," *IEEE Commun. Mag.*, vol. 53, no. 11, Nov. 2015, pp. 108–09.
- [4] K. Wang *et al.*, "Virtual Resource Allocation in Software-Defined Informationcentric Cellular Networks with Device-to-Device Communications and Imperfect CSI," *IEEE Trans. Vehic. Tech.*, vol. PP, no. 99, Feb. 2016, p. 1.
- [5] S. Zhou *et al.*, "Software-defined Hyper-Cellular Architecture for Green and Elastic Wireless Access," *IEEE Commun. Mag.*, vol. 54, no. 1, Jan. 2016, pp. 12–19.
- [6] C. Fang *et al.*, "A Survey of Green Information-Centric Networking: Research Issues and Challenges," *IEEE Commun. Surveys Tutorials*, vol. 17, no. 3, 3rd qtr., 2015, pp. 1455–72.
- [7] C. Liang, F. R. Yu, and X. Zhang, "Information-Centric Network Function Virtualization over 5G Mobile Wireless Networks," *IEEE Network*, vol. 29, no. 3, May, 2015, pp. 68–74.
- [8] M. Armbrust *et al.*, "A View of Cloud Computing," *Commun. ACM*, vol. 53, no. 4, Apr. 2010, pp. 50–58.
- [9] N. Saxena, A. Roy, and H. Kim, "Traffic-Aware Cloud RAN: A Key for Green 5G Networks," *IEEE JSAC*, vol. 34, no. 4, Apr. 2016, pp. 1010–21.
- [10] S. Sarkar, S. Chatterjee, and S. Misra, "Assessment of the Suitability of fog Computing in the Context of Internet of Things," *IEEE Trans. Cloud Computing*, vol. pp, no. 99, Oct. 2015, p. 1.
- [11] ETSI, "Mobile-Edge Computing – Introductory Technical White Paper," Sept. 2014.
- [12] L. Zhang *et al.*, "Named Data Networking," *ACM SIGCOMM Comp. Commun. Rev.*, vol. 44, no. 3, July 2014, pp. 66–73.
- [13] G. Pallis, "Cloud Computing: The New Frontier of Internet Computing," *IEEE Internet Computing*, vol. 14, no. 5, Sept. 2010, pp. 70–73.
- [14] T. Chen *et al.*, "Software Defined Mobile Networks: Concept, Survey, and Research Directions," *IEEE Commun. Mag.*, vol. 53, no. 11, Nov. 2015, pp. 126–33.
- [15] C. Fang *et al.*, "An Energy-Efficient Distributed In-Network Caching Scheme for Green Content-Centric Networks," *Computer Networks*, vol. 78, Feb. 2015, pp. 119–29.

BIOGRAPHIES

RU HUO (EthelyHuo@gmail.com) received her B.S. degree in information engineering from Harbin Engineering University, Heilongjiang, China, in 2011. She is currently working toward her Ph.D. degree in the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications (BUPT), China. From September 2015 to September 2016, she studied at the University of British Columbia (UBC), Vancouver, Canada, as a visiting Ph.D. student. Her current research interests include wireless networks, software-defined networking, information-centric networking, and resource management and allocation.

FEI RICHARD YU (richard.yu@carleton.ca) received his Ph.D. degree in electrical engineering from UBC in 2003. From 2002 to 2006, he was with Ericsson, Lund, Sweden, and a start-up in California. He joined Carleton University in 2007, where he is currently a professor. He received the IEEE Outstanding Service Award in 2016, IEEE Outstanding Leadership Award in 2013, Carleton Research Achievement Award in 2012, Ontario Early Researcher Award (formerly Premier's Research Excellence Award) in 2011, Excellent Contribution Award at IEEE/IFIP TrustCom 2010, Leadership Opportunity Fund Award from the Canada Foundation of Innovation in 2009, and Best Paper Awards at IEEE ICC 2014, IEEE GLOBECOM 2012, IEEE/IFIP TrustCom 2009, and the International Conference on Networking 2005. His research interests include cross-layer/cross-system design, security, green IT, and QoS provisioning in wireless-based systems. He serves on the Editorial Boards of several journals, and is a Co-Editor-in-Chief for *Ad Hoc & Sensor Wireless Networks*, and is a Lead Series Editor for *IEEE Transactions on Vehicular Technology* and *IEEE Communications Surveys & Tutorials*. He has served as a Technical Program Committee (TPC) Co-Chair of numerous conferences.

TAO HUANG (htao@bupt.edu.cn) received his B.S. degree in communication engineering from Nankai University, Tianjin, China, in 2002, and his M.S. and Ph.D. degrees in communication and information systems from Beijing University of Posts and Telecommunications in 2004 and 2007, respectively. He is currently an associate professor at Beijing University of Posts and Telecommunications. His current research interests include network architecture and software-defined networking.

RENCHAO XIE (Renchao_xie@bupt.edu.cn) received his Ph.D. degree from the School of Information and Communication Engineering, BUPT, in 2012. From July 2012 to September 2014, he worked as a postdoctoral researcher at China Unicom. From November 2010 to November 2011, he visited Carleton

University as a visiting scholar. He is an associate professor at BUPT. His current research interests include content delivery network, information-centric networking, and 5G networks. He has published more than 30 journal and conference papers. He has served on the Technical Program Committees (TPCs) of Chinacom 2016 and the 2012 IEEE Vehicular Technology Conference (VTC)-Spring. He has also served for several journals and conferences as a reviewer, including *IEEE Transactions on Communications*, *ACM/Springer Wireless Networks*, the *EURASIP Journal on Wireless Communications and Networking*, *(Wiley) Wireless Communications and Mobile Computing*, *IEEE Communications Letters*, 2011 IEEE GLOBECOM, and so on.

JIANG LIU (liujiang@bupt.edu.cn) received his B.S. degree in electronics engineering from Beijing Institute of Technology, China, in 2005, his M.S. degree in communication and information systems from Zhengzhou University, China, in 2009, and his Ph.D. degree from BUPT in 2012. He is currently an assistant professor at BUPT. His current research interests include network architecture, network virtualization, software-defined networking, information-centric networking, and tools and platforms for networking research and teaching.

VICTOR C. M. LEUNG [S'75, M'89, SM'97, F'03] (vleung@ece.uvc.ca) is a professor of electrical and computer engineering and holder of the TELUS Mobility Research Chair at UBC. His research is in the areas of wireless networks and mobile systems. He has co-authored more than 800 technical papers in archival journals and refereed conference proceedings, several of which have won best-paper awards. He is a Fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada. He serves on the Editorial Boards of *IEEE JSAC-SGCN*, *IEEE Wireless Communications Letters*, and several other journals. He has provided leadership to the technical program committees and organizing committees of numerous international conferences. He was the recipient of the 1977 APEBC Gold Medal, NSERC Postgraduate Scholarships from 1977 to 1981, a 2012 UBC Killam Research Prize, and an IEEE Vancouver Section Centennial Award.

YUNJIE LIU (liuyj@chinaunicom.cn) received his B.S. degree in technical physics from Peking University, Beijing, China, in 1968. He is currently the Academician of the China Academy of Engineering, Chief of the Science and Technology Committee of China Unicom, and Dean of the School of Information and Communication Engineering, BUPT. His research interests include next generation networks, and network architecture and management.

Green DataPath for TCAM-Based Software-Defined Networks

Huawei Huang, Song Guo, Jinsong Wu, and Jie Li

A TCAM-based flow table is a power-hungry hardware that can provide high-speed lookup operations for packet switching networks. A number of energy-efficient TCAM usage approaches have been proposed, but this topic still has not been well studied in the context of traffic engineering for SDN. Aiming to find energy-efficient routing paths for traffic sessions in SDN networks, the authors propose a novel Green DataPath architecture, where the dynamic voltage and frequency scaling power management technique is introduced.

ABSTRACT

A TCAM-based flow table is power-hungry hardware that can provide high-speed lookup operations for packet switching networks. A number of energy-efficient TCAM usage approaches have been proposed in recent literature, but this topic still has not been well studied in the context of traffic engineering for SDN. To this end, aiming to find energy-efficient routing paths for traffic sessions in SDN networks, we propose a novel Green DataPath architecture, where the dynamic voltage and frequency scaling (DVFS) power management technique is introduced. The dedicated SDN controlling module and DVFS-enabled switches are devised for the control plane and data plane, respectively. A framework for energy-efficient routing algorithms is also developed under the proposed architecture and evaluated by extensive simulations using three traffic scheduling schemes.

INTRODUCTION

In the last decade, ternary content addressable memory (TCAM) has become the dominant hardware, providing super high-speed forwarding operation in packet switching networks. For example, a commercial TCAM chip named R8A20410BG can support 20 Mbits density working at 360 MHz per table, which suggests that it can perform up to 360 million searches per second per table. While a TCAM has line-rate speed lookup benefits, it also comes with disadvantages such as the high cost-to-density ratio (US\$350 for a 1 Mbit chip) and high power consumption (15–30 W/Mb).

Due to these reasons, TCAMs have been limited to wild card storage in packet switching devices, and must be carefully scheduled to use. Therefore, a number of energy-efficient TCAM usage approaches have been proposed for packet switching networks in the state-of-the-art literature. These approaches can be classified into three categories: *TCAM usage reduction* [1, 2], *TCAM partial utilization* [3–5], and *forwarding rule compression* [6–10], while they are still in the premature stage in the context of traffic engineering for software defined networking (SDN).

In this article, we call the traffic flows between the data source transmitter and the client receiver a *session*. To find the energy-efficient routing paths for traffic sessions in SDN networks, we

present a novel Green DataPath architecture, where the dynamic voltage and frequency scaling (DVFS) power management technique is used. In particular, we have also devised the dedicated SDN controlling logic and DVFS-enabled switches. Under the Green DataPath architecture, a routing algorithm is proposed and evaluated under various network settings.

PRELIMINARIES

SDN

SDN simplifies network management via decoupling the control and data planes using a logically centralized network operating system called a controller. Thus, complicated controlling logics are no longer necessarily installed in packet forwarding devices such as switches or routers. Therefore, SDN has been viewed as the next generation network paradigm [11]. In SDN networks, each SDN switch at the data plane conducts data forwarding according to the flow table entries (also called rules) installed by the controller. Each forwarding rule can be expressed in the form of $\langle Match, Action \rangle$, in which the *Match* field is used to match against the packet header. If a rule is matched, the switch executes the specified actions in the *Action* field to the packet. For example, the rule $\langle Match:\{ip_nw_src = 100.0.0.1, nw_dst = 100.0.0.2\}, Action=output:3 \rangle$ indicates that the packets from a host with source IP address 100.0.0.1 and a destination IP address 100.0.0.2 will be forwarded to the *output* port 3 of the switch.

TCAM

In Ethernet networks, switches and routers must deliver bandwidth-hungry services such as voice over Internet Protocol (VoIP), IP television (IPTV), video on demand (VOD), and wireless third/fourth generation (3G/4G) with the appropriate quality of service (QoS) levels. In order to build the platforms necessary to optimally manage large amounts of network traffic quickly and effectively, system designers are increasingly relying on advanced content addressable memory (CAM), especially TCAM devices, to perform ultra-fast data packet searches.

CAM compares input search words, such as the match fields in packet headers, against a table of stored forwarding rules, and returns the address of the matched data. CAM can finish a complete lookup operation over all stored rules

in a single clock cycle. Therefore, it is popular in high throughput systems. Figure 1a illustrates an example of the lookup operation. When a packet with the source IP address 100.0.0.1 arrives at a switch, the packet header will be compared against the rule prefixes stored in CAM based table. The matched prefix, such as the shadowed one, will activate the corresponding matchline, which generates an encoding signal. After decoding such a mapping signal by Decoder, the predefined action, such as 100.0.0.1:Output 3, will be duplicated to the Action execution module. Finally, the processed packet leaves the current switch.

In general, there are two types of CAMs: binary CAM (BCAM) and TCAM. The former can be used to store full entries and perform exact 0/1 lookup against each bit of the search data, while the latter can store wildcard entries and do more beyond the binary comparison. In each wildcard entry, the “X” value, called a “don’t care” bit, can also be represented, indicating that a particular bit in the search data will not be taken into consideration when comparing with a stored rule. This feature is very useful in many applications such as the prefix matching in IP-lookup and range queries for packet classification. In order to support three states of each bit in a rule, that is, *match 0*, *match 1*, and *don’t care*, each TCAM cell requires encoding using two physical bits. For example, Fig. 1b illustrates a NOR-type-based TCAM cell, which contains two static random access memory (SRAM) cells representing two physical bits, D_0 and D_1 . Since each physical bit can represent two binary states, the combination of D_0 and D_1 can denote four logical possible states, but only three of them are required by the ternary storage. On the other hand, Fig. 1c shows the ternary encoding table for the NOR-type-based TCAM cell, where we set $D_0 = 0$, $D_1 = 1$, and $D_0 = 1$, $D_1 = 0$ to store logic ternary symbols “0” and “1,” respectively. Additionally, the cell allows searching for an “X” symbol by setting both SL_0 and SL_1 to logic “0.” This is an external “don’t care” that forces a match of a bit regardless of the stored bit. Therefore, using TCAM, a packet forwarding device can do *wildcard* lookup operations.

In the early stage of IP routers, the lookup speed was unable to match the growth of link bandwidth. TCAMs have been adopted to design high throughput forwarding engines on routers and switches [6].

Due to the realization of the logical ternary symbol, TCAMs are more expensive and consume much more circuit board space than SRAMs. In the fast lookup operation, TCAM chips also generate a large amount of heat. Therefore, a TCAM cell is far more complicated and power-consuming than an SRAM cell. For example, 1 Mbit TCAMs consume 15–30 W of power, about 50 times higher than SRAMs.

POWER CONSUMPTION OF TCAM

The power consumption of a hardware component mainly depends on the voltage supply (V_{dd}), equivalent capacitance (C_{eq}), and operating frequency (f). Most of the power consumption in TCAM is due to charging and discharging of various control lines, such as the searchlines and

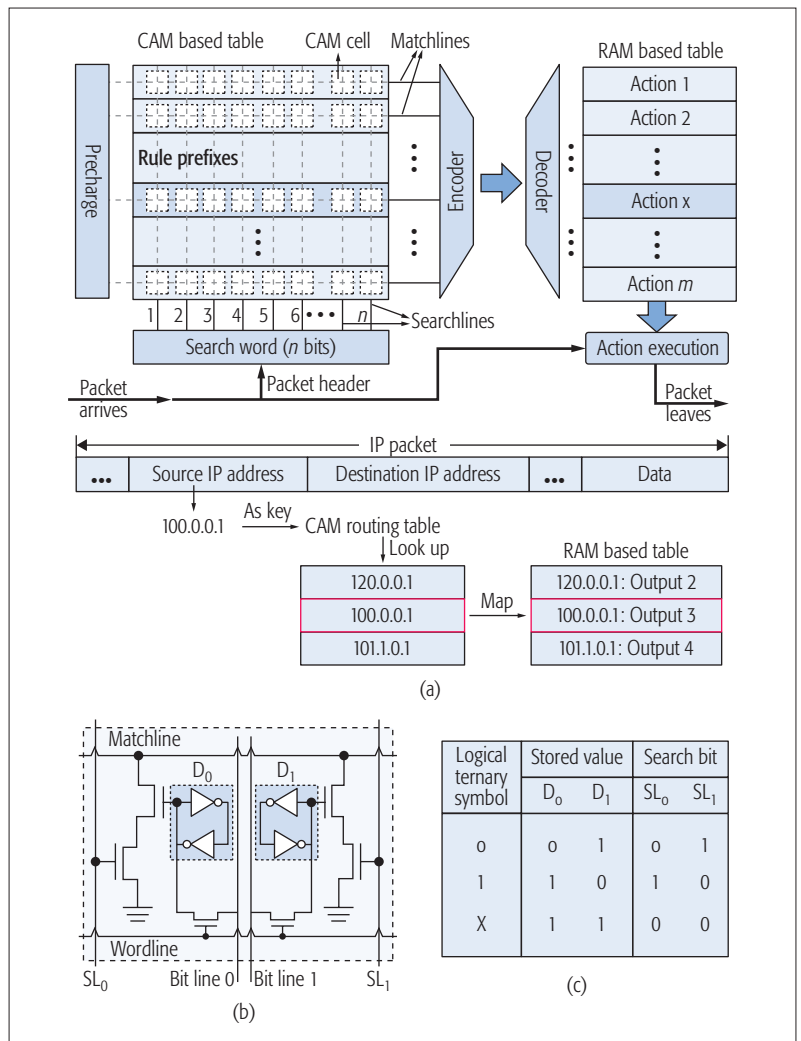


Figure 1. The rationale of CAM-based lookup operation and TCAM cell: a) rationale of packet lookup operation in a switch/router; b) a NOR-type TCAM cell; c) ternary encoding for a NOR cell.

matchlines shown in Fig. 1a. According to the TCAM power consumption model presented in [12], $P \propto V_{dd}^2 \cdot f$, we can infer that three parameters (i.e., V_{dd} , C_{eq} , and f) can be tuned directly to reduce power consumption. Based on the impact of parameter variations on circuits and micro-architecture [13], the following insights of parameter configuration are obtained regarding power reduction on a transistor:

- The capacitance of transistors in an off-the-shelf TCAM chip is already ossified and cannot be tuned.
- The variation of supply voltage takes a significant percentage of the entire supply voltage range. Too high voltage leads to unreliable processing upon packets, while too low voltage increases the probability of failure during read and write operations.
- Too low frequency leads to low performance (e.g., low processing speed), and too high frequency results in high leakage power. Note that a wide spread of reasonable frequency distribution does exist.

In summary, the reduction of power consumption can be achieved by reasonably tuning V_{dd} and f .

Once a dynamic traffic flow arrives, the expense of refreshing the current solution is significantly high or even intractable. Therefore, designing energy-efficient strategies that can handle the highly dynamic arriving flows in an online fashion is a critical challenge.

Category	Literature	Critical component(s)	Power awareness	Rule compression ratio	Dynamic rule update
TCAM usage reduction	Yamanaka [1]	Match field translator	No	Not applicable (N/A)	Good
	Congdon [2]	Signature CAM, prediction circuitry	Yes	(N/A)	Poor
TCAM partial utilization	Panigrahy [3]	ASIC based prefix indexer	Yes	(N/A)	Poor
	CoolCAMs [4]	Bit-selection logic	Yes	(N/A)	(N/A)
	Ma [5]	CAM based pre-classifier	Yes	(N/A)	Fair
Rule compression	EaseCAM [6]	Prefix aggregation and expansion techniques	Yes	Fair	Fair
	Meiners [7]	TCAM Razor approach	No	High	Poor
	Sun [8]	Tree representation	No	High	Poor
	Compact TCAM [9]	Shorter tags	Yes	Fair	Fair
	Multiplexer [10]	Rule-multiplexing scheme	No	High	Fair
DVFS-based	This article	Chip-indexer, DVFS module	Yes	(N/A)	Good

Table 1. Comparisons on the energy-efficient TCAM usage.

STATE-OF-THE-ART ENERGY-EFFICIENT TCAM USAGE

When applying energy-efficient lookup operation in physical packet-switching networks, recent related works can be generally classified into the three aforementioned categories. The remarkable properties of these existing proposals are summarized in Table 1.

TAXONOMY OF ENERGY-EFFICIENT TCAM USAGE

Category-A: TCAM Usage Reduction: In order to offload TCAM usage, Yamanaka *et al.* [1] built a “matching field translator” architecture, in which a list of exact match rules are generated for a corresponding wildcard rule in the first step, and then a controller translates exact matching fields into the source medium access control (MAC) addresses based on the correspondence between them. As a result, only the shorter rules that contain MAC addresses are necessary and can be stored in BCAM of a switch. Similarly, Congdon *et al.* [2] created a signature CAM and RAM-based packet parser, which works as prediction circuitry. According to the prediction logic results (i.e., prediction hit, incorrect prediction, and prediction miss), the TCAM utilization manners are attributed to no-TCAM, only using master-TCAM and full-TCAM usage, respectively.

Category-B: TCAM Partial Utilization: Panigrahy *et al.* [3] partitioned TCAMs into several groups first, and then used an application-specific integrated circuit (ASIC)-based hash table to perform lookup in only one TCAM chip, others remaining inactive. Similarly, Zane *et al.* [4] proposed a bit-selection logic to reduce power consumption. In the proposed architecture, TCAMs are partitioned to different blocks, and the hashing bits are selected to point to specified TCAM

subtables. Recently, Ma *et al.* [5] introduced a smart pre-classifier which classifies a packet in advance such that only a small portion of TCAM will be activated and searched for a given packet.

Category-C: Forwarding Rule Compression: Ravikumar *et al.* [6] introduced prefix aggregation and expansion techniques, aiming to activate a limited number of TCAM arrays during IP lookup. In such a way, the effective TCAM size in a router can be compacted. To address the range expansion problem of TCAM installation, Meiners *et al.* [7] considered how to generate a semantically equivalent packet classifier that requires the minimum number of rules for a given set of original TCAM entries. Using tree representation of rules, Sun *et al.*, [8] proposed a redundancy removal algorithm, which removes redundant rules and combines overlapping rules to build an equivalent and smaller rule set for a given packet classifier. Kannan *et al.* [9] used shorter tags for identifying flows than the original ones used to store the flow entries. As a result, the size of forwarding rules can be reduced. In order to efficiently use TCAM space, a rule multiplexing scheme in [10] was proposed with joint optimization on traffic engineering in SDN networks. Using this scheme, the original same set of rules deployed on each node for the whole flow of a session but toward different paths can be compacted in some particular overlapped switch nodes, such that the occupied TCAM space is reduced.

CHALLENGES

Update Rules for Dynamic Traffic Flows: As observed from the related work presented above, most existing approaches can only be operated offline.

For example, in the approaches belonging to Category-B, ASIC-based hash tables were used

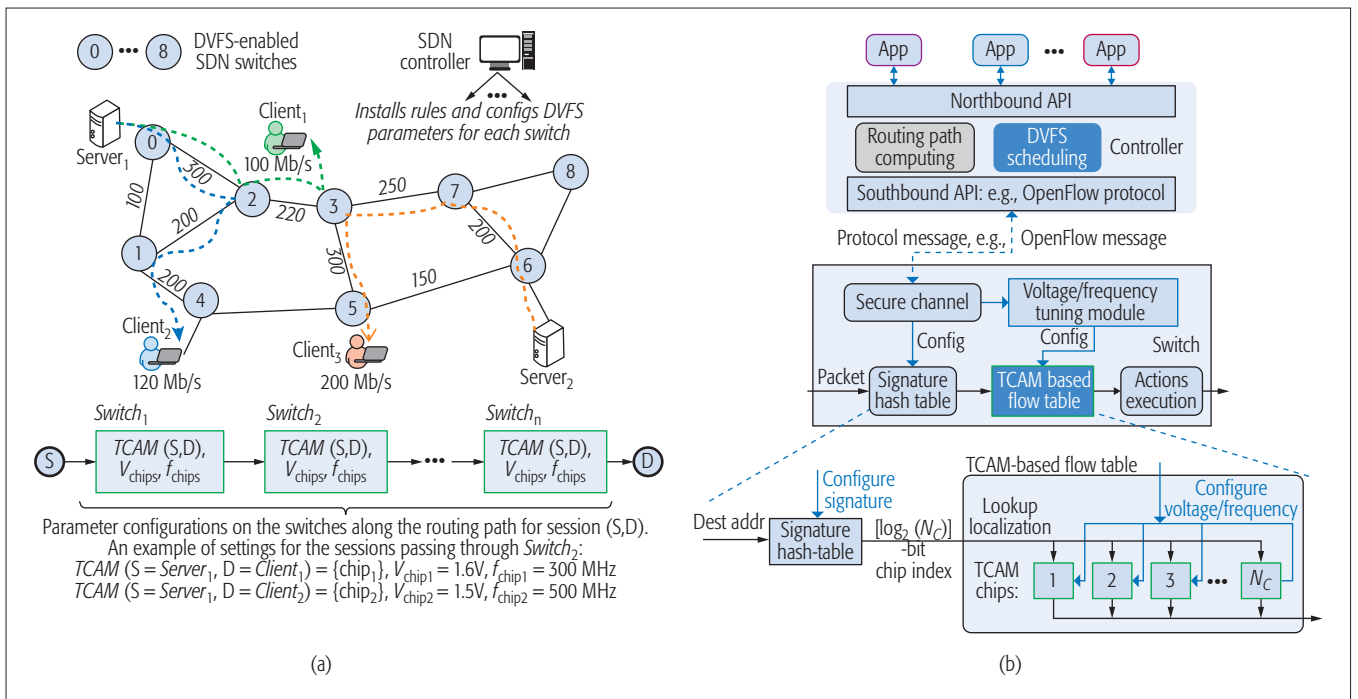


Figure 2. Green DataPath architecture: a) a use case under Green DataPath architecture; b) DVFS-enabled controller/switch.

to map the arriving packets to the specified TCAM chips [3, 4], and the indexing entries were used in a TCAM based pre-classifier [5]. However, how to update these hash-table/pre-classifier entries when new traffic flows arrive still has not been well resolved.

In practice, traffic flows usually arrive at a switch randomly and dynamically, resulting in the controller having to install time-varying rule sets on the traversed switches. Once a dynamic traffic flow arrives, the expense of refreshing the current solution is significantly high or even intractable. Therefore, designing energy-efficient strategies that can handle the highly dynamic arriving flows in an online fashion is a critical challenge.

QoS Guarantee: The other challenge is how to guarantee the QoS requirement of client flows while saving energy.

Most existing approaches enforce additional processing before lookup operations using TCAMs. For example, the method proposed in [1] requires rule set translation. Moreover, all proposals in Category-C have to execute complicated algorithms to compact the original rule set into a smaller one. All such pre-lookup operations introduce a non-negligible delay and potentially degrade the QoS performance of client flows.

GREEN DATAPATH ARCHITECTURE

In this section, we present a novel Green DataPath architecture, which is a totally different approach in the context of traffic engineering for SDNs compared to the earlier reviewed three categories. The reason is that the DVFS technique [14] is particularly introduced to the proposed architecture.

OVERVIEW

We first present an overview of the Green DataPath Architecture by displaying a use case demonstrated in Fig. 2a. We consider an SDN

network with a centralized controller and several DVFS-enabled SDN switches/routers, in which three sessions are being served under the Green DataPath architecture. When a client S requests an app service from data source D , the controller decides a path for session (S, D) along with a set of rules installed on the traversed switches. Particularly, it also specifies dynamic frequency and voltage parameter configurations for the TCAMs in each switch. When a switch receives a DVFS parameter configuration message, its embedded TCAMs shall work in the specified state through a voltage/frequency tuning module.

When the controller installs forwarding rules into the destined TCAM chip of each passing switch during a routing path, the operating frequency and voltage of the target TCAM chip need to be specified, even if this TCAM chip is empty (there are no rules being stored). The bottom figure in Fig. 2a illustrates an example of parameter configurations for two sessions, in which for the $Client_1$ oriented session, the parameter configurations toward $Switch_2$ are given as TCAM ($S = Server_1, D = Client_1$): {chip₁}, $f_{chip_1} = 300$ MHz, $V_{chip_1} = 1.6$ V. Such a configuration indicates that the forwarding rule for this session decides to install on TCAM chip₁, with an operating frequency of 300 million Hz and an operating voltage of 1.6 V. Note that TCAM could perform one packet search per cycle clock per table/chip. That is, the corresponding search speed is 300 million searches per second (MSPS) in TCAM chip₁. The frequency/voltage configuration of a TCAM chip can be tuned adaptively. Specifically, the frequency of a TCAM chip can be increased without exceeding the reliable threshold when new traffic sessions pass through. Therefore, which TCAM chip should be chosen to place the rules for each session is also a critical problem in the Green DataPath architecture.

If a TCAM chip is being installed with rules, we call it an activated chip. All activated TCAM chips are working under the scheduled voltage/frequency settings. If a match is hit, the corresponding action will be executed to the target packet, which is finally delivered out of the switch. Otherwise, the packet may be dropped or sent to a controller for further consultation.

DVFS-ENABLED CONTROLLER

In order to support our proposed architecture, a dedicated controlling logic and a DVFS tuning module need to be supported by the controller and switch, respectively. Figure 2b displays the design of the DVFS-enabled controller and switch. The logically centralized controller governs the global information of the entire network such as the connectivity map, link bandwidth, and TCAM chip occupation status. Furthermore, it implements the DVFS mechanism through the following two major modules shown in Fig. 2b.

- The routing path computing module can be implemented by a path searching algorithm, for example, Dijkstra's shortest path algorithm, aiming to find a routing path for traffic flows based on the global network overview.

- The DVFS scheduling module is responsible for generating the frequency and voltage configurations according to a certain energy efficiency policy for TCAM chips. When the first packet that indicates a new traffic session arrives at the controller via a protocol message such as the OpenFlow *packet_in* message, another function of the DVFS scheduling module is to put a unique binary signature in the specific header field of this packet and send it to the ingress switch via a protocol message like the OpenFlow *packet_out* message.

It is worth noting that a unique signature with a $\lceil \log_2(N_C) \rceil$ -bit, where N_C is the number of TCAM chips in a switch, is used to specify the target TCAM chip and can be parsed by a switch. All the packets belonging to a specific session will be directed to the indexed TCAM chip for lookup processing. We call such a pre-classification the *lookup localization*.

DVFS-ENABLED SWITCH

On the other hand, DVFS-enabled switches conduct data forwarding according to the rules installed by the controller. Corresponding to the critical components, that is, the *Signature hash-table* and *voltage/frequency tuning module*, which are shown in the bottom of Fig. 2b, we describe the relevant operations as follows.

- The forwarding rules, packet signature, as well as the voltage/frequency configurations are delivered from the controller to switches via protocol messages (e.g., the OpenFlow *packet_out* messages), and finally arrive at the *secure channel* which resides in a switch.
- The *signature hash table* stores the TCAM chip indexed hash table entries, each of which consists of the destination address prefix and $\lceil \log_2(N_C) \rceil$ -bit signature. The hash table is used to perform the *lookup localization* discussed above.
- The *voltage/frequency tuning module* is designed to assign the voltage and frequency setting for TCAM chips.
- The received rules are cached in the specified TCAM chips.
- Finally, after lookup operation, packets are processed by the *Actions execution* module.

When the first packet is returned from the controller through the *packet_out* message, the specified short destination address filed in its header will be bound with the $\lceil \log_2(N_C) \rceil$ -bit unique signature. It forms a *signature hash table*

entry and will be stored in the *signature hash table*. For example, if there are 8 TCAM chips in a switch, we need 3 bits to denote the index of each TCAM chip. If a new packet with IP destination address 10.0.0.2 sent to the controller triggers the installation of a forwarding rule, the controller may create a 3-bit signature, say 001, to specify the target installation TCAM chip of the associated forwarding rule. When this packet is received by the ingress switch through an OpenFlow *packet_out* message, the signature 001 can be parsed, and then a *signature hash table entry* $\langle \text{dest_addr}:10.0.0.2, \text{chip_index}:001 \rangle$ will be created together with the destination address and stored in the *signature hash table*. Once the successive packets belonging to the same session arrive at the hash table, they will be directed to the target TCAM chip with ID 001. Note that, such a signature hash table can be realized by an ASIC-based comparator [3] or a very small TCAM chip such as the pre-classifier proposed in [5].

After the *lookup localization*, the packet is directed to a designated TCAM chip and compared to the stored rules. Furthermore, the hash table entries can also be updated by the controller flexibly and dynamically when the assigned target TCAM chip of a session is changed according to a certain traffic engineering policy.

If a TCAM chip is being installed with rules, we call it an *activated* chip. All activated TCAM chips work under the scheduled voltage/frequency settings. If a match is hit, the corresponding action will be executed to the target packet, which is finally delivered out of the switch. Otherwise, the packet may be dropped or sent to a controller for further consultation.

OPEN ISSUES

Some open issues in our proposed Green DataPath architecture are summarized as follows.

Implementation of DVFS-Enabled Switch: In order to support DVFS working mode, new hardware modules, that is, the *signature hash table* and *voltage/frequency tuning* modules, are expected to be embedded into commercial SDN switches as the next generation of switches.

New Protocol Design: Corresponding to the proposed architecture, new supporting protocols or extensions of existing SDN protocols aim to enable SDN networks to work under energy-efficient DVFS control. For instance, the control plane requires new application programming interfaces (APIs) that can collect the DVFS parameters from switches and distribute the individual voltage/frequency configurations to each switch.

Fine-Grained Traffic Engineering: It is also desired to perform energy-aware fine-grained traffic engineering by jointly considering the DVFS parameter settings under our proposed architecture, aiming to achieve low latency, high throughput, and low power consumption.

GREEN DATAPATH FINDING PROBLEM

This section focuses on a green datapath finding problem under the proposed architecture. First, we describe this problem, and then propose a heuristic algorithm to find the energy-efficient data path for each target traffic session in the context of traffic engineering.

PROBLEM STATEMENT

We consider an SDN network $\mathcal{G} = (N, E)$, which consists of controllers (one or more), DVFS-enabled switch set N , and edge set E . Without loss of generality, as shown in Fig. 2b, we assume that all the switches are homogeneous, and each of them is equipped with N_c TCAM chips. Given the data rate requirement of a set U of traffic sessions, as well as the reliable voltage/frequency tuning ranges, a controller calculates a routing path for each of them, and decides on which TCAM chip to install the corresponding rules and the voltage/frequency configuration parameters of the active TCAM chips. Based on the system model described above, we study a *Green data path finding problem* (GDP) with the objective of minimizing the total energy consumption of all traffic sessions such that the end-to-end traffic rate constraints and TCAM chip's rule space constraints are obeyed.

ALGORITHMS

In order to solve GDP, we propose a framework of a DVFS-based algorithm that can be executed on an SDN controller. The flowchart of such an algorithm framework is shown in Fig. 3. Since the target is to save energy over all sessions while TCAM-based flow tables are being used, we fix the operating voltage of TCAM chips at the lowest reliable value and tune the operating frequency of TCAM chips in the proposed algorithm.

At first, the controller finds a candidate path set P_i using the Depth-First-Search (DFS) algorithm for each session $i \in U$, and picks up the current feasible shortest one for each session subject to the bandwidth constraints on links. Then all the sessions are sorted in a specific order according to their demanded traffic rates (increasing or decreasing) or their original arrival sequence. The corresponding traffic scheduling schemes are denoted as first-fit-increasing (FFI), first-fit-decreasing (FFD), and first-in-first-serve (FIFS). Next, each session in the sorted set U' is going to be provisioned in sequence by checking each traversed switch in the selected shortest routing path. If there is no feasible TCAM chip in the current switch, the infeasibility notification will be returned. Otherwise, assign the required forwarding rules in the first feasible TCAM chip for the current target session. Note that we call a TCAM chip feasible if its available frequency and remaining flow table space are able to hold the current target session. Afterward, the controller calculates the lowest voltage/frequency configuration that incurs the lowest power consumption to the current TCAM chip subject to the traffic rate constraints. The algorithm finishes successfully once all the sessions have been processed.

PERFORMANCE EVALUATION

This section presents the simulation results of the performance evaluation under the proposed architecture. In simulation we adopt the CORONET [15] topology, which consists of 60 switch nodes and 79 bidirectional links. Using the CORONET CONUS topology, we enforce four data source servers connecting to four switch nodes that are located in Salt Lake City, Utah; Dallas, Texas; Louisville, Kentucky; and Scranton, Pennsylvania. Then, in total, 56 traffic sessions

Using the CORONET CONUS topology, we enforce four data-source servers connecting to four switch nodes that are located in Salt Lake City, Dallas, Louisville, and Scranton. Then, in total, 56 traffic sessions are generated between server-connected switch nodes and other switch nodes.

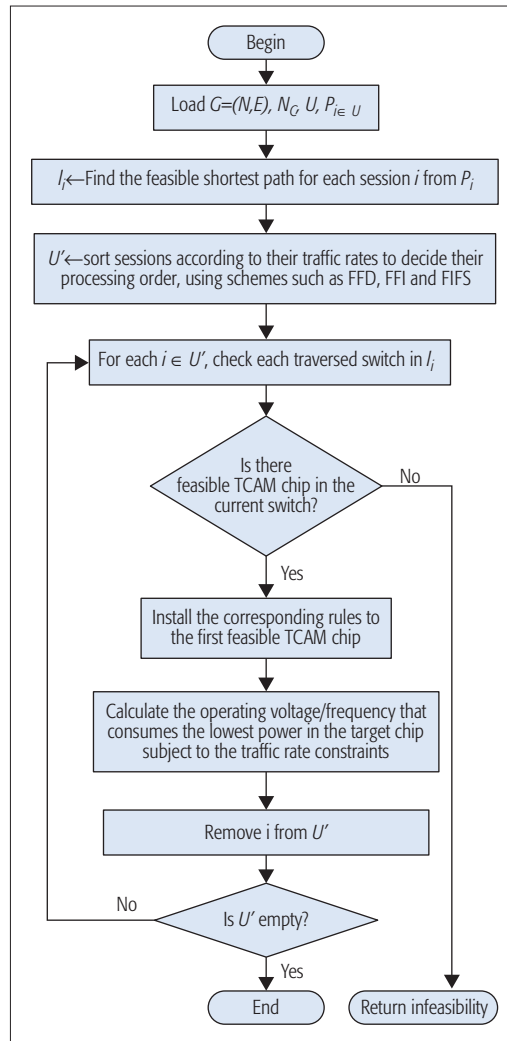


Figure 3. The DVFS-based algorithm framework.

are generated between server-connected switch nodes and other switch nodes.

The traffic rate of each session is randomly generated within the range [12, 1200] Gb/s. We assume that packets are transmitted via TCP, in which the size of each packet is 1500 bytes. In that case, if the traffic rate of a session is 1200 Gb/s, the corresponding packet rate is calculated as $(1200 \text{ Gb/s}) / (1500 \times 8 \text{ b/packet}) = 100$ million packets per second (Mpps). By invoking the DFS algorithm, we provide each session with five candidate paths. Furthermore, each session is assumed to consume only one forwarding rule in every traversed switch, and each switch is assumed to be assembled with eight TCAM chips. For each chip, we fix the reliable operating voltage as 1.5 V referring to the parameter settings shown in [13], and vary the maximum tunable operating frequency (MOP-frequency) within a reliable range. In our simulation study, four performance metrics are considered: throughput, the number of activated TCAM chips, power consumption, as well as the feasible solution ratio (FSR), which is defined as the feasible sessions over all sessions.

In the first group of simulations, we aim to compare the FSR and throughput performance among three traffic scheduling schemes (i.e., FFI, FFD, and FIFS). The MOP-frequency of each TCAM

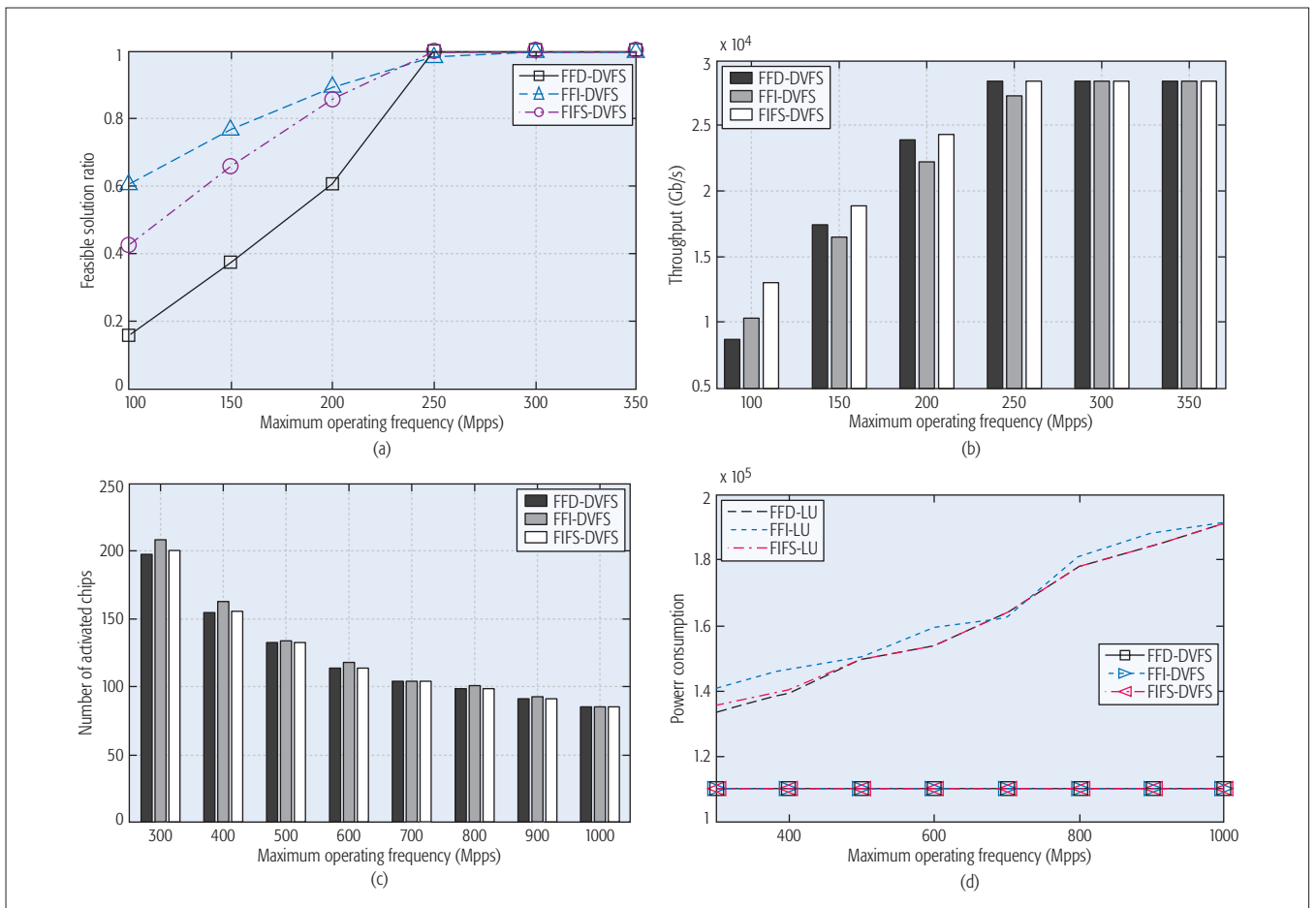


Figure 4. Performance evaluation while tuning the maximum tunable operating frequency of TCAM chips: a) feasible solution ratio; b) throughput; c) total number of activated TCAM chips; d) power consumption.

chip varies within [100, 350] Mpps or MHz. As demonstrated in Fig. 4a, the FFI always has the highest capability to find solutions for sessions, while FFD is shown to be the worst. The reason is that when the maximum operating frequency of a TCAM chip is too low, say less than 300 Mpps, it is infeasible to find a solution for the unassigned sessions once there is no available TCAM chip. In the FFD scheme, the session with the larger traffic rate has higher priority to be processed by the proposed DVFS-based algorithm. As a result, most critical TCAM frequency is consumed by a small number of large-sized sessions. Under the FFI scheme, the situation is exactly opposite. The performance of the FIFS scheme is in between.

Figure 4b shows the total throughput of the three DVFS-based schemes. It can be observed that the FIFS always has the highest throughput, and its advancement is significant especially when the MOP-frequency is less than 250 Mpps. The throughput of FFD is lower than FFI only when MOP-frequency is 100 Mpps, but higher afterward. Although we already know that FFI has a higher FSR than FFD from Fig. 4a, the accumulated traffic rate over all satisfied sessions would be lower than FFD, due to the fact that the giant-sized sessions can be satisfied with higher priority under the FFD scheme. Finally, as the MOP-frequency grows high enough, all the schemes achieve the same throughput because the feasible solution of all sessions can always be found.

Next, we increase the MOP-frequency of TCAM chips to 1000 Mpps, aiming to evaluate the number of activated chips and power consumption when all sessions can be satisfied with feasible routing solutions. We can observe from Fig. 4c that the total number of activated TCAM chips shows as a decreasing function of MOP-frequency. Although all schemes perform similarly, FFI activates slightly more chips than the other two schemes. We attribute this to the fact that the small-sized sessions are always provisioned first in each TCAM chip, and there will be more waste of frequency resource in the activated chips. As a result, more chips need to be activated to satisfy the giant-sized sessions. However, when the MOP-frequency becomes very high, the differences among schemes disappear.

Finally, to evaluate the power saving introduced by DVFS, we compare our proposed algorithms with a non-DVFS-based solution, denoted as the least unified (LU) strategy, while the MOP-frequency of a TCAM chip varies within [300, 1000] Mpps. In LU, the voltage/frequency setting in all TCAM chips is unified. Furthermore, the TCAM chip should be configured with the fewest operating settings that ensure all sessions are processed without queuing delay in switches. The power consumption under the two policies on TCAM chips are shown in Fig. 4d, based on the power consumption model given above. We can observe that the proposed

DVFS-based algorithms save power consumption by 20–40 percent, and the power consumption of the LU strategy increases with MOP-frequency in an approximately linear manner. The reason is that the total power consumption is only determined by the summation of traffic rates under the proposed algorithms. In contrast, the frequency capability of each activated TCAM chip will be almost fully exploited when the MOP-frequency increases from 300 Mpps to 1000 Mpps under the LU strategy. We also notice that the aggregated traffic rate in the switch nodes near servers is significantly higher than the averaged workload that is assigned to other switch nodes. In order to satisfy the traffic rate requirement in the heavy-loaded nodes, the *lowest* frequency must be tuned to a very high value, thus increasing the overall power consumption under the unified setting policy.

In summary, we have the following observations:

- Too low frequency settings may degrade the feasible solution ratio and the total throughput.
- The number of the activated TCAM chips decreases while increasing the chip's operating frequency.
- The power consumption does not grow even when the maximum tunable operating frequency is large enough under the proposed Green DataPath architecture.

CONCLUSION

In this article, a novel Green DataPath architecture has been proposed to find energy-efficient routing paths for traffic sessions. We focus on how to achieve energy-efficient traffic engineering in the TCAM-based SDN networks via introducing the DVFS power management technique. We have designed a DVFS-enabled Green DataPath architecture for SDN networks, where the dedicated SDN controlling a DVFS scheduling module and DVFS-enabled switches is presented. Then, for the proposed architecture, an energy-efficient data path finding algorithm framework has been proposed. Finally, we have evaluated this framework under three traffic scheduling schemes (i.e., FFD, FFI and FIFS) by extensive simulations. Some useful insights of parameter settings have been discussed and summarized.

ACKNOWLEDGMENT

This work was partially supported by Grant-in-Aid for Japan Society for the Promotion of Science (JSPS) Fellows, Grant Number 16J07062.

REFERENCES

- [1] H. Yamanaka *et al.*, "Openflow Networks with Limited I2 Functionality," *Proc. 13th Int'l. Conf. Networks*, 2014, pp. 221–29.

- [2] P. T. Congdon *et al.*, "Simultaneously Reducing Latency and Power Consumption in Openflow Switches," *IEEE/ACM Trans. Networking*, vol. 22, no. 3, 2014, pp. 1007–20.
- [3] R. Panigrahy and S. Sharma, "Reducing Tcam Power Consumption and Increasing Throughput," *Proc. 10th Symp. High Performance Interconnects*, 2002, pp. 107–12.
- [4] F. Zane, G. Narlikar, and A. Basu, "Coolcams: Power-Efficient tcams for Forwarding Engines," *Proc. IEEE 22nd Annual Joint Conf. IEEE Computer and Commun.*, vol. 1, 2003, pp. 42–52.
- [5] Y. Ma and S. Banerjee, "A Smart Pre-Classifier to Reduce Power Consumption of Tcams for Multi-Dimensional Packet Classification," *Proc. ACM SIGCOMM 2012 Conf. Applications, Technologies, Architectures, and Protocols for Computer Commun.*, 2012, pp. 335–46.
- [6] V. Ravikumar, R. N. Mahapatra, and L. N. Bhuyan, "Easecam: An Energy and Storage Efficient Tcam-Based Router Architecture for IP Lookup," *IEEE Trans. Comp.*, vol. 54, no. 5, 2005, pp. 521–33.
- [7] C. R. Meiners, A. X. Liu, and E. Torng, "Tcam Razor: A Systematic Approach Towards Minimizing Packet Classifiers in Tcams," *Proc. IEEE Int'l. Conf. Network Protocols*, 2007, pp. 266–75.
- [8] Y. Sun and M. S. Kim, "Tree-Based Minimization of Tcam Entries for Packet Classification," *Proc. 7th IEEE Consumer Commun. and Networking Conf.*, 2010, pp. 1–5.
- [9] K. Kannan and S. Banerjee, "Compact Tcam: Flow Entry Compaction in Tcam for Power Aware SDN," *Proc. Distrib. Comp. and Networking*, Springer, 2013, pp. 439–44.
- [10] H. Huang *et al.*, "Joint Optimization of Rule Placement and Traffic Engineering for Qos Provisioning in Software Defined Network," *IEEE Trans. Computers*, vol. 64, no. 12, 2015, pp. 3488–99.
- [11] N. McKeown *et al.*, "Openflow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Comp. Commun. Rev.*, vol. 38, no. 2, 2008, pp. 69–74.
- [12] K. Pagiamtzis and A. Sheikholeslami, "Content-Addressable Memory (Cam) Circuits And Architectures: A Tutorial And Survey," *IEEE J. Solid-State Circuits*, vol. 41, no. 3, 2006, pp. 712–27.
- [13] S. Borkar *et al.*, "Parameter Variations and Impact on Circuits and Microarchitecture," *Proc. 40th Annual Design Automation Conf.*, 2003, pp. 338–42.
- [14] D. C. Snowdon, S. Ruocco, and G. Heiser, "Power Management and Dynamic Voltage Scaling: Myths and Facts," *Proc. 2005 Wksp. Power Aware Real-Time Computing*, vol. 12, 2005.
- [15] A. L. Chiu *et al.*, "Architectures and Protocols for Capacity Efficient, Highly Dynamic and Highly Resilient Core Networks," *IEEE/OSA J. Optical Commun. Net.*, vol. 4, no. 1, 2012, pp. 1–14.

BIOGRAPHIES

HUAWEI HUANG received both his Bachelor's and Master's degrees in computer science and technology from the China University of Geoscience, Wuhan, in 2011 and 2013, respectively. He is currently a Ph.D. student at the University of Aizu, Japan. His research interests are in software-defined networking, network functions virtualization, and wireless networks. He is now a research fellow of the Japan Society for the Promotion of Science.

SONG GUO [M'02, SM'11] received his Ph.D. degree in computer science from the University of Ottawa, Canada. He is a full professor at the School of Computer Science and Engineering, University of Aizu. His research interests are mainly in the areas of wireless communication and mobile computing, cloud computing and networking, and cyber-physical systems. He serves as Associate Editor of *IEEE TPDS* and *IEEE TETC*. He is a Senior Member of ACM.

JINSONG WU is an associate professor in the Department of Electrical Engineering at the Universidad de Chile, Santiago. He is the founding Chair of the IEEE Technical Committee on Green Communications and Computing. He is an Editor of the *IEEE JSAC* Series on Green Communications and Networking. He was the leading editor and co-author of the comprehensive book *Green Communications: Theoretical Fundamentals, Algorithms, and Applications* (CRC Press, 2012).

JIE LI [M'94, SM'04] received his Dr. Eng. degree from the University of Electro-Communications, Tokyo, Japan. He is currently a professor at the University of Tsukuba, Japan. His current research interests include mobile distributed computing and networking, big data and cloud computing, the Internet-of-Things, operation systems, and modeling and performance evaluation of information systems. He is a Senior Member of the ACM and a member of the Information Processing Society of Japan.

We also notice that, the aggregated traffic rate in the switch nodes near servers is significantly higher than the averaged workload that is assigned to other switch nodes. In order to satisfy the traffic rate requirement in the heavy-loaded nodes, the least frequency must be tuned to a very high value, thus increasing the overall power consumption under the unified setting policy.

Toward the Development of a Techno-Social Smart Grid

S. N. Akshay Uttama Nambi and R. Venkatesha Prasad

SG envisions developing user-centric distributed systems to offer cost-effective and reliable power supply. Its effectiveness depends highly on consumer awareness and engagement. SG deployments and programs have been found to be lacking in active consumer participation. We address this by proposing a TSSG wherein technologies related to energy infrastructure interface with social activities of consumers.

ABSTRACT

Advancement in communication and computing technology is driving the next-generation electrical smart grid, SG. SG envisions developing user-centric distributed systems to offer cost-effective and reliable power supply. Its effectiveness depends highly on consumer awareness and engagement. SG deployments and programs have been found to be lacking in active consumer participation. We address this by proposing a *techno-social* framework for SG, TSSG, wherein technologies related to energy infrastructure interface with social activities of consumers. Social interactions play a crucial role in preferences of consumers and the decisions they make. The rapid evolution of social networks now enables us to model and capture these interactions. The proposed framework combines social networks with energy networks to understand individual and collective behavior of consumers in order to change the energy demand patterns. We describe several mechanisms to enable harnessing useful data from such a framework, including its applicability to various SG applications. Specifically, we illustrate the benefits of collective modeling of techno-social aspects by developing goal-oriented virtual communities. This is one of the first articles to consider both energy consumption information and characteristics of consumers to determine such communities. We employed data from a real-world SG deployment with more than 4000 households along with their preferences, opinions, and interests to evaluate our proposal.

INTRODUCTION

Smart grids (SGs) take advantage of information and communication technologies (ICT) to integrate power infrastructure with information infrastructure [1]. To increase energy efficiency and sustainability, SG assembles many features such as advanced metering infrastructure (AMI), demand response (DR), demand side management (DSM), demand forecast (DF), and emergency management (EM). Smart meters are widely being deployed to monitor energy consumption of consumers. Bidirectional communication between these devices and energy companies (energy utilities or utilities) enable immediate feedback to consumers on power usage, power quality, and pricing details. The

current literature addresses dynamic pricing to provide incentives to users to balance their energy demands matching the generation. Renewable energy sources are also a part of SG to encourage generation of energy and selling surplus [2], thereby converting consumers into prosumers. Ultimately, these techniques should help lower carbon emissions, smooth peak demands, and increase usage of renewable resources. The effectiveness and adoption of these techniques highly depend on consumer awareness, participation, and engagement. Prevalent SG deployments and programs have been found to be lacking in consumer awareness and engagement [3]. Hence, understanding *what* consumers want and *how they behave* is fundamental for developing a sustainable future-proof SG.

Energy utilities need to develop innovative energy services to *understand* how the energy supply is perceived by consumers and *engage* them to actively participate in the functioning of the grid. Current research efforts [4, 5] try to design feedback mechanisms to promote awareness of energy consumption by providing detailed energy breakdown and appliance-specific energy/cost details.

While these mechanisms are necessary and valuable, it is not sufficient to motivate pro-environmental behavior. Hence, along with energy consumption characteristics, consumer preferences (pro-environmental, pro-reduction, pro-behavioral change) and their social context (social ties, influence, and relationship) need to be considered during the development of SG programs.

Hitherto, a growing number of efforts have been applying the principles of social network analysis to home energy management. These mechanisms are aimed at revolutionizing the understanding of energy usage characteristics and helping lessen the impact on the environment.¹ Rich data from social networks capturing human interactions has closed an important loop [6, 7]. This has enabled us to model social interactions and use these models in the design of SG programs to develop robust and consumer-centric services.

Most of the current studies do not consider the interplay between the social context of consumers and their energy information in developing consumer-centric services. In contrast, we try to fill the gap by modeling and analyzing the social context of consumers along with the energy

¹ C. Chima. <http://mashable.com/2011/02/08/smart-grid-social-media/>

network. We propose a techno-social framework to model both technological and social aspects of the SG. The framework determines *how* the social context of consumers can be obtained and *which* social network models can be used along with energy consumption details to develop consumer-centric services. For example, aspects such as interests and preferences of consumers can be deduced from social networks through posts, comments, likes, products purchased, and so on. These aspects along with consumer energy usage patterns can be used by utilities to target specific *groups* of consumers in DR programs to either reduce or shift their usage. We also see that social contexts of consumers constantly change and evolve (e.g., new friends, change in relationships, preferences). Modeling the technological and social aspects jointly is a challenging task as the models developed should capture the *dynamics*, not only energy consumption patterns but also the social context of consumers.

We propose a *techno-social framework for smart grids* (TSSG), where infrastructure composed of various technologies and social aspects are studied together. A social network overlay is proposed to capture the behavior and preferences of consumers vis-à-vis SG. The framework utilizes traditional social network models [8] to analyze this along with energy consumption profiles. We illustrate the benefits of modeling the techno-social aspects by forming communities directed toward particular goals. The novelty in formation of communities lies in fusing the technological and social data. These communities can now be targeted to promote energy awareness, and provide tailored recommendations and community-specific tariffs.

The main contributions of this article are:

- A novel TSSG to enable effective energy coordination, management, and awareness amongst consumers
- Showing advantages of having a social network overlay for different SG programs to support active participation of consumers
- Illustration of goal-oriented community formation by fusing the techno-social data from a real SG deployment with more than 4000 households

To the best of our knowledge we are the first to show the interaction between energy consumption and consumer characteristics to determine communities for targeted recommendations.

TECHNO-SOCIAL SMART GRIDS

The techno-social framework integrates physical (energy network), cyber (ICT network), and social dimensions of the SG (Fig. 1). The ICT network comprises IEEE 802.15.4 (ZigBee), 802.11 (WiFi), and 802.16 (WiMAX) technologies for communication between consumers and utilities [9]. We now go further into the details of individual components of this framework.

TSSG FRAMEWORK

The proposed TSSG framework consists of three layers: *physical grid*, *smart grid*, and *social grid* (Fig. 2).

Physical Grid: An interconnected network for delivering electricity to consumers. The physical grid consists of several power generation stations,

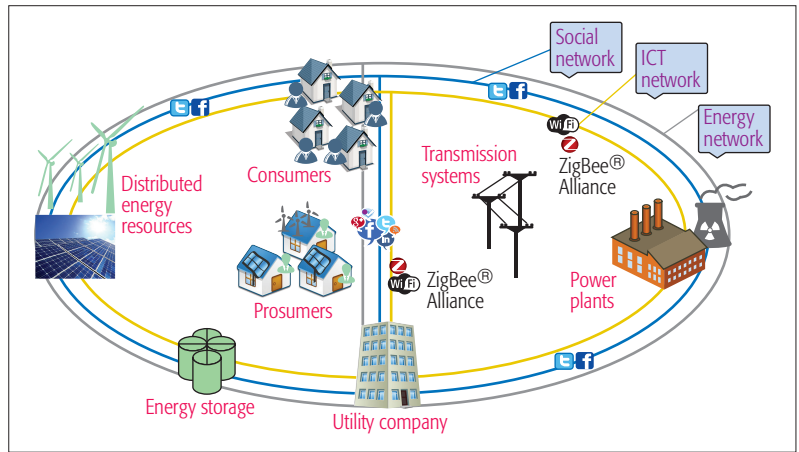


Figure 1. Techno-social framework of the SG with its entities.

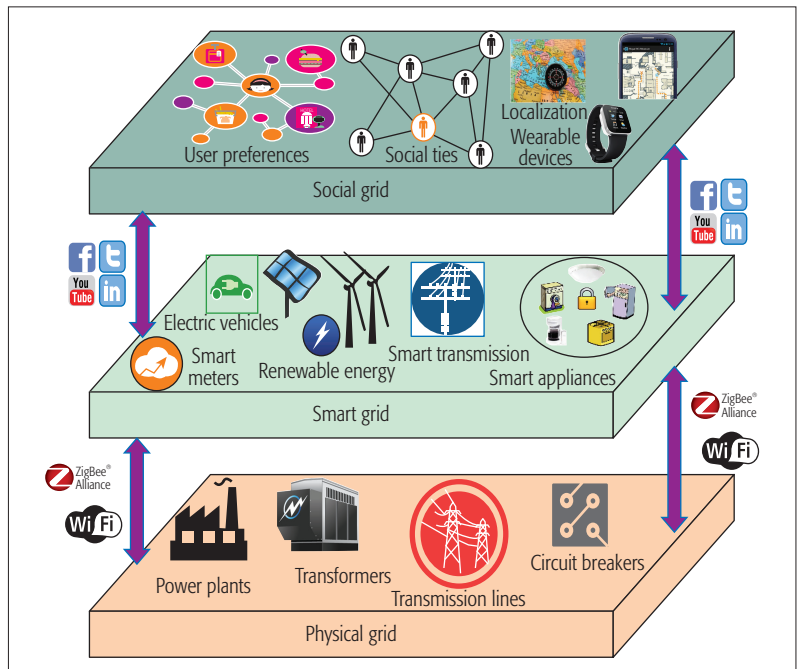


Figure 2. Techno-social framework for SG.

power plants, high-voltage transmission lines to carry power to the consumers, and distribution lines to interconnect various entities of the grid. The different actors in physical grid include transmission system operators (TSOs) and distribution system operators (DSOs) for operating, maintaining, and developing the grid. Current research on physical grid involves improving generation and transmission techniques.

Smart Grid: An intelligent power system that uses ICT to enhance efficiency, reliability, and sustainability of power generation and distribution networks [1]. SG utilize several mechanisms to curtail and balance load, flatten peak demands, automate load control, apply adaptive pricing, and bring awareness to its consumers. The various actors in SG are energy utilities, consumers, and prosumers who utilize technologies such as integrated real-time communication, smart meters, phasor measurement units, advanced load control, and virtual power plants. Current research involves demand response mechanisms with dynamic pricing, and integra-

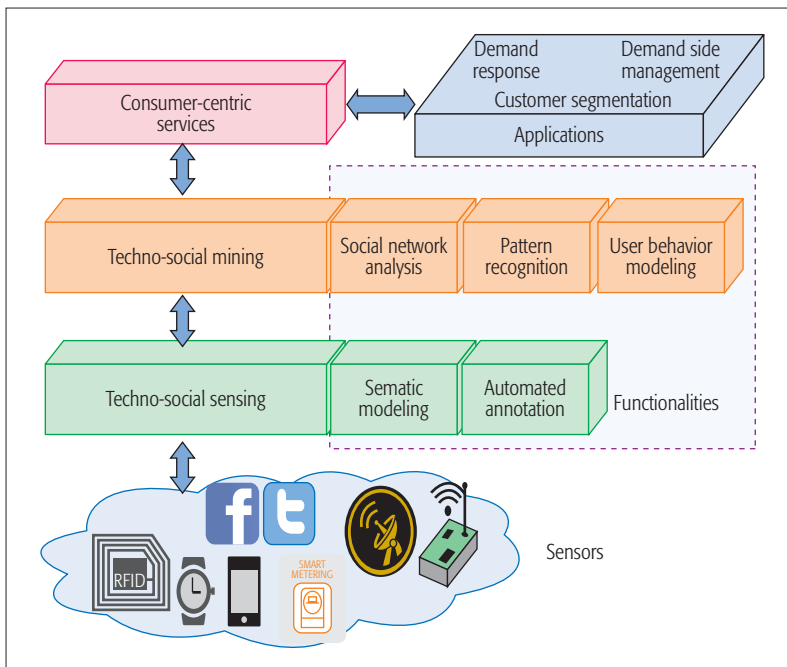


Figure 3. Interactions of core components of techno-social SG.

tion of renewable resources and electric vehicles into the grid.

Social Grid: Promotes interactions among different SG entities to support coordination of energy consumption, to trade energy between prosumers and consumers, and to promote awareness. With the increased growth in deployment of embedded sensors in the environment, the social grid can now accurately monitor consumers in different dimensions [6, 10]. In social grid, third-party services such as Facebook, Twitter, LinkedIn, and Google, along with ubiquitous sensors play a crucial role in collecting consumer social activities.

The TSSG framework supports development of several consumer-centric services like personalized recommendation, forming communities with similar consumer characteristics and energy profiles, and so on, by modeling and analyzing the data from the techno-social ecosystem. The introduction of social overlay on SG supports two-way information flow:

- TSSG derives information about consumer norms, preferences, ties, and interactions to understand the perception of consumers regarding energy pricing, demand reduction, and so on.
- TSSG analyzes these preferences to derive best practices toward sustainable behavior and disseminates them using the social overlay.

Moreover, it is easy to see that this feedback itself could be used to gather more inferences and to understand whether consumers follow certain practices. Further, it is easy to gather which feedback information was useful or followed, and so on. This helps in further improvement of information dissemination. Furthermore, the proposed TSSG framework is decentralized, that is, it is implemented at a neighborhood/community level, where techno-social data from SG and social grid are available.

² Social media APIs: <http://www.programmableweb.com/category/social/apis?category=20087>

CORE COMPONENTS

To enable interactions between the SG and social grid layer, several challenges need to be addressed in gathering and modeling both energy and consumer data. Some of these challenges are:

- Processing the raw, fragmented, and unstructured data collected from the techno-social ecosystem
- Analyzing and modeling multidimensional data from different data sources
- Adapting to changing and evolving social context (preferences, relationships, and ties) of consumers over time
- Striking a trade-off between quality of collected data and privacy
- Deducing aggregated behavior of consumers to promote collective awareness
- Adjusting to temporal dynamics of the techno-social data at different resolution (hourly, daily, monthly, seasonal)

To address these challenges, in this article, we define two core components: *techno-social sensing* and *mining*. Figure 3 shows the core components of the TSSG framework and their interactions toward development of consumer-centric services.

Techno-Social Sensing: Techno-social sensing is responsible for collectively harvesting data from the SG and social grid layers. From the *smart grid layer* data regarding energy consumption at households, appliance-specific consumption, power factor, voltage, and current are collected. From the *social grid layer* data regarding consumer activities, social ties, behavior, interactions, preferences, and opinions are collected with the help of wearable devices, smartphones, RFID tags, online social networks, GPS logs, and Internet of Things (IoT) sensors embedded in the environment [11]. Furthermore, several participatory sensing and social network application programming interfaces (APIs) can now be used to search for and gather information about consumers. For example, social media such as Facebook, Twitter, Google+, Pinterest, and Youtube has a wealth of information on preferences of consumers. Data collection from these networks is now possible with the help of open-source APIs.²

Techno-Social Mining: Techno-social mining aims to develop models to understand consumer behavior by identifying underlying patterns, rules, and beliefs from the data collected by techno-social sensing. At the *smart grid layer*, user-specific energy consumption profiles, appliance energy profiles, cost-aware appliance usage schedule, load shifting strategies can be derived by applying different pattern recognition and machine learning algorithms. At the *social grid layer*, consumer behavior, beliefs, and social ties can be derived by mining consumers' content on social media using social network analysis. Many ways have been proposed for social mining, for example, social network analysis, social media mining, and sentiment analysis. Recently, *sentiment analysis* is gaining popularity to analyze consumer sentiments. For example, sentiment analysis on consumers' posts on Facebook and Twitter can help in determining the sentiment of a user toward a technology (electric vehicles, energy reduction, sustainable energy usage, etc.). We now illustrate how consumer preferences on some of the energy related topics can be derived

using sentiment analysis across social networks. Sentiment analysis aims to identify and extract subjective information from data shared across social networks. We utilize open source sentiment analysis mechanisms to determine positive or negative sentiments with respect to energy related topics across various social networking sites. Figure 4 shows the sentiments of several consumers (more than 1000) over a week toward green energy campaigns.

Figure 4a shows the percentage of positive and negative sentiments associated with “sustainable energy” on Twitter for a week. Among the tweets related to sustainable energy, 85 percent had positive sentiment and 15 percent had negative sentiment. Similarly, Fig. 4b shows the strength, sentiment, passion, and reach for three topics: energy reduction, sustainable energy, and electric vehicles. Electric vehicles have a strength of 50 percent — likelihood of it being discussed in social media; sentiment of 15:1 — ratio of positive to negative sentiments; passion of 75 percent — likelihood of repetition of the topic; and reach of 11 percent — range of influence of the topic in social media. Figure 4c shows the classification of tweets corresponding to various sentiments associated with energy reduction on Twitter. Here a tweet is classified into one of the 20 categories derived from social behavior models. Figure 4d shows the percentage of posts with positive, neutral, and negative sentiments across social networking sites such as Facebook, Google+, Reddit, and news/blogs for energy reduction. Several metrics have been proposed in the literature to derive accurate and reliable sentiments over several days. Techno-social mining uses these to derive preferences of consumers, which can be employed to develop consumer-centric energy services.

ROLE OF TSSG IN CONSUMER-CENTRIC SERVICES

This section describes how TSSG can be employed toward developing consumer-centric SG programs.

CONSUMER AWARENESS

Smart meter data, when analyzed effectively, can help consumers reduce both energy consumption and cost. In-house displays, interactive games, and smartphone applications are being proposed to increase awareness among consumers. However, recent research efforts show that this approach fades away after a few initial weeks of deployment [5]. With constant increase in number of consumers using social networks (SNs), information propagation on SNs is a promising approach to seek responses from consumers, interact with them, and also disseminate information [10]. The proposed TSSG framework enables smart meters to directly connect to social networks and provide energy consumption information to online social networks. The role of TSSG is to provide the energy consumption information to consumers using SN in an unobtrusive way. Furthermore, SN models can be utilized to analyze who is the most influential consumer, how information spreads among consumers, and consumer-interest-based groups in SNs.

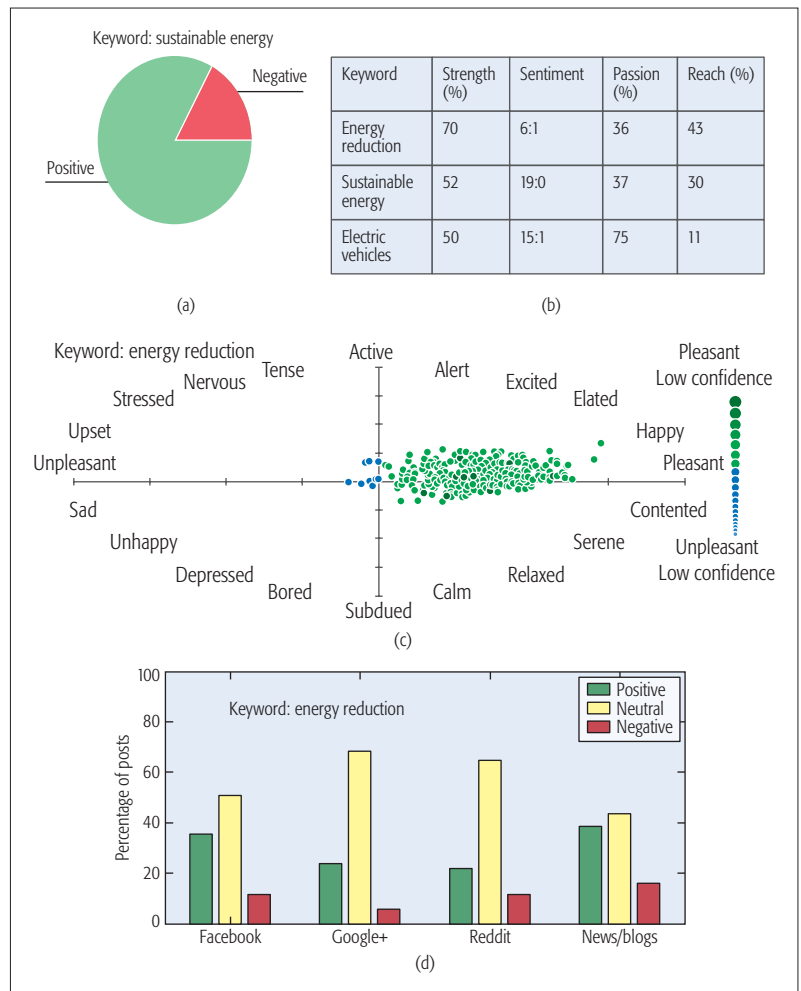


Figure 4. Sentiments of several consumers over a week towards various energy campaigns.

CONSUMER COORDINATION

The coordination of energy consumption among consumers in a neighborhood helps in balancing the temporal energy usage. Current approaches for active coordination require a central coordinator and prior agreements among consumers. TSSG framework can support completely distributed coordination with the help of SN overlay on SG. TSSG identifies and enables interactions among consumers with similar interests to balance aggregate consumption. These consumers can coordinate among themselves to address their energy demands over SN. Consequently, other consumers can schedule their demands such that total energy consumed may not exceed the generated energy. Identifying such a group of consumers is highly challenging due to the evolving nature of social context. TSSG utilizes the data from the techno-social ecosystem to derive consumer groups that can effectively coordinate and support energy management programs.

ENERGY TRADING

With the penetration of electric vehicles and renewable energy sources such as wind turbines, and solar panels in residential settings, households not only consume but also produce their own energy (“prosumers”). Concepts such as vehicle to home, vehicle to grid, and energy trad-

Currently there is no secure, privacy-aware, marketplace to enable real-time coordination for energy traders. TSSG supports energy trading by allowing consumers to coordinate on SN. Specifically, TSSG aims to enable trading of energy by mining the techno-social data where consumers with similar social contexts and beliefs are allowed to negotiate.

ing require active coordination among consumers to enable direct selling of excess energy. Currently, there is no secure, privacy-aware marketplace to enable real-time coordination for energy traders. TSSG supports energy trading by allowing consumers to coordinate on SN. Specifically, TSSG aims to enable trading of energy by mining the techno-social data where consumers with similar social contexts and beliefs are allowed to negotiate.

GOAL-ORIENTED COMMUNITIES

Heterogeneity in consumer characteristics hinders identifying target consumers for specific SG programs. Hitherto, utilities used methods based on sociology, psychology, and behavioral economics to determine target consumers for DR programs. These efforts mainly considered data reported by consumers and did not model the dynamics of energy consumption and social contexts. TSSG supports identification of consumers by deriving virtual communities — groups of consumers such that intra-community associations are denser than inter-community associations. The associations could be derived based on various features explained earlier. Based on the requirement, utilities can identify the features and consequently form communities; we call this “goal-oriented communities.” The goal could be identifying consumers who are pro-environmental, pro-behavioral change, or pro-energy reduction. Consequently, utilities can devise different incentives such as community-specific recommendations and tariffs. TSSG can support formation of these communities by applying community detection algorithms based on correlation among energy consumption, user content, social ties, and relationships.

The key requirement in determining the techno-social mining techniques that could be used in TSSG depends on the objective of the SG program. For example, to promote energy awareness and sustainability in energy usage, we need to understand consumer behavior, their preferences, and also their energy consumption pattern. Several models, such as the theory of planned behavior, health belief model, social practice theory (SPT), and diffusion of innovation theory [12], exist to understand and model behavior changes. However, SPT is increasingly being applied to analyze behavior in the context of energy management, transportation, and waste management.

The principle of SPT is that the behavior of consumers (vis-à-vis energy consumption) arises from the interactions between three components [12]:

- Norms — individual and shared expectations of comfort levels, social aspirations, and so on
- Material culture — physical aspects of a home, that is, building type, heating devices, and energy-related technologies
- Energy practices — actions of and processes used by consumers, that is, temperature settings, maintenance of technologies, and so on

SPT argues that the focus should not be on individual behavior, but rather on social practices to understand why certain practices are performed;

how and why others are prevented from carrying out some tasks; and the evolution of technology with respect to societal behavior. Furthermore, behavior change is most likely due to careful scrutiny of norms and practices, and then the promotion of the best practices.

Recently, a framework for energy culture [12] was proposed to understand the energy consumption behavior of consumers. This framework utilizes the basic components of SPT and adapts them to understand the energy consumption behavior of consumers. For example, space-heating inefficiencies might be the result of ineffective heating technologies (material culture) or inappropriate heat settings (practices) or unrealistic expectations about warmth (norms). The combination of norms, material culture, and energy practices can create self-reinforcing habitual patterns. Achieving behavioral change involves altering one or more of these components, noting that change in one will almost inevitably lead to change in the others.

TSSG extends the current social science models and can provide feedback to the consumers on the best practices to support sustainability. Specifically in TSSG:

- Norms about consumers are derived from techno-social sensing, that is, deriving consumer preferences and values from social networks such as Facebook, LinkedIn, and Twitter, as described earlier.
- Material culture and current practices can be derived from the electricity consumption patterns of the households.

Furthermore, TSSG also supports integration of other models from social sciences to analyze the behavior of consumers and promote sustainability.

ILLUSTRATION: FORMATION OF GOAL-ORIENTED COMMUNITIES

In TSSG, SNs help us to derive consumer preferences, which can be used in designing better energy management services. Some SG applications may be location-dependent, such as energy sharing between neighboring houses and energy shifting in a neighborhood. However, energy reduction and pro-environmental energy usage apply to a larger group of consumers irrespective of their locations. To show the benefits of collecting preferences of consumers, we develop new goal-oriented virtual communities to promote energy awareness, and provide tailored recommendations and community-specific tariffs. Utilities currently employ techniques that offer the same information, such as how to reduce cost and new energy policies, to its entire consumer base irrespective of their preferences and energy consumption profiles. This information may not be valid or be mostly redundant to some consumers. Forming goal-oriented communities enables tailored feedback and tariffs to consumers with similar preferences and interests. Virtual communities formed by analyzing both the social and energy contexts help bring effectiveness in the campaign.

We employ the CER dataset [13] collected during a trial in Ireland. The dataset contains energy consumption measurements from 4232

households every 30 min between July 2009 and December 2010. The objective of the trial was to investigate the effect of feedback on energy consumption in households. Each participating household filled out a questionnaire before and after the trial. The questionnaire contained questions about the socio-economic status of the consumers, appliances, properties of the dwelling, and the consumption behavior of the occupants.

Figure 5 shows a block diagram of goal-oriented community formation using TSSG core components. The techno-social sensing gathers data about average energy consumption using the smart meter (SM) and consumer preferences, beliefs, opinions, and interests using the survey data collected during the trial. The techno-social mining component tries to find structures in the sensed data to detect communities that match the goals set by the utilities. The community detection is performed by applying an *unsupervised clustering* technique called expectation-maximization (EM) clustering [14]. EM clustering applies maximum likelihood estimation to determine the optimal number of communities. The techno-social data is used by EM clustering to determine different communities based on the goals defined by the utilities. This article studies three goal-oriented community formations: pro-behavioral change, pro-environment, and pro-energy reduction. Thus, utilities can now devise tailored recommendations to the communities formed based on these goals.

Figure 6a shows the communities formed by traditional approaches, which utilize only average energy consumption of the consumers. Each community here indicates the group of consumers who have similar average energy consumption. Utilities currently use this mechanism to provide incentives and feedback to each community to reduce or change their usage patterns. However, these recommendations are not tailored based on consumer preferences, which may result in a lower adoption rate of SG programs. Hence, the TSSG framework considers data from both the energy network and social contexts of consumers to provide tailored recommendations and promote awareness. In this article, we deduce consumer preferences, opinions, and interests from the survey data. Note that the social context of consumers from other sources such as online SNS, mobile phones, and other sensors embedded in the environment can also be utilized. The questions considered from the survey are:

- Pro-behavioral change: Am/I/we interested in changing the way we use electricity if it results in reduction of the bill?
- Pro-environment: Am/I/we interested in changing the way we use electricity if it helps the environment?
- Pro-energy reduction: Is it too inconvenient to reduce our usage of electricity?³

The answers to the above questions are in the range [1, 5], where 1 means strongly agree and 5 means strongly disagree. The techno-social mining (i.e., EM clustering) uses both the average energy consumption and social context data of consumers to derive accurate goal-oriented communities. Figure 6 shows the various communities formed based on the above questions and average energy consumed.

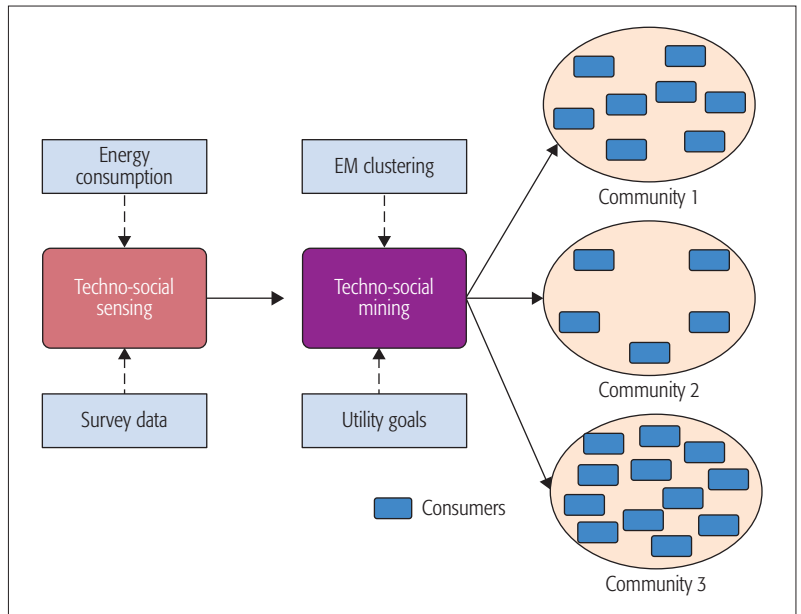


Figure 5. Block diagram of goal-oriented community formation.

Figure 6b shows six communities formed based on consumers who are *pro-behavioral change*. Mean average energy consumed in each community is shown over each bar, and the bar color shows the average response by all the consumers in that community. It can be seen that communities 1, 4, and 6 are more for behavioral change. Communities 2 and 5 are moderate toward change in their behavior, and community 3 strongly disagrees with changing the usage pattern to reduce bills. Even though consumers in communities 1, 4, and 6 have similar interests, their average energy consumption is different. Hence, utilities can now target each of these communities separately with tailored recommendations, feedback, and tariffs. For example, utilities can target community 6, with members open to behavioral change and consuming high energy, to change/modify their energy usage pattern. These recommendations and feedback have high potential to reduce consumption. Moreover, the same recommendations may not be applicable to community 1 as their energy consumption is low even though they have similar interests as community 6. Thus, TSSG can support utilities to segment consumers to promote sustainable energy usage.

Similarly, Fig. 6c shows five communities formed based on consumers who are *pro-environment*. Members of communities 2 and 5 strongly agree toward change in usage behavior to help the environment. Communities 1 and 3 are moderate toward environmental impact, and members of community 4 do not worry about the environment. It is interesting to see that even though the average energy consumed in community 2 is lower than communities 3 and 4, members of community 2 are more environment-friendly and agree to change energy usage patterns. Members of community 5 have the highest average energy consumption and are willing to change their usage patterns to become environment-friendly. This can be used by utilities to target members of community 5 by provid-

³ Questions 1, 2, and 3 correspond to questions 4332, 4333, and 4352, respectively, in the residential pre-trial survey.

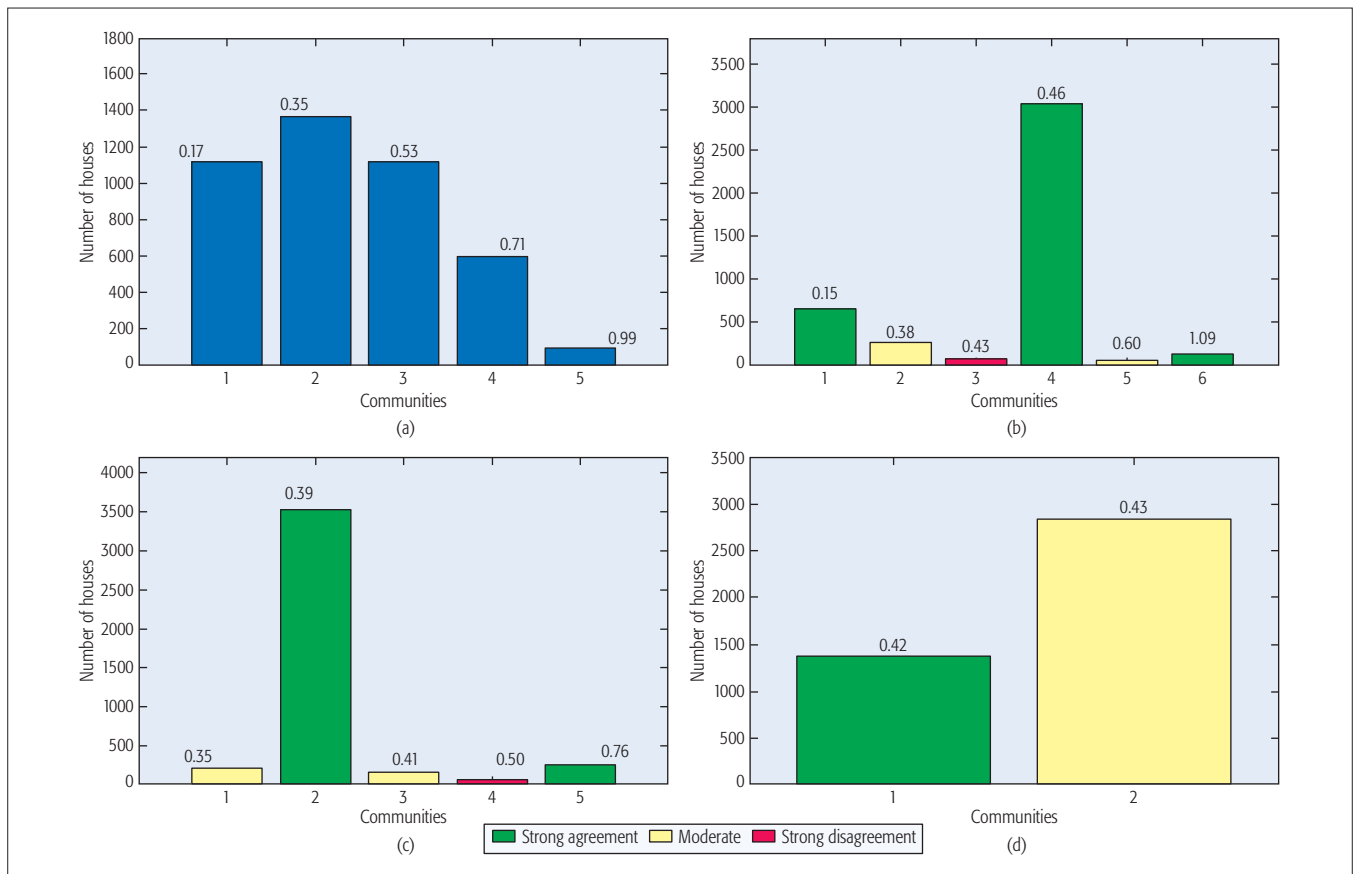


Figure 6. Different goal-oriented community formation strategies: a) communities based on average energy consumption; b) communities based on pro-behavioral change; c) communities based on pro-environment; d) communities based on pro-energy reduction. (Numbers on the bar indicate average energy consumed per household in kW).

ing recommendations on different usage patterns to reduce consumption.

Figure 6d shows the communities formed based on consumers who are *pro-energy reduction*. Only two communities are formed based on the answer and the average energy consumption. Both communities have similar average energy consumption. However, members of community 1 strongly think there is no inconvenience in reducing their energy usage, unlike members of community 2, who are moderate toward reduction in energy usage. In total, 33 percent of 4232 consumers think that there is no inconvenience in reducing energy usage, and the remaining 67 percent are moderate toward reduction in energy usage. Utilities can provide different tariffs to these communities as incentives to reduce their average energy consumption.

Deriving communities with consumers having various levels of social activities is one of the open problems that need to be addressed to guarantee fairness. Recent work by Muhammad *et al.* [15] proposes several fairness constraints to adapt the clustering and classification algorithms to consider consumers with varying levels of participation in SNs. TSSG can incorporate these constraints during community formation to guarantee fairness among consumers with varying levels of social activity. Additionally, data from surveys, face-to-face interviews, and campaigns can be used to guarantee fairness. Furthermore, our community formation strategy can be applied to DR algorithms such as load shifting, load

reduction, and energy sharing by using data from physical grid and consumers' locations.

CONCLUSIONS

Development of sustainable future-proof SG depends heavily on active participation and engagement of consumers. We proposed a novel techno-social framework for smart grids (TSSG) where infrastructure composed of various SG technologies interact with social activities of consumers. The TSSG framework uses traditional social network models to analyze consumer behavior along with energy consumption information to develop consumer-centric services. The role of TSSG toward various SG applications and its benefits are described. Furthermore, we illustrate goal-oriented community formation with data from more than 4000 households. We also showed how groups of consumers can be targeted differently by considering the heterogeneity in consumer population and their consumption.

To the best of our knowledge, TSSG is possibly the first approach that has tried to bring a holistic view, inclusive of consumers, prosumers, utilities, and SG infrastructure. We believe the integration of technological and social aspects can lead to development of sustainable and future-proof SG.

REFERENCES

- [1] X. Fang *et al.*, "Smart Grid – The New and Improved Power Grid: A Survey," *IEEE Commun. Surveys & Tutorials*, vol. 14, no. 4, 2012, pp. 944–80.

[2] Y. Kim *et al.*, "A Secure Decentralized Data-Centric Information Infrastructure for Smart Grid," *IEEE Commun. Mag.*, vol. 48, no. 11, 2010, pp. 58–65.

[3] Council of European Energy Regulators, "2020 Vision for Europe's Energy Customers, A Discussion Paper," Ref. C12-SC-02-04, 2012.

[4] Y. Strengers, "Designing Eco-Feedback Systems for Everyday Life," *Proc. ACM SIGCHI Conf. Human Factors in Computing Systems*, 2011.

[5] S. Karjalainen, "Consumer Preferences for Feedback on Household Electricity Consumption," *Energy and Buildings*, vol. 43, 2011, pp. 458–67.

[6] J. Kleinberg, "The Convergence of Social and Technological Networks," *Commun. ACM*, vol. 51, no. 11, 2008, pp. 66–72.

[7] F. Giannotti *et al.*, "A Planetary Nervous System for Social Mining and Collective Awareness," *Euro. Physical J. Special Topics*, vol. 214, no. 1, 2012, pp. 49–75.

[8] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*, Cambridge Univ. Press, 2010.

[9] S. N. A. U. Nambi *et al.*, "A Cost-Benefit Analysis of Data Processing Architectures for the Smart Grid," *Proc. MobiHoc, Wireless and Mobile Technologies for Smart Cities*, 2014.

[10] E. Bakshy *et al.*, "The Role of Social Networks in Information Diffusion," *Proc. 21st ACM Int'l. Conf. World Wide Web*, 2012.

[11] S. N. A. U. Nambi, A. Reyes Lua, and R. Prasad, "LocED: Location-Aware Energy Disaggregation Framework," *Proc. 2nd ACM BuildSys*, 2015.

[12] J. Stephenson *et al.*, "Energy Cultures: A Framework for Understanding Energy Behaviors," *Energy Policy*, vol. 38, no. 10, 2010.

[13] The Commission for Energy Regulation (CER), "Electricity Customer Behavior Trial," *Irish Social Science Data Archive*, 2012.

[14] B. D. Chuong, and B. Serafim, "What Is the Expectation Maximization Algorithm?" *Nature Biotechnology*, vol. 26, no. 8, 2008, pp. 897–99.

[15] M. Zafar *et al.*, "Fairness Constraints: A Mechanism for Fair Classification," *Proc. ICML*, 2015.

BIOGRAPHIES

AKSHAY UTTAMA NAMBI S. N. received a B.E. degree in computer science and engineering from Visvesvaraya Technological University, Belgaum, India, and is currently pursuing a Ph.D. degree at Delft University of Technology, the Netherlands. He is currently with the Embedded Software Group, Delft University of Technology. Prior to this, he was a visiting researcher with the Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland, and the Indian Institute of Science (IISc), Bangalore. His research interests are in the areas of data mining, machine learning, the Internet of Things, cyber physical systems, smart grids, and energy harvesting.

VENKATESHA PRASAD R. [SM] received a B.E. degree in electronics and communication engineering and an M.Tech. degree in industrial electronics from the University of Mysore, India, in 1991 and 1994, respectively, and a Ph.D. degree from IISc in 2003. He has been a consultant with the ERNET Laboratory, IISc. From 1999 to 2003, he was also a consultant with CEDT, IISc, involved with VoIP application developments. In 2003, he was a team leader with Esqube Communication Solutions Pvt Ltd., Bangalore, India, where he was involved in the development of various real-time networking applications. From 2005 to 2012, he was a senior researcher, and since 2012, he has been an assistant professor with the Delft University of Technology. His work at Delft University of Technology has resulted in 180+ publications. He is a Senior Member of the ACM.

To the best of our knowledge, TSSG is possibly the first approach that has tried to bring a holistic view, inclusive of consumers, prosumers, utilities, and SG infrastructure. We believe the integration of technological and social aspects can lead to development of sustainable and future-proof SG.

IEEE 5G Spectrum Sharing Challenge: A Practical Evaluation of Learning and Feedback

Sreeraj Rajendran, Bertold Van den Bergh, Tom Vermeulen, and Sofie Pollin

We present the results, experiences and takeaways from comparing a diverse set of dynamic spectrum access methods during the IEEE DySPAN 2015 spectrum challenge. Five solutions for coexistence with a given wireless link were implemented and tested in an unknown environment during the conference in Stockholm. The challenge was framed broadly, enabling participants to use their own hardware, antennas, physical layer or medium access control solutions to compete in a unified setup.

ABSTRACT

We present the results, experiences, and takeaways from comparing a diverse set of dynamic spectrum access methods during the IEEE DySPAN 2015 spectrum challenge. Five solutions for coexistence with a given wireless link were implemented and tested in an unknown environment during the conference in Stockholm. The challenge was framed broadly, enabling participants to use their own hardware, antennas, physical layer, or medium access control solutions to compete in a unified setup. Each solution was run two times and ranked using a single metric. Between the two runs the teams were allowed to improve their solution. The metric considered wanted throughput and unwanted interference. In addition to the metric, all solutions were evaluated by a jury. In this article, we give a detailed overview of the challenge, how we organized it, the participating teams, and finally the winners. We conclude with some takeaways on dynamic spectrum access.

INTRODUCTION

It has been nearly two decades since Joseph Mitola coined the term cognitive radio in 1998. A cognitive radio is an intelligent and autonomous system that can adapt its transmission and reception parameters based on the environment. Since then, wireless technology has become much more sophisticated. Capacity is boosted by adding more configurable algorithms and by utilizing various spectrum bands. Still, it remains an open question how much learning is needed and how relevant it is to tune those knobs and bands dynamically in response to the environment. Learning the environment requires feedback about the environment. This feedback can be obtained by spectrum sensing or from a database. In essence, the main questions related to the design of dynamic spectrum access radios, or cognitive radios as Mitola framed them, are:

- How much benefit can be achieved by learning and adapting, compared to other solutions at the physical or medium access layer?
- What feedback information should be available to facilitate learning and adaptation?

To start the debate toward answering these questions, IEEE DySPAN 2015 organized a spectrum challenge,¹ which was designed in such a way that any wireless research group could participate. It was up to the teams to decide how much effort they would spend on antenna

design, novel hardware, physical layer solutions, or medium access protocols. Each team could individually decide how much adaptation and learning to include. As a novel enhancement to the learning process, a database was implemented with real-time feedback about the throughput of the incumbent link. This would give each team instantaneous feedback about interference caused to the primary receiver, enabling detailed learning of the optimal transmission parameters.

While receiver feedback is currently not present in any spectrum sharing database, it is not unrealistic to explore. First, it would give an upper bound on what learning and feedback would bring. Second, we see a trend toward more and more receiver based regulation, giving the receiver a larger role in the spectrum optimization problem.² Third, many wireless systems already implement some kind of acknowledgment, which is receiver feedback that can be used to optimize transmitter parameters.

A high-level overview of the IEEE DySPAN 2015 challenge setup is given in Fig. 1. The database provides packets and performance metrics in real-time to the transmitter and collects statistics about the system performance. Both the primary user (PU) and the secondary user (SU) radio were connected to the database and could at most use two antennas. The PU system was based on IEEE 802.15.4, which was chosen because it is one of the cheapest wireless systems one can find. Therefore, even on a very limited budget, each team would be able to set up the PU system in their lab. For the challenge, the PU system was implemented using a software defined radio (SDR) IEEE 802.15.4 implementation running on two USRPs. This implementation follows the standard, but has different RF properties compared to the cheap dongle, the main idea being that while some properties are given, some remain to be learned and calibrated by using the feedback in the database or spectrum sensing.

To understand the current state of these spectrum sharing algorithms, various competitions were held in the past. The DARPA spectrum challenge,³ a prominent one among them, put state of the art (SoA) cognitive algorithms to the test by designing competitive and cooperative competitions. This challenge gave insight to the adaptability of the SoA radio algorithms in using the spectrum aggressively or sharing it among other users based on the environment. The dif-

¹ <http://dyspan2015.ieee-dyspan.org/content/5g-spectrum-sharing-challenge>

² www.whitehouse.gov/sites/default/files/microsites/ostp/pcast_spectrum_report_final_july_20_2012.pdf

³ www.darpa.mil/program/spectrum-challenge

ference with the IEEE 5G spectrum sharing challenge is that the DARPA challenge focused mainly on cooperation using a given radio and physical layer challenges, and not so much on the learning or feedback information requirements of a cognitive radio network. By comparison, the IEEE 5G spectrum sharing challenge focused on learning and feedback information requirements of a cognitive radio network using database aided spectrum sharing. The participants were given complete freedom in their radio and physical layer design, within the spectrum constraints. Another spectrum sharing challenge is the SPECTRUM-SHARC Student Cognitive Radio Contest,⁴ where students are given access to the CORNET cognitive radio testbed. Here the hardware was fixed but teams were allowed to make waveform changes. The goals of the challenge are similar to the 5G spectrum sharing challenge in that the SU needs to maximize its throughput while minimizing its interference to the PU. However, realtime feedback of the PU and SU throughput and physical design freedom made the IEEE 5G spectrum sharing challenge different from the SPECTRUM-SHARC challenge.

The rest of this article is organized as follows. We define the challenge setup and the winning parameters. A brief overview of participating teams and their radio designs are presented. We detail the actual challenge results and discuss shortcomings of the used methods. Finally, take-aways and conclusions from the challenge are presented.

CHALLENGE IMPLEMENTATION

The challenge was designed to meet a range of criteria: enable breadth in the solutions, enable teams with no hardware or expertise to participate, and enable learning with real-time feedback about the primary user throughput statistics. Given those high-level criteria, a standard IEEE 802.15.4 stack for a widely available wireless dongle was provided to ensure the participation of teams with only MAC experts who want to avoid physical layer implementation hassles. The heart of the solution was a database, that could in real-time talk to the dongle's IEEE 802.15.4 standard stack, as well as GNU Radio and LabVIEW software defined radio systems. As a result, the PU and SU could be any system, from a cheap dongle with an IEEE 802.15.4 stack to custom implementations on SDRs.

To make the competition more challenging, the PU simultaneously transmitted on four predefined frequency bands with channel spacing of 5 MHz. The secondary user had to maximize its throughput over the same 20 MHz that the PU was using. The center frequency of the band is 2.3 GHz, which was a dedicated interference-free band used for the challenge. As a result, teams had to learn only the room characteristics (fading and pathloss parameters), PU traffic patterns, and the exact PU RF properties of the USRP-X310s used.

PRIMARY USER SETUP

The PU radio had a four-channel GNU Radio based IEEE 802.15.4 stack [1] with an O-QPSK physical layer connected to a USRP front-end via an Ethernet interface, as shown in Fig. 2. The PU used four independent streams that were configured to transmit packets independently on

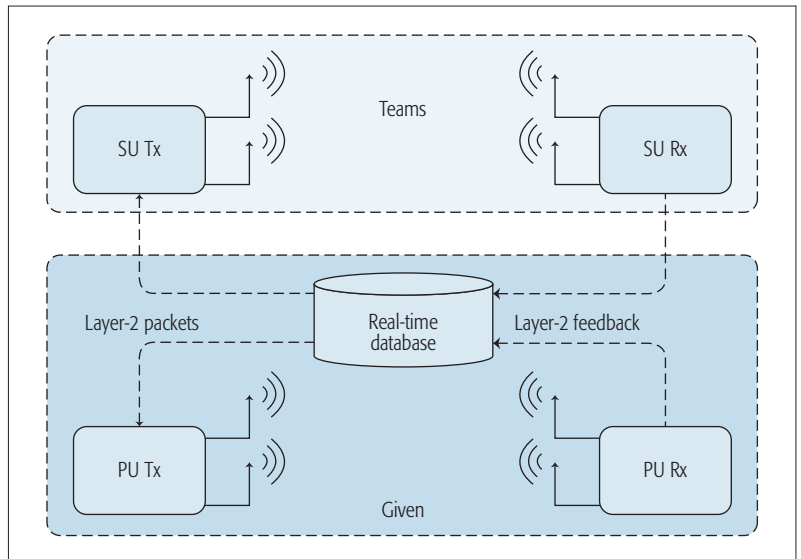


Figure 1. Spectrum challenge setup showing the real-time feedback from the database.

the four channels using a fixed packet length of 127 bytes. The packet generator block controlled these four streams by requesting data from the database, re-sizing it, and pushing the packets to the independent 802.15.4 MAC modules of the streams. The timing of these packets was controlled by the packet generator and was based on four instantiations of the same random distribution. The packet generator switched to a new distribution every three minutes to mimic the non-stationary nature of channel occupancy.

The random distributions used in the challenge consisted of:

- A uniform distribution with minimum and maximum inter-packet duration of 8ms and 150ms, respectively.
- Two Poisson distributions with means 20ms and 150ms.
- A back-to-back transmission scheme with minimal inter-packet period of 5ms.

All the parameters of these distributions were configured before the start of the challenge. A common starting seed was used for all distributions, which helped keep the randomness fixed across the teams during both phases of the challenge. The implemented code can be easily adapted to follow different distributions to test more complex channel occupancy scenarios in the future. In reality, there can be many different spectrum occupancy scenarios, so a fairly large set of possible random distributions were chosen for the challenge.

DATABASE AND FEEDBACK

Giving feedback about the PU performance to the SU radio should allow for much faster and more fine-grained adaptation of the SU settings. This feedback mechanism was implemented using a central database server. The server delivers random packets to both the PU and SU. Received packets are delivered back to the database server, where they were verified and statistics were updated. These statistics were accessible from both the PU and the SU.

To generate packets with strong randomness

⁴ <http://radiocontest.wireless.vt.edu/index.html>

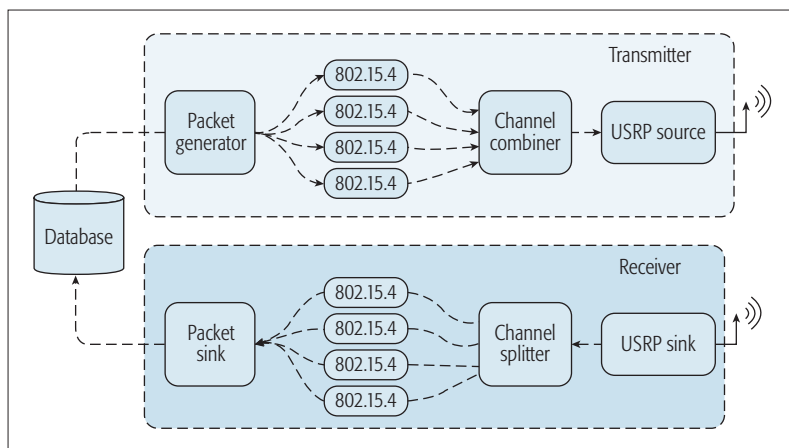


Figure 2. Primary user setup.

at high throughput, AES-128 was used in counter mode with a random key. It is important that frames are sufficiently random, since otherwise they could be compressed before transmission. A secure message authentication code (MAC) is calculated over the frame to detect if it has been corrupted. A sequence number is added to prevent packets from being counted multiple times.

Since we feel that this database platform coupled with client libraries for both GNU Radio and LabVIEW may be useful to others, it is made publicly available.^{5,6} The PU implementation used in the challenge is also made available in this package.

CHALLENGE VISUALIZATION

A challenge visualization was set up for live monitoring and analysis of the status of the challenge. A screen capture of the used visualization is shown in Fig. 3. In the left column, graphs showing instantaneous PU and SU throughput are plotted along with the elapsed time. A live FFT and waterfall

⁵ <http://claws.be/spectrum-challenge/>

⁶ https://github.com/networked-systems/dyspanchallenge_2015

⁷ <http://gqrx.dk/>

plot, as shown in the right column of the figure, was used for monitoring the channel occupancy patterns of both pairs of radios. A custom visualization module connected to the real-time database was used for displaying the PU and SU performance. A software defined radio receiver powered by the GNU Radio and the Qt graphical toolkit Gqrx⁷ was used to display spectrum details.

CHALLENGE PHASES

The challenge consisted of two phases, a learning phase and a test phase, as given below.

- Learning phase (10 min):
 - SU can learn PU statistics.
 - PU feedback is provided to optimize the SU parameters.
 - No scoring in this phase.
- Test Phase (10 min):
 - Actual scoring phase.
 - PU transmission statistics same as during the learning phase.
 - SU penalized for interference.

Each phase had a duration of 10 minutes. The SU radio could learn about the environment, the PU transmission statistics, and the exact PU transmitter and receiver RF properties impacting the interference sensitivity. This acquired knowledge could then be used to calibrate the SU parameters and algorithms to improve its performance. During the learning phase, the SU could also test its algorithms and validate them efficiently by making use of the real-time feedback from the database. The final scores were only calculated during the test phase of the challenge.

CHALLENGE METRIC

Two winners were selected. One was selected based on a single metric combining both SU and PU throughput, which could be measured

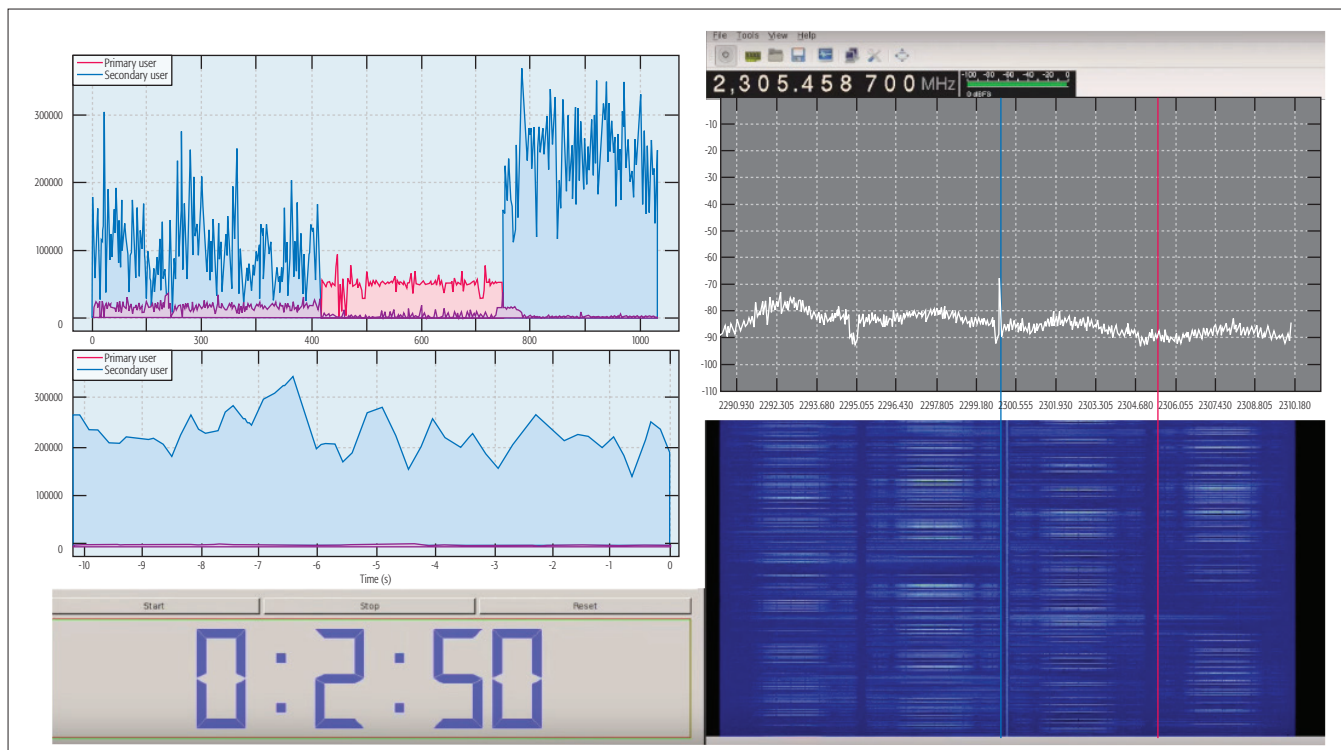


Figure 3. Challenge visualization: Instantaneous PU and SU throughput (left) with live FFT (top-right) and waterfall plots (bottom-right).

Team	Physical layer	Coexistence techniques
KIT	Fully parameterizable filter bank multicarrier (FBMC) PHY. 256 subcarriers over four channels with 5 MHz each, 2 b/s/Hz. Low adjacent channel interference. High spectral efficiency. Short packets for low probability of interference.	Adaptive noise floor estimation, no calibration needed. Energy detection gives reliable information about PU beyond database. Past knowledge from SU transmitter sensing used to improve detection probability. Locally sensed busy channels are temporarily blacklisted. Omnidirectional antennas ensure consistent performance in dynamic scenarios.
CNCT/TUI	OFDM PHY Receive on 4x5MHz channels at Rx Directional antennas	Efficient multi-threaded solution for simultaneous multi-channel transmission. Noise floor estimation on each channel, which improves PU detection performance. Best modulation and coding scheme selected during the learning phase. PU channel occupancy modeled as a Markov chain. SU's transmission channel selected based on this Markov model.
AIT	Tx-Rx directional Yagi-Uda planar parasitic antennas OFDM modulation with QPSK or 16QAM	Modulation selected based on PU and the SU throughput feedback. SU transmits between two adjacent PU channels to avoid interference (e.g., 2.295–2.3 GHz). Dynamic gain control to minimize PU interference and boost SU's throughput.
FORTH-ICS	4x IEEE 802.11a/g @ 5 MHz (SU Tx/Rx) 4x IEEE 802.15.4 (PU detector) Abstraction in OS – everything appears as typical interface	Virtualization in time and frequency domain. PU frame, inter-frame and channel transition durations are decoded. Decoded parameters used to select packet length and modulation parameters.
FR	Custom physical layer waveform FSK modulated tones used as the base waveforms	SU waveform exhibits negligible projection on the PU waveforms. No mutual interaction between PU and SU waveforms. Transmit gain control based on the feedback.

Table 1. Secondary user features.

unambiguously from the database. The final challenge score was represented by the product of SU throughput (T_U) and PU (S_{PU}) satisfaction. The PU satisfaction was calculated from the offered PU throughput (\hat{T}_{PU}) and the delivered PU throughput (T_{PU}) as given in Eq. 1. A maximum throughput loss tolerance of 10 percent was allowed. More than 10 percent PU throughput loss would be counted as zero PU satisfaction.

$$Score = T_{SU} \times S_{PU}$$

$$S_{PU} = \max\left(0, \frac{10}{9} T_{PU} - \hat{T}_{PU}\right). \quad (1)$$

In addition to this single performance metric, which can be measured but does not capture the novelty of the solution, a second winner selected by a jury was announced. The jury evaluated mainly the maturity of the solution, and most importantly the breadth, i.e., how many cognitive aspects were considered in the final solution. This would exclude non-adaptive teams, or teams failing to take advantage one way or another of the PU or SU feedback enabled in the system.

SECONDARY USER ALGORITHMS

The challenge was designed to enable a large variety of teams to participate in the event. Eventually, five teams were selected to compete, using techniques ranging from special waveforms to advanced statistical PU profiling techniques. A brief overview of each team's implementation is given below. The summary of the used physical layer parameters and co-existence techniques can be found in Table 1. Each team published the details of their solution as an IEEE DySPAN 2015 challenge paper [2–6].

Team 1–KIT: The team from Karlsruhe Institute of Technology (KIT) used a cross layer optimized secondary user system [2] for the challenge. They employed a secondary user waveform with high spectral efficiency, sensing

based on energy detection with thresholds that are learned during the learning phase of the challenge, and diversity gain that is achieved using multiple antennas. The channel occupancy knowledge is accommodated in the MAC layer of the SU system. The SU transmitter synchronously sensed all four channels. The measured power levels are then used to adapt the energy detection threshold, and one packet is transmitted on every detected free channel. Static learning is employed in the learning phase of the challenge to learn about the channel utilization probabilities, and reinforcement learning is used in the test phase of the challenge.

Team 2–CNCT/TUI: A joint team from CONNECT/CTVR, Trinity College Dublin, and Technische Universität Ilmenau (TUI) devised a system that makes use of a state machine consisting of sensing, learning, decision making, and transmit/receive [3]. Four-channel frequency domain energy detection is used during the sensing state. The best modulation and coding schemes are learned during the learning phase of the state machine. The PU channel occupancy parameters are also sensed, and a channel distribution is built with the sensed information. This is used to model the PU pattern as a Markov chain, which is used for SU transmission channel selection.

Team 3–AIT: The team from Athens Information Technology (AIT) used parasitic directional antennas and tried to transmit in the blind spots of the primary user receiver [4]. The system selects the best beam pattern from a set of patterns that is learned during the learning phase of the challenge. The system employs adaptive power control based on PU throughput feedback from the database.

Team 4–FORTH-ICS: The team from the Foundation for Research and Technology–Hellas (FORTH) used a custom developed PCIe SDR device with a standard 802.11g stack [5]. The

A group from San Diego State University and University of California San Diego designed a SU radio waveform to overlay the PU channels without employing any monitoring or learning [6]. The designed waveform creates negligible interference to the PU waveform avoiding the need for any learning about the primary user statistics.

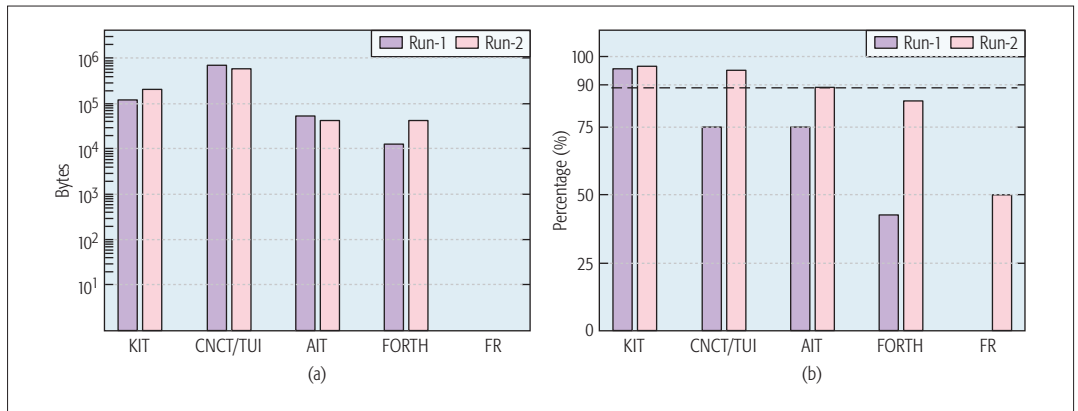


Figure 4. Primary and Secondary user performance: a) SU Bytes transferred and b) PU packet success percentage.

available 20MHz spectrum is virtualized, and the system can transmit in four adjacent 5MHz channels in parallel. During the learning phase the SU discovers available SNR on each channel. Furthermore, the maximum transmission power that will not trigger the PU CCA mechanism is measured. The PU inter-frame time period, frame size distributions, and PU channel switching time are learned during the learning phase of the challenge by monitoring and decoding PU transmissions. The transmission power, minimum packet size, and frame distributions thus learned are used in the test phase of the challenge.

Team 5–FR: A group from San Diego State University and the University of California San Diego designed a SU radio waveform to overlay the PU channels without employing any monitoring or learning [6]. The designed waveform creates negligible interference to the PU waveform, avoiding the need for any learning about the primary user statistics.

CHALLENGE RESULTS AND DISCUSSION

Two runs (20 minutes for each team per run) of the challenge were conducted, and the winner was selected based on the best scores out of two runs. The number of runs were fixed to two in order to give the participants a second opportunity to fine tune their algorithms. The availability of dedicated spectrum and the reproducible PU spectral occupancy pattern made the performance results repeatable. Figures 4a and 4b detail the SU and PU performances, respectively. Figure 4a summarizes the bytes transferred by each team during both runs of the challenge. Even though most of the teams were able to achieve high SU throughput, only two teams managed to keep the PU throughput loss below 10 percent (Fig. 4b). Some teams tried to overcome this using short packet lengths and fine transmit power control, while others used directional systems to keep the interference level low.

WINNERS

From the final scores computed from both runs, team CNCT/TUI was selected as the winner based on the metric. CNCT/TUI's high throughput PHY-MAC combination along with beamforming techniques helped in achieving the winning score. Team KIT was selected by the jury as the winner based on the breadth and con-

sistency of the solution. Team KIT performed consistently well with their spectrally efficient PHY and short length packets. Even though team AIT employed many features, including adaptive power control based on the PU feedback, they failed in suppressing the carrier leakage from their radio front-end, which interfered continuously with the PU. Team FORTH also had difficulties in controlling the interference with the PU, as their algorithm assumed only a single channel usage by PU at a time. The non-interfering signal design from team FR also performed badly, as the implemented adaptive power control mechanism failed in reducing the interference.

Referring to the feature list, we can conclude that a radio with local sensing-based learning, a spectrally efficient physical layer, and directional antennas or equivalent adaptive power control for reducing interference, are promising building blocks for dynamic and cognitive spectrum access.

The implementation challenges faced by the teams and feedback from each team after the challenge is summarized in Table 2. Even though the participating radios were not fully context aware, the challenge showed that promising technological building blocks exist that can design dynamic spectrum access (DSA) solutions for challenging PU systems.

TAKEAWAYS

GENERAL DISCUSSION

Given the breadth of all possible solutions for realizing cognitive radio, the challenge entries explored several technologies for DSA. Teams designed their solution, sometimes focusing on a single technique (e.g., FR) or by combining many known techniques (e.g., CNCT/TUI). Although it is difficult to draw general conclusions from the results of the challenge, the following observations are worth discussing.

- Given the metric, it made sense to go for high throughput designs to have a chance to win based on the metric. Yet, teams failed to also sufficiently focus on interference avoidance to improve the PU satisfaction measure.

- Directional antennas were advantageous during the challenge, as the setup was static and hence it was very easy to point to the SU receiver, and away from the PU receiver. Some teams

achieved this by manual configuration, others learned it during the learning phase.

- Due to the dynamic nature of cognitive radio for spectrum sharing, hard-coded assumptions about the PU can be dangerous. For example, FORTH-ICS had a brilliant solution that, however, assumed that the PU only used a single channel at a time. It was difficult to adapt the solution on the spot.

- While all solutions had some form of intelligence and adaptation (minimally power control), most teams failed to take advantage of the learning phase to make significant improvements in the system. A main reason for this is that the PU was very dynamic, and learning this was of course very challenging. While gains by automated learning were not logged, we did log gains by human intervention in between the two runs. Indeed, all the teams did better the second time, which proves that by some human or manual calibration, all designs could be improved. However, the ideal cognitive radio should be able to learn its configuration autonomously, or ideally even outperform human calibration. We are still far from that reality, but some promising building blocks to enable this were tested during the challenge.

- While there was real-time feedback of PU interference, one would expect that it should be easy to design a scheme that meets the interference constraints, by simply disabling the transmitter when the PU interference ratio became too high. Most teams did not strongly rely on this feedback and hence did not manage to keep interference below the target levels.

SUGGESTIONS FOR FUTURE WORK

The challenge results revealed that solutions utilizing time domain spectral vacancies yielded good results. The SoA research also suggests devising policies that make use of spectral vacancies [7, 8]. The main challenges in realizing such a system include limitations in observing the channel in a half duplex radio receiver, modeling limitations of the non-stationary PU transmissions [7], and the inherent non-determinism of channel occupancy. The solutions presented during the challenge tried to tackle some of these challenges, for example predicting the PU channel occupancy pattern using a hidden Markov model (HMM) [3]. However, models based on such dynamic Bayesian networks (DBNs) like Kalman filters and HMMs may be sub-optimal, mainly due to relatively simple state transition structures or internal state space structures. Models employing deep learning techniques that have rich internal state representations and flexible non-linear functions are shown to outperform these conventional techniques. DSA systems employing such models are yet to be investigated. An in-band full duplex system can also benefit DSA as it improves the quality of observation of the environment, which aids instantaneous response. For example, a SU can stop an ongoing transmission if it detects a PU transmission in the same channel.

The database feedback was used only by a few teams during the learning phase to test the robustness of their sensing schemes or for transmitter power control. While most of the

Team	Challenges and takeaways
KIT	(-) Non-deterministic latency in PC host based SDR is a serious issue, but can be mitigated by moving control algorithms to FPGA. (+) Adaptive modulation and code rate along power control are key required features. (+) Short packets can easily reduce interference.
CNCT/TUI	(+) Exploiting directional antennas really improves DSA performance. (+) Moving the host code close to the radio helps to solve SDR latency issues. (+) Noise floor calibration with simple energy detection works in practice.
AIT	(+) Beamforming techniques using parasitic antennas greatly improve PU-SU isolation. (+) Transmitting in the PU spectrum nulls is effective with proper power control. (-) RF frontend calibration is a must; carrier leakages can result in high interference.
FORTH-ICS	(-) Non-practical assumptions about PU statistics can harm the entire system. (-) Forced to use small constellations (BPSK) as the interference with PU was high as a result of wrong assumptions. (+) SDR virtualization over existing standards can be used with proper adaptations [5].
FR	(+) Knowledge about the PU physical layer can be used to design interference free waveform. (+) Automatic transmit power control is essential even with interference free waveform design. (+) Complex PU statistics learning can be avoided with these designs. (-) SU transmit power should be tightly controlled using PU feedback, which can otherwise result in high interference.

Table 2. Team feedback.

challenge entries relied on local sensing and directional antennas for improving the metric, the use and value of feedback from the PU is not thoroughly investigated. Such an investigation makes sense, as the feedback helps in controlling SU interference in mission critical PU scenarios. It has been shown that reinforcement learning techniques like Q-learning, which makes use of PU feedback, can improve the system performance [9]. On the other hand, the practicality of enabling feedback and deciding the type of feedback should be debated more before reaching a proper consensus. For example, if the SU can decode the PU transmissions, it may be possible to infer the extent of interference caused. Alternatively, the PU could provide this information over a side-channel, e.g., the Internet. In the future, it would also be interesting to continue building on the initial setup by adding more challenges that focus more on the metric, such as the PU latency, the scenario, the PU traffic patterns, the PU RF properties, the learning phase, and availability of detailed feedback.

CONCLUSION

Spectrum is scarce, and to make optimal use of it we will have to share it. Sharing needs adaptive techniques that enable radios to adapt their configurations to the exact properties of the interfered system, packet transmission parameters as well as detailed RF properties and sensitivities. In the IEEE DySPAN 2015 challenge, a system was implemented giving the competing users optimal feedback about their impact on the legacy, primary system. Surprisingly, it was learned that keeping interference below well communicated limits was still a major challenge. The reasons for this were very broad, from hardware non-idealities that were underestimated, to failure to take advantage of the exact feed-

Empirical results show that effective time domain utilization of spectral vacancies yielded good results, which is also backed by SoA research. Policies making use of time domain sharing along with viable feedback schemes for PU interference reduction should be investigated further to enable DSA systems in practice.

back. Nevertheless, two teams achieved a very high throughput, showing that with existing technology it is possible to design a DSA system in practice, even for a very challenging and dynamic primary user scenario. However, this was enabled by some human interventions and algorithm changes in between the two runs. Even though we understand the key features required, the main challenge remains in adding full context awareness in radios that can adapt without much human intervention.

As a main conclusion for the IEEE DySPAN 2015 challenge, we can say that participating teams learned a lot, from the challenge and from each other, and went home with an improved design and solution. The design and organization of such small, targeted challenges can serve the community tremendously, as it makes it possible to benchmark solutions, compare them, learn from each other, and give teams prime opportunities to encounter challenges that they had not yet considered when testing in the lab only. While it is hard to generalize the winning solutions as the best or must-have DSA technology, it is possible to generalize some lessons learned about learning, adaptability, human intervention, and how challenging wireless communication really is. Empirical results show that effective time domain utilization of spectral vacancies yielded good results, which is also backed by SoA research. Policies making use of time domain sharing along with viable feedback schemes for PU interference reduction should be investigated further to enable DSA systems in practice.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the article. They are also grateful to National Instruments for the hardware support during the challenge and for the final prizes. Finally, the authors would like to thank all challenge participants for their enthusiasm and co-operation.

REFERENCES

- [1] B. Bloessl *et al.*, "A GNU Radio-based IEEE 802.15.4 Testbed," *Proc. 12. GI/ITG KuVS Fachgespräch Drahtlose Sensornetze (FGSN 2013)*, Cottbus, Germany, Sept. 2013, pp. 37–40.
- [2] A. Kaushik *et al.*, "Spectrum Sharing for 5G Wireless Systems (Spectrum Sharing Challenge)," *2015 IEEE Int'l. Symp. Dynamic Spectrum Access*

Networks (DySPAN), Stockholm, 2015, pp. 1–2.

- [3] J. Tallon *et al.*, "Coexistence Through Adaptive Sensing and Markov Chains," *2015 IEEE Int'l. Symp. Dynamic Spectrum Access Networks (DySPAN)*, Stockholm, 2015, pp. 7–8.
- [4] D. Ntaikos *et al.*, "Low-Complexity Air-Interface-Agnostic Cooperative Parasitic Multi-Antenna Spectrum Sharing System," *2015 IEEE Int'l. Symp. Dynamic Spectrum Access Networks (DySPAN)*, Stockholm, 2015, pp. 5–6.
- [5] S. Papadakis *et al.*, "Robust spectrum sharing through virtualization," *2015 IEEE Int'l. Symp. Dynamic Spectrum Access Networks (DySPAN)*, Stockholm, 2015, pp. 9–10.
- [6] F. Harris, R. Bell, and V. K. Adsumilli, "Spectrum Sharing Between a ZigBee Frequency Hopper and an FSK Modem," *2015 IEEE Int'l. Symp. Dynamic Spectrum Access Networks (DySPAN)*, Stockholm, 2015, pp. 3–4.
- [7] Y. Liu and A. Tewfik, "Primary Traffic Characterization and Secondary Transmissions," *IEEE Trans. Wireless Commun.*, vol. 13, no. 6, June 2014, pp. 3003–16.
- [8] S. Huang, X. Liu, and Z. Ding, "Opportunistic Spectrum Access in Cognitive Radio Networks," *IEEE 27th Conf. Computer Commun. INFOCOM 2008*, Phoenix, AZ, 2008, pp. 2101–09.
- [9] K. A. Yau, P. Komisarczuk, and P. D. Teal, "Reinforcement Learning for Context Awareness and Intelligence in Wireless Networks: Review, New Features and Open Issues," *J. Network and Computer Applications*, vol. 35, no. 1, Jan. 2012, pp. 253–67.

BIOGRAPHIES

SREERAJ RAJENDRAN (sreeraj.rajendran@esat.kuleuven.be) received his masters degree in communication and signal processing from the Indian Institute of Technology, Bombay, in 2013. He is currently pursuing the Ph.D. degree in the Department of Electrical Engineering, KU Leuven, Belgium. Before joining KU Leuven, he worked as a senior design engineer on the baseband team of Cadence, and as an ASIC verification engineer at Wipro Technologies. His main research interests include machine learning algorithms for wireless and low power wireless sensor networks.

BERTOLD VAN DEN BERGH (bertold.vandenbergh@esat.kuleuven.be) is a Ph.D. student in the Department of Electrical Engineering, KU Leuven, Belgium. His research is focused on reliable communication systems for UAVs and other applications. Before starting his Ph.D. he obtained his masters degree in microelectronics at KU Leuven in 2013, and did an internship at Flanders Make on wireless localization and tracking of mechatronic systems. He also worked at ETH Zurich on the Swarmix project.

TOM VERMEULEN (tom.vermeulen@esat.kuleuven.be) obtained his B.Sc. and M.Sc. degrees in electrical engineering from KU Leuven, Belgium in 2011 and 2013, respectively. In between he did an internship at National Instruments developing courseware on software defined radios. Currently he is a Ph.D. candidate at KU Leuven, focusing on in-band full duplex communication. In 2016 he was a visiting scholar at UCLA, where he worked on simultaneous transmissions and collision detection. His main research interests are software defined radio, low power wireless sensor networks, and in-band full duplex.

SOFIE POLLIN (sofie.pollin@esat.kuleuven.be) obtained her Ph.D. degree at KU Leuven with honors in 2006. From 2006 to 2008 she continued her research on wireless communication, energy-efficient networks, cross-layer design, coexistence, and cognitive radio at UC Berkeley. In November 2008 she returned to imec to become a principal scientist on the green radio team. Since 2012 she has been a tenured track assistant professor in the Department of Electrical Engineering, KU Leuven, Belgium. Her research centers around networked systems that require networks that are ever more dense, heterogeneous, battery powered, and spectrum constrained. She is a BAEF and Marie Curie fellow, and an IEEE senior member.

New Technologies and Trends for Next Generation Mobile Broadcasting Services

Alejandro de la Fuente, Raquel Pérez Leal, and Ana García Armada

ABSTRACT

It is expected that by the year 2020, video services will account for more than 70 percent of mobile traffic. It is worth noting that broadcasting is a mechanism that efficiently delivers the same content to many users, not only focusing on venue casting, but also distributing many other media such as software updates and breaking news. Although broadcasting services are available in LTE and LTE-A networks, new improvements are needed in some areas to handle the demands expected in the near future. In this article we review the actual situation and some of the techniques that will make the broadcast service more dynamic and scalable, meeting the demands of its evolution toward the next generation. Resource allocation techniques for broadcast/multicast services, integration with new waveforms in 5th generation mobile communications (5G), initiatives for spectrum sharing and aggregation, or the deployment of small cells placed together with the existing macro cells, are some enhancements that are examined in detail, providing directions for further development. With this evolution, 5G broadcasting will be a driver to achieve the spectral efficiency needed for the 1000 times traffic growth that is expected in upcoming years, leading to new applications in 5G networks that are specifically focused on mobile video services.

INTRODUCTION

Mobile data traffic has been growing rapidly in recent years, and this growth is expected to accelerate in the near future. For that reason, the mobile industry is preparing for 1000 times data traffic growth, where richer content will be delivered, including more video traffic used in multiple emerging applications. Furthermore, 28 billion interconnected devices are expected in 2021 [1]. It is widely accepted that this 1000 times capacity increase will be achieved by the combination of three approaches:

- Spectral efficiency improvements achieved through the introduction of new signal processing and coordination techniques.
 - The optimized use of the spectrum, adequately combining licensed and unlicensed frequency bands.
 - An increase in the density of base stations.
- Mobile broadcasting is one of the major driv-

ers to achieve these goals, since it provides an efficient use of the spectrum. Indeed, current trends in broadcasting will make it more dynamic, making possible the efficient use of on-demand spectrum. In addition, the deployment of small cells, thus increasing the number of base stations, enhances venue casting and allows the use of unlicensed spectrum in some areas.

Mobile broadcasting is an important ingredient in the evolution of digital television (TV) broadcasting. This market has traditionally been fragmented, with five different standards in the market: advanced television systems committee (ATSC) in North America; digital video broadcasting (DVB) in Europe; digital terrestrial multimedia broadcast (DTMB) in China; integrated services digital broadcasting (ISDB) in Japan; and digital multimedia broadcasting (DMB) in Korea. From this starting point, Future of Broadcast Television (FOBTv), an initiative to create a common ecosystem for the future generation of broadcast systems, was launched in 2012. The development of this common working scenario will allow the different actors in the broadcasting services sector to take advantage of global economies of scale.

As an alternative, the mobile industry has adopted the 3rd Generation Partnership Project (3GPP) technologies, Long Term Evolution (LTE) and LTE-Advanced (LTE-A), to cope with the growing demand of data traffic in mobile networks, combining unicast and multicast transmissions based on evolved multimedia broadcast and multicast service (eMBMS).

A hybrid network approach, combining the existing infrastructure of both broadcasters and mobile operators, has also been proposed as a solution. This synergy may benefit from the mobile infrastructure deployed on dense low power low tower (LPLT) networks complemented with existing terrestrial broadcasting deployed on high power high tower (HPHT) networks. The concept of future extension frames (FEF) in DVB-second generation terrestrial (DVB-T2) systems is defined to achieve the convergence of eMBMS and DVB-next generation handheld (DVB-NGH). Several projects are working on the hybrid approach based on this concept, for example with the definition of a common physical layer (CPHY), or using the FEF concept together with the carrier aggregation proposal for LTE-A.

Although broadcasting services are available in LTE and LTE-A networks, new improvements are needed in some areas to handle the demands expected in the near future. The authors review the actual situation and some of the techniques that will make the broadcast service more dynamic and scalable, meeting the demands of its evolution toward the next generation.

This work was supported in part by the Spanish Ministry of Economy and Competitiveness, National Plan for Scientific Research, Development and Technological Innovation (INNPACTO subprogram), LTEXtreme project (IPT-2012-0525-430000), and the project TEC2014-59255-C3-3-R (ELISA).

The authors are with the Universidad Carlos III de Madrid.

Digital Object Identifier:
10.1109/MCOM.2016.1600216RP

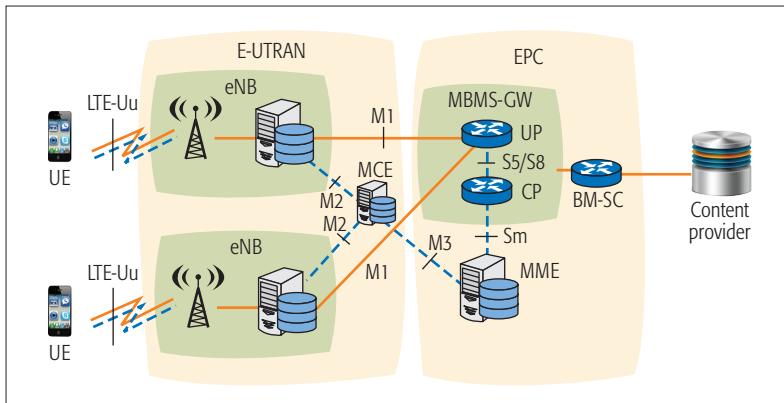


Figure 1. Architecture required for the introduction of eMBMS in LTE showing the different nodes that constitute the radio access and core networks.

Today the different actors have not reached a consensus, and it is not clear which one of the three options would be the main component of future broadcasting networks. This article is focused on the 3GPP eMBMS proposal, which can be also part of a hybrid approach. This technology presents several limitations today that require an evolution. Unlike unicast, broadcast/multicast transmissions by their nature guarantee fairness among all users at the cost of a decrease in throughput. Consequently, new radio resource management strategies are required to achieve a good trade-off between fairness and throughput, enabling higher throughput for users with good channel conditions. These new strategies should take into account both the traditional broadcast/multicast scenarios and new trends such as ultradense networks (UDN). New waveforms are being proposed for the evolution of mobile radio access with several new requirements in mind, such as the emergence of machine-type communication (MTC) in the context of the Internet of Things (IoT). Also, the massive deployment of small cells is being proposed. Even though small cells will play an important role in broadcasting, the ability to cover wide areas should be preserved in some scenarios, such as TV broadcasting. Therefore, the ability to cope with broadcast requirements should be considered in this new waveform design. New spectrum usage schemes are emerging based on spectrum sharing [2]. Novel techniques for spectrum sharing and aggregation must be designed for broadcasting to take advantage of this new paradigm. In this article we describe the technologies that are being envisioned to improve broadcasting, providing an overview of the actual challenges and potential solutions.

The rest of the article is structured as follows. The next section describes broadcasting in LTE and LTE-A standards, followed by a discussion of new technologies and solutions for the evolution of broadcasting services. Then we discuss the new applications that will make use of these technologies. The article finishes with some concluding remarks.

BROADCASTING IN LTE AND LTE-A

Traditionally, mobile systems have been using unicast transmissions to every user, even to deliver certain services such as radio or TV con-

tent in wide areas. However, the utilization of unicast transmissions presents clear limitations regarding radio resource allocation in this context. Broadcast/multicast transmissions where the same content is delivered simultaneously to a certain amount of users in a determined area have inherent advantages because of the use of common resources.

3GPP LTE defined eMBMS in [3] to deliver broadcast/multicast services in mobile networks, therefore combining unicast and broadcast services in the same network. The architecture required for the introduction of eMBMS in LTE is shown in Fig. 1, where new entities are included. The broadcast multicast service center (BM-SC) is responsible for authorization, authentication, billing, and global configuration. The MBMS gateway (MBMS-GW) manages the sending of multicast IP packets from the BM-SC to the evolved Node B (eNodeB). The control signaling of the session is managed by the mobility management entity (MME). Finally, the synchronization among the cells in the multicast/broadcast over single frequency network (MBSFN) area is carried out by the multi-cell/multicast coordination entity (MCE).

The solution based on MBSFN has been adopted by 3GPP LTE to improve the performance of eMBMS. MBSFN improves the signal to interference plus noise ratio (SINR) of the users located in the areas where the coverage overlaps from different base stations by using common radio resources in all the cells belonging to the same single frequency network (SFN) area. The whole SFN behaves as a single cell with the only interference caused by the signals arriving outside the duration of the cyclic prefix.

Broadcasting in LTE was commercially launched in South Korea in January 2014. After that, the BBC deployed an experimental system for the 2014 Commonwealth Games. Another representative example is the 2014 Superbowl, where the mobile operator Verizon successfully trialed live eMBMS technology to a selected group of end users in New York, with video streaming of 1.8 Mb/s that could be received directly in full-screen format on mobile devices and also in a four-panel mosaic.

NEW TECHNOLOGIES FOR NEXT GENERATION BROADCASTING SERVICES

Nowadays, many new technologies are being considered to support next generation broadcasting services in LTE-A mobile networks and beyond. In this section we detail some of them, summarizing their current status and ongoing research.

ADAPTIVE RESOURCE ALLOCATION IN BROADCAST TRANSMISSION

Resource allocation strategies are becoming an important issue for broadcast and multicast services. New techniques are required to achieve high performance, both in terms of total service throughput and fairness among all the users.

The different channel conditions between the users have traditionally forced system designers to adopt a conservative approach, which maximizes the fairness among the users. However, this is achieved at the cost of limiting the ser-

vice throughput by the user that suffers the worst channel conditions. The conventional multicast scheme (CMS) introduces inefficiencies when some users experience poor channel conditions, and as a result the available spectrum is not fully exploited.

Recent research on broadcast/multicast resource allocation in orthogonal frequency division multiplexing (OFDM)-based systems proposes solutions based on an efficient distribution of physical resource blocks (PRB) among different broadcast/multicast groups to improve the trade-off between service throughput and fairness. The block diagram of a packet scheduler that performs the resource allocation of an OFDM system is depicted in Fig. 2. The frequency domain scheduler is used to split all the broadcast or multicast users into several subgroups according to the channel quality indicator (CQI) reported by the terminals, and after that to design a resource allocation strategy based on multicast subgroups using different MCS. These strategies allow the system to improve the fairness among users with different channel conditions, delivering the service at different rates to the users that belong to each subgroup, at the time the total throughput is maximized.

Figure 3 shows a performance evaluation of different multicast resource allocation strategies studied in [4]. The use of CMS results in maximizing fairness among the users; on the other hand, by using opportunistic multicast scheduling (OMS), spectral efficiency is maximized. Furthermore, the subgroup-based strategy has been analyzed using the optimization of three different utility functions, i.e. maximum throughput (MT), proportional fairness (PF), and minimum dissatisfaction index (MDI). Creating multicast subgroups with MDI utility function is shown to provide high fairness with a slight decrease in the throughput performance of the users belonging to the same multicast group.

Practical systems demand low-complexity schemes, where the time required to adapt the resource allocation to channel variations is minimized. Solutions to find optimal resource allocation in very few steps have been studied for a single cell with one multicast group in [4].

Other research is focused on resource allocation strategies for multi-flow delivery among the users in multicast environments [5]. Additionally, the eNodeB requires the CQI feedback from the users, but CQI is a private information generated by the users with their own measurements and can be untruthful. Based on these statements, a mechanism to elicit the true CQI from the user is proposed.

All these works are optimizing resource allocation focusing on the trade-off between throughput and fairness. However, there are other important parameters related to the quality of service (QoS), such as latency or guaranteed bit rate (GBR), that must be taken into account. Also, the 3GPP has specified quality of experience (QoE) metrics that may be reported on a voluntary basis [6] and that may also be considered. Therefore, there is the need to develop new algorithms considering these requirements. Furthermore, the resource allocation problem must

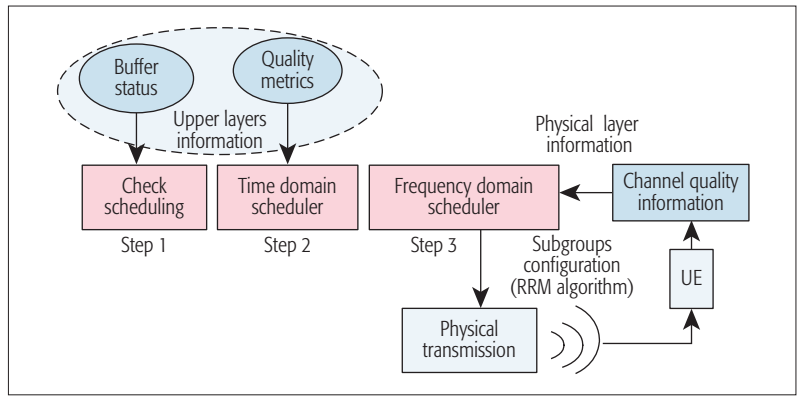


Figure 2. Block diagram of a packet scheduler that performs the resource allocation of an OFDM system. An adequate design of the scheduler will allow the system to improve the fairness among users while the total throughput is maximized.

be addressed for a single-cell scenario, where a single eNodeB is delivering multiple multicast services to users. In addition, new resource allocation strategies must be developed for heterogeneous multi-cell scenarios working in a coordinated SFN that exploits the benefits of improving the channel quality of the users by using MBSFN areas.

On the other hand, dynamic switching between unicast and multicast/broadcast transmissions makes possible the provision of broadcast services on demand, taking advantage of their scalability. First, they can be geographically localized, using broadcast transmission only where it is needed. Second, broadcasting can be used as much as it is needed, reserving a certain amount of bandwidth resources for these transmissions. Finally, the service can be switched to broadcast transmission solely in cases when it brings efficiency. This feature is called multicast operation on demand (MooD) and is introduced in [6]. MooD enables certain content that is initially delivered over the unicast network to be turned into a multicast transmission, in order to efficiently use network resources when the traffic volume exceeds a certain threshold. Since the current research on resource allocation strategies has been focused exclusively on broadcast/multicast transmissions, there is a need to consider dynamic switching in the algorithms to be developed.

NEW WAVEFORMS FOR CONVERGED SERVICES

It is well known that OFDM, the waveform used in the 4th generation of mobile communications (4G) (LTE and LTE-A), presents interesting characteristics to be used in wireless networks. However, the use of a time-domain rectangular window in OFDM has the disadvantage of requiring very strict time and frequency synchronization. This means that the addition of a cyclic prefix is mandatory, resulting in throughput loss. LTE is able to achieve this tight synchronization since the users are allowed to transmit, at the expense of exchanging energy-costly messages. These assumptions are no longer feasible when looking at expected use cases in future mobile communication systems such as MTC, IoT, or user-centric deployments. A system incorporating IoT devices should preferably allow them to

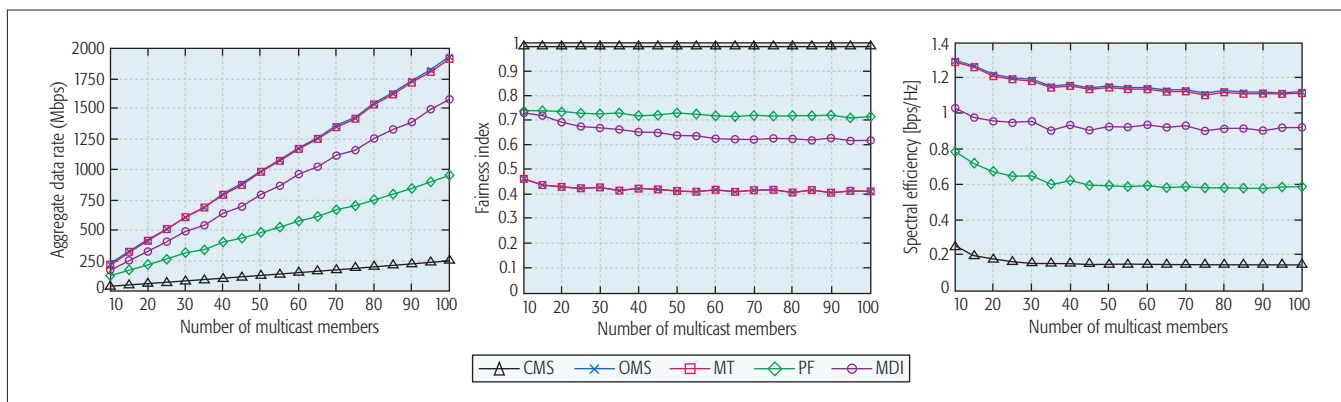


Figure 3. Performance analysis of different multicast resource allocation strategies studied in [4]: CMS, OMS, and multicast subgrouping with MT, PF, and MDI. The effect of increasing the multicast group size is shown for each of them in terms of: a) aggregate data rate; b) fairness index; c) spectral efficiency.

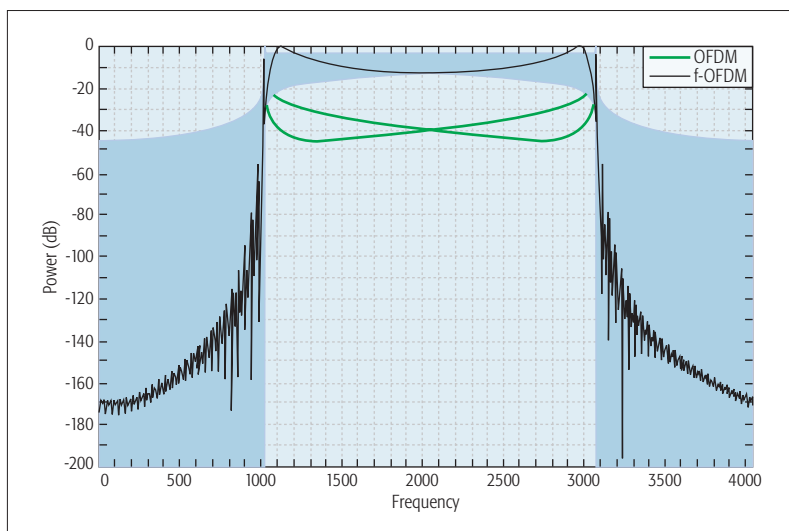


Figure 4. Comparison of the spectrum of classical OFDM and f-OFDM showing a better frequency confinement for f-OFDM.

transmit their messages without tight synchronization and using cheaper components [7].

4G deployments are based on devices connected to the network in a cell-centric way. It is worth mentioning that some important features of mobile communications such as coordinated multi-point (CoMP) [8], handover management, and offloading, may take advantage of the cell-centric concept. On the contrary, user-centric processing, when devices belong to multiple cells, leads to a disparity in the distances between the device and all access points whose respective carrier frequencies are also different. Consequently, tight synchronization may not be possible or cost-effective in a user-centric system.

These upcoming trends are leading the international community to the conclusion that the air interface of the 5th generation of mobile communications (5G) needs to lower the degree of synchronization that OFDM is currently demanding. As a consequence, in the last few years some alternatives to OFDM have been proposed, such as filter bank multi-carrier (FBMC), universal filtered multi-carrier (UFMC), or generalized frequency division multiplexing (GFDM) [7].

Nevertheless, the way these new waveforms

can be used to improve broadcasting in range, capacity, robustness, and utility remains today an open issue. For instance, longer ranges will be needed to provide converged TV services in 5G networks. To achieve the delivery of TV service up to a range of 60 km using the traditional OFDM waveform, an extended cyclic prefix of $200\mu\text{s}$ would be required, not only when traditional macrocells are used, but also when massive small cells are deployed, in order to benefit from SFN gain. With the current LTE parameters, this would imply only one OFDM symbol per subframe, consequently decreasing the LTE bit rate more than ten times. How this can be managed with the new waveforms is yet to be defined. An alternative waveform that may be able to cope with the requirements of IoT, and at the same time maintain the benefits of the cyclic prefix for broadcasting scenarios, is the recently proposed filtered OFDM (f-OFDM) [9]. In Fig. 4, the spectrums of classical OFDM and f-OFDM are compared, showing that the latter has better frequency confinement, which is more suitable for asynchronous IoT communications.

SPECTRUM SHARING AND AGGREGATION

The traditional regulation of spectrum use is based on two ends: exclusive use and license-exempt access. The idea of flexible licensing provides new opportunities in spectrum use for 5G systems by reusing parts of unused spectrum. New strategies are needed to support the variety and density of the upcoming wireless services and users, taking into account a good trade-off between cost and interference resilience, while ensuring service priority and spectrum availability. In particular, broadcasting presents the ability to use unpaired spectrum for the delivery of mass media or content. Therefore, spectrum sharing using flexible licensing allows the system to share some spectral resources to support the upcoming demand with the required QoS. At the same time, it opens up access to the unused spectrum bands in any location to be used in an opportunistic way by other 5G actors. Different regulatory solutions to achieve such flexibility have been proposed, such as light licensing, licensed shared access (LSA), and pluralistic licensing (PL) [2].

The light licensing approach consists of coordinated sharing between primary users (those

with higher priority or legacy rights on spectrum usage) and secondary users (those allowed to use the spectrum without interfering with primary users). The scalability needed in mobile services makes this approach inappropriate due to its limitations in transmission power [2]. As a consequence, this strategy can only be used in systems where interference is controlled.

The LSA approach authorizes additional licensed users to access primary users' spectrum under tight controls to prevent interference. The main goal of LSA is to provide additional shared spectrum usage in specific bands, while QoS for all rights holders is guaranteed.

PL is an innovative spectrum licensing approach that takes into account the requirements of primary and secondary users, providing fair use to both of them. With this technique, primary users can choose from a set of licenses, according to different rules that depend on the amount of interference they can tolerate. On the other hand, secondary users access the band in a cognitive way, observing the primary users' requirements.

These licensing approaches, in particular PL, make use of cognitive radio (CR) technology to improve spectrum use by means of dynamic spectrum access (DSA), where the unlicensed users access unoccupied licensed bands in an opportunistic way. It is worth noting that wide continuous spectrum bands are not often available, due to current regulations and policies. In this scenario, spectrum aggregation is an interesting solution. Different algorithms have been proposed to optimize spectrum assignment, maximizing the number of users, fulfilling the bandwidth requirements for secondary users, or combining with adaptive modulation and coding (AMC) to achieve higher network throughput [10]. However, these algorithms have not been designed with broadcasting applications in mind. Indeed, TV broadcasting has traditionally been considered a primary use of licensed bands. Broadcast/multicast, as part of future mobile services, opens up a new paradigm of secondary use of the spectrum, providing a more efficient use.

This cognitive approach can facilitate the use of small parts of unpaired spectrum. In the context of mobile broadcasting, as Fig. 5 illustrates, the use of supplemental downlink (SDL) has been proposed to increase downlink capacity by aggregating paired and unpaired spectrum [11]. In conclusion, the use of CR technology and spectrum aggregation brings new opportunities to enhance the capacity of mobile broadcasting.

SMALL CELLS WITH BROADCASTING FOR VENUE CASTING

The deployment of small cells in a SFN, under the umbrella of the existing macro cells, enhances venue casting performance. This deployment, which is presented in Fig. 6, yields robust coverage across the venue, maximizing the SFN gain by using more overlapping cells and improving the coverage offered by macro cells. This coverage improvement enables the transmission of higher-order MCS in the SFN area, hence increasing capacity.

The introduction of UDN in broadcasting scenarios offers more advantages, such as the opportunity to localize a specific broadcast service,

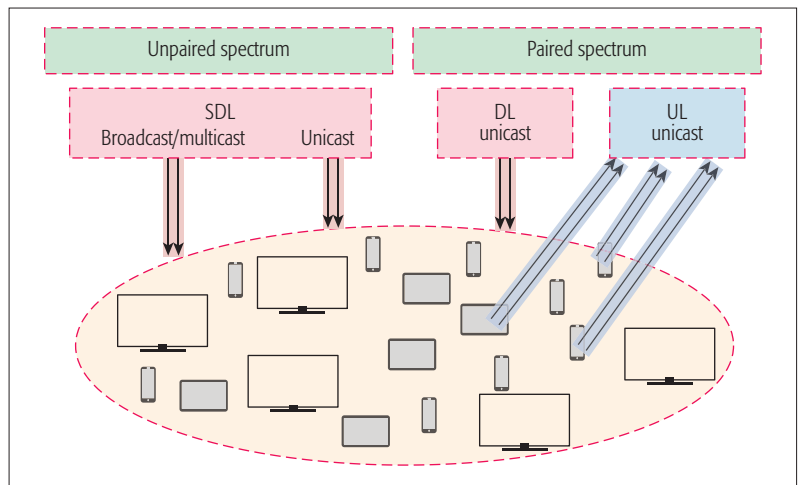


Figure 5. SDL deployment with paired and unpaired spectrum to deal with asymmetric downlink traffic in broadcasting services and increase the downlink capacity.

requiring only small cells for broadcasting and freeing up macro cells for unicast transmissions [12]. Small cells enhance the users' experience, providing better streaming performance consisting of more channels and content availability. The relevance of small cells when trying to provide ubiquitous coverage is shown in [1], where a new approach to coverage and subscribers' QoE analysis, the so called application coverage, is presented. As they show, different coverage is achieved depending on the application taken into account, with video being the most critical application. It is shown that with a macro cell deployment, it is not possible to offer the required quality for the video service. Including a micro cell improves the situation with 21 percent coverage of video, while the use of indoor small cells achieves 100 percent coverage of video services.

These enhancements in user experience affect not only venue users but also macro cell users of different services. Indeed, small cells can offload traffic from nearby macro cells, improving the availability of radio resources.

To fully leverage the advantages of small cells, adaptive resource allocation techniques must be designed for these specific scenarios. In a previous section, we discussed the resource allocation strategy of creating subgroups based on the link quality feedback of the users. In the context of using SFN in heterogeneous deployments, new constraints and conditions arise as compared to the macro cell scenario. When users are close to any of the small cell base stations, they will all share a high quality link, therefore they will be allocated to the same multicast subgroup. Consequently, the location information of the users, if available, can be an interesting input to enhance resource allocation strategies. In addition, this multi-tier deployment offers the flexibility of allocating some tiers to broadcast/multicast transmissions while some others are restricted to unicast. This flexibility must be considered when designing resource allocation algorithms to fully exploit these potential benefits.

Radio resource management may be facilitated by the emerging cloud radio access network (CRAN) architecture, where baseband data are

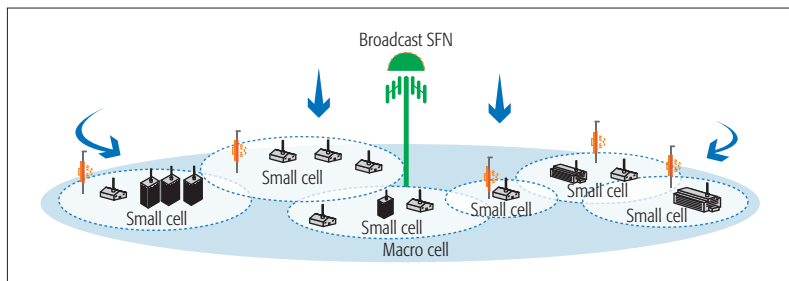


Figure 6. The deployment of small cells surrounded by existing macro cells enhances venue casting performance.

processed in a centralized way and distributed through a fiber or wireless front-haul to the small base stations. The possibility to centrally coordinate transmission has clear advantages for the deployment of SFNs. On the other hand, the limited capacity of the front-haul/back-haul may increase the latency and constitute a bottleneck for multicasting services. The joint optimization of caching and multicasting is a potential solution to improve the efficiency of massive content dissemination as shown in [13], where spatial content diversity is achieved in a large-scale heterogeneous network with back-haul constraints.

NEW APPLICATIONS ENABLED BY 5G BROADCASTING

These new technologies and trends for next generation mobile broadcasting will enable a wide variety of applications and services. The first 5G trials will be related to broadcasting at global events: in 2018, when the XXIII Winter Olympic Games begin in Pyeong Chang, South Korea intends to introduce a prototype 5G network, and a 5G trial network is planned to be launched for the Tokyo Summer Olympic Games in 2020.

Video services are driving the evolution toward 5G, opening up new business opportunities. A new class of services has been introduced with interactive and personalized TV services with cost-efficiency. This yields new business models based on revenue sharing, a leased/hosted network model, and others. These new business models can create new partnerships between operators, content owners, spectrum holders, and advertisers.

The Next Generation Mobile Networks (NGMN) Alliance, a joint effort of the major mobile operators, has specified the following minimum requirements for broadcast-like services:

- User experienced data rate of up to 200 Mb/s in the downlink (the maximum data rate is foreseen to distribute 4K/8K video).
- End to end latency smaller than 100 ms.
- User mobility up to 500 km/h.
- Device autonomy of several days up to years, depending on the use case (the longest autonomy is required for MTC devices).

In addition, 5G broadcasting can be extended for public safety applications, because mobile networks can provide these services efficiently in critical situations. This feature is called a group communication system enabler (GCSE), and its standardization has started with [14]. This service can be accessed by using push-to-voice/data/text, and the recipients can be dynamically

moved between broadcast and unicast depending on which transmissions achieve higher efficiency.

The new 5G architecture and radio interface will enable enhanced quality video broadcasting with ultra high definition television (UHDTV). Online content providers such as Netflix, Amazon, and YouTube are planning to release series and films in initial UHDTV formats such as 4K, as well as the evolution path toward 8K that is already defined. Indeed, emerging applications related to video, either complementing UHDTV or independently, such as telepresence, augmented reality (AR), and virtual reality (VR), will be enabled in the near future, and are considered to be one of the key use cases driving the requirements for 5G [15].

AR is a technique where the virtual world complements the real world. Its main playground is the real environment, where a virtual environment created by computer graphics is added. Overlaying digital information onto the real world, viewed through a camera-phone, is already possible today with multiple applications, but the business models and usage patterns are still evolving. Telepresence has received much attention in recent years as a means to provide innovative options to engage people and create a more collaborative environment. The mobile telepresence is a part of a thriving consumer robotics industry that is forecasted to reach US\$6.5 billion by 2017, according to a 2013 report of Oyster Bay. VR services can help people experience their presence in an imaginative world that looks real, while also giving them a chance to communicate with that world. Since VR emulates the real world, real-time video streaming should be very high in quality, so it is estimated that a 4–28 Gb/s data rate will be needed to transmit such video over the air, given that no or slight compression would be feasible.

All these new applications require very high bit rates and low latencies. As explained before, broadcast/multicast is a key enabler for achieving these requirements, making efficient use of the spectrum.

SUMMARY AND CONCLUDING REMARKS

Broadcasting, as we have seen, is an important driver to achieve efficient use of the spectrum in future mobile networks. The evolution of the broadcasting service will make it more dynamic and useful.

This article has detailed emerging technologies that need to be continuously enhanced to achieve the goals of next generation broadcasting. Dynamic resource allocation strategies for multicast are being developed to maximize the benefits of using multicast and broadcast transmissions in heterogeneous networks, where the service is delivered to users with different SINR. In addition, using dynamic switching between broadcast and unicast transmissions makes broadcasting ideal when and where it is needed, even if there are few users per cell demanding the service. This service can be delivered on demand, it can be used in more applications, and it can become more scalable. The requirements of the upcoming 5G networks are motivating the development of new waveforms that improve the efficiencies of traditional OFDM for some foreseen use cases. The development of a 5G

standard to deal with both mobile and broadcast industry demands requires the analysis of the implications of waveforms on broadcasting performance. It also requires novel schemes to enable spectrum sharing to optimize spectrum usage. The use of coordinated spectrum access and the sharing of infrastructure are required, instead of using competitive and interfering access. The utilization of small cells, together with the existing macro cells, enhances venue casting, improving coverage and thereby capacity everywhere, including the opportunity for better utilization of unlicensed spectrum. Finally, an efficient use of unpaired spectrum can be carried out with broadcasting services.

Together with these technological advances, it is worth mentioning that the broadcast industry and academia have been jointly working in 3GPP and DVB, proposing several technical approaches combining the existing infrastructure of both traditional broadcasters and mobile broadband providers to address both fixed and mobile delivery of the service.

With this work we would like to stimulate the ongoing research toward next generation broadcasting, making it capable of providing new services and fulfilling user demands.

REFERENCES

- [1] P. Cerwall *et al.*, "Ericsson Mobility Report: On the Pulse of the Networked Society," *Ericsson*, June 2015.
- [2] A. Kliks *et al.*, "Spectrum and License Flexibility for 5G Networks," *IEEE Commun. Mag.*, vol. 53, no. 7, July 2015, pp. 42–49.
- [3] TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2," 13.2.0 3, 3GPP Rel. 13, Dec. 2015.
- [4] G. Araniti *et al.*, "Evaluating the Performance of Multicast Resource Allocation Policies over LTE Systems," *2015 IEEE Int'l. Symp. Broadband Multimedia Systems and Broadcasting*, June 2015, pp. 1–6.
- [5] C. H. Ko *et al.*, "Strategy-Proof Resource Allocation Mechanism for Multi-Flow Wireless Multicast," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, 2015, pp. 3143–56.
- [6] TS 26.346, "Multimedia Broadcast/Multicast Service (MBMS); Protocols and Codecs (Release 13)," 13.2.0, 3GPP Rel. 13, June 2015.
- [7] G. Wunder *et al.*, "5G NOW: Non-Orthogonal, Asynchronous Waveforms for Future Mobile Applications," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 97–105.
- [8] R. Irmer *et al.*, "Coordinated Multipoint: Concepts, Performance, and Field Trial Results," *IEEE Commun. Mag.*, vol. 49, no. 2, Feb. 2011, pp. 102–11.
- [9] J. Abdoli, M. Jia, and J. Ma, "Filtered OFDM: A New Waveform for Future Wireless Systems," *2015 IEEE 16th Int'l. Wksp. Signal Processing Advances in Wireless Communications (SPAWC)*, June 2015, pp. 66–70.

- [10] S. Ping *et al.*, "SACRP: A Spectrum Aggregation-Based Cooperative Routing Protocol for Cognitive Radio Ad-Hoc Networks," *IEEE Trans. Commun.*, vol. 63, no. 6, June 2015, pp. 2015–30.
- [11] S. Yrjölä *et al.*, "Strategic Choices for Mobile Network Operators in Future Flexible UHF Spectrum Concepts?" *Int'l. Conf. Cognitive Radio Oriented Wireless Networks*, Springer, 2015, pp. 573–84.
- [12] "LTE Broadcast: Evolving and Going Beyond Mobile," *Qualcomm Technologies, Inc.*, Tech. Rep., 2014.
- [13] Y. Cui, D. Jiang, and Y. Wu, "Analysis and Optimization of Caching and Multicasting in Large-Scale Cache-Enabled Wireless Networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, July 2016, pp. 5101–12.
- [14] TS 23.468, "Group Communication System Enablers for LTE (GCSE_LTE); Stage 2," 13.3.0, 3GPP Rel. 13, Dec. 2015.
- [15] 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.

BIOGRAPHIES

ALEJANDRO DE LA FUENTE [S] (afuente@tsc.uc3m.es) is a Ph.D. candidate at the University Carlos III of Madrid in the multimedia and communications program. He received B.S. and M.S. degrees in telecommunications engineering from the Technical University of Madrid and an M.S. in multimedia and communications from University Carlos III of Madrid. He has worked for Ericsson Spain as a technical expert and product manager of telephony solutions. He joined the Communications Research Group of University Carlos III of Madrid in 2011. His research is focused in multicast/broadcast transmissions in 4G and 5G networks, especially in adaptive resource allocation strategies and cross-layer optimization techniques, with several contributions to international conference proceedings.

RAQUEL PÉREZ LEAL (rpleal@tsc.uc3m.es) received M.S. and Ph.D. degrees in telecommunications engineering from the Technical University of Madrid, and a postgraduate degree in management development from the IESE, Navarra University. Among other positions, she has worked as an expert on network and services solutions, as a manager of the Department of Network Technology at the Alcatel Corporate Research Center in Madrid, and as a manager of the Department of Advanced Development and R&D Strategy Areas at Alcatel Espacio. She joined the University Carlos III of Madrid in 2011. She has undertaken industrial and research projects and participated in a number of international projects (European Space Agency and European Framework Programs for Research and Technological Development), and nationally-funded projects in collaboration with research and industrial organizations. Her present interests include wireless communications networks and multimedia and multiuser distribution over wireless networks.

ANA GARCÍA ARMADA [SM] (agarcia@tsc.uc3m.es) received the Ph.D. degree in electrical engineering from the Technical University of Madrid in February 1998. She is currently a professor at University Carlos III of Madrid, Spain, where she has occupied a variety of management positions. She is leading the Communications Research Group at this university and the Laboratory of Communication Systems for Security and Space at the Scientific Park of the same university. She has participated in several national and international research projects related to wireless communications. She is the co-author of eight book chapters on wireless communications and signal processing. She has published more than 40 papers in international journals and more than 50 contributions to international conference proceedings. She has contributed to international organizations such as ITU and ETSI. Her main interests are OFDM and MIMO techniques and signal processing applied to wireless communications.

The broadcast industry and academia have been jointly working in 3GPP and DVB, proposing several technical approaches combining existing infrastructure of both traditional broadcasters and mobile broadband providers to address both fixed and mobile delivery of the service.

ADVERTISING SALES OFFICES

Closing date for space reservation: 15th of the month prior to date of issue

NATIONAL SALES OFFICE

Marion Delaney
Sales Director, IEEE Media
EMAIL: md.ieeemedia@ieee.org

Mark David
Sr. Manager Advertising & Business Development
EMAIL: m.david@ieee.org

NORTHERN CALIFORNIA

George Roman
TEL: (702) 515-7247
FAX: (702) 515-7248
EMAIL: George@George.RomanMedia.com

SOUTHERN CALIFORNIA

Marshall Rubin
TEL: (818) 888 2407
FAX: (818) 888-4907
EMAIL: mr.ieeemedia@ieee.org

MID-ATLANTIC

Dawn Becker
TEL: (732) 772-0160
FAX: (732) 772-0164
EMAIL: db.ieeemedia@ieee.org

NORTHEAST

Merrie Lynch
TEL: (617) 357-8190
FAX: (617) 357-8194
EMAIL: Merrie.Lynch@celassociates2.com

Jody Estabrook
TEL: (77) 283-4528
FAX: (774) 283-4527
EMAIL: je.ieeemedia@ieee.org

SOUTHEAST

Scott Rickles
TEL: (770) 664-4567
FAX: (770) 740-1399
EMAIL: srickles@aol.com

MIDWEST/CENTRAL CANADA

Dave Jones
TEL: (708) 442-5633
FAX: (708) 442-7620
EMAIL: dj.ieeemedia@ieee.org

MIDWEST/ONTARIO, CANADA

Will Hamilton
TEL: (269) 381-2156
FAX: (269) 381-2556
EMAIL: wh.ieeemedia@ieee.org

TEXAS

Ben Skidmore
TEL: (972) 587-9064
FAX: (972) 692-8138
EMAIL: ben@partnerspr.com

EUROPE

Christian Hoelscher
TEL: +49 (0) 89 95002778
FAX: +49 (0) 89 95002779
EMAIL: Christian.Hoelscher@husonmedia.com

COMPANY	PAGE
Franhofer Heinrich Hertz Institute	Cover 4
National Instruments	3
IEEE GLOBECOM 2016	13
IEEE Member Digital Library	Cover 2
IEEE ComSoc Membership	Cover 3
IEEE ComSoc Membership	7
IEEE ComSoc Membership	63
IEEE ComSoc Training	15
IEEE ComSoc Training	115
IEEE ComSoc Training	135

CURRENTLY SCHEDULED TOPICS

TOPIC	ISSUE DATE	MANUSCRIPT DUE DATE
EDUCATION AND TRAINING: TELECOMMUNICATION STANDARDS EDUCATION	MAY 2017	DECEMBER 1, 2016
SOFTWARE DEFINED VEHICULAR NETWORKS: ARCHITECTURES, ALGORITHMS AND APPLICATIONS	JULY 2017	DECEMBER 1, 2016
BEHAVIOR RECOGNITION BASED ON WI-FI CHANNEL STATE INFORMATION (CSI)	OCTOBER 2017	FEBRUARY 1, 2017
ADVANCES IN NETWORK SERVICES CHAIN	SEPTEMBER 2017	FEBRUARY 1, 2017
EDUCATION & TRAINING: SCHOLARSHIP OF TEACHING AND SUPERVISION	NOVEMBER 2017	MAY 1, 2017

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

TOPICS PLANNED FOR THE DECEMBER ISSUE

RESEARCH TO STANDARDS – NEXT GENERATION IOT/M2M APPLICATIONS, NETWORKS, AND ARCHITECTURES

INTEGRATED COMMUNICATIONS, CONTROL, AND COMPUTING TECHNOLOGIES FOR ENABLING AUTONOMOUS SMART GRID

INTERNET OF THINGS (IOT)

RADIO COMMUNICATIONS

CONSUMER COMMUNICATIONS AND NETWORKING

AUTOMOTIVE NETWORKING AND APPLICATIONS

THE GLOBAL COMMUNITY OF COMMUNICATIONS PROFESSIONALS

Special Member Rates

50% off - Membership for new members. Offer valid through 15 August 2016.

Member Benefits

IEEE Communications Magazine
(electronic & digital delivery)

IEEE Communications Surveys and Tutorials
(electronic)

Online access to IEEE Journal of Lightwave Technology, IEEE OSA Journal of Optical Communications and Networking and IEEE RFID Virtual Journal

Member Discounts

Valuable discounts on conferences, publications, IEEE WCET Certification program, IEEE Training courses and other exclusive member-only products.



Join Now!

<http://bit.ly/1sdsNno>



If your technical interests are in communications, we encourage you to join the IEEE Communications Society (IEEE ComSoc) to take advantage of the numerous opportunities available to our members.

www.comsoc.org

HIGH-SPEED DSP PLATFORM



Flexible Platform for 100G Real-time Digital Signal Processing

Key Features

- Multi-100G real-time data access
- Compact and flexible solution, based on Micro-TCA chassis and plug-in boards
- 4-channel 56-GSa/s, 8-bit ADCs
- 4-channel 65-GSa/s, 8-bit DACs
- High-performance Virtex Ultrascale FPGAs
- Multi-100G QSFP28 interface to external hardware
- Multi-purpose, built-in Gigabit Ethernet (GbE) connection
- Ready-to-use synchronization, interface and control IP cores included

Applications

- Test of digital signal processing algorithms in real-time
- Online transmission performance evaluation of optical transmission systems
- Realization of FPGA-based real-time transceivers