May 2017, Vol. 55, No. 5

IEEE Communications Magazine

- Enabling Mobile and Wireless Technologies for Smart Cities
- 5G Network Slicing: Concepts, Principles, and Architectures
- Green Communications and Computing Networks
- Telecommunication Standards Education

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

IEEE Communications Magazine

MAY 2017, vol. 55, no. 5
www.comsoc.org/commag

Enabling Mobile and Wireless Technologies for Smart Cities: Part 3

Guest Editors: Ejaz Ahmed, Muhammad Imran, Mohsen Guizani, Ammar Rayes, Jaime Lloret, Guangjie Han, and Wael Guibene

5G Network Slicing – Part 1: Concepts, Principles, and Architectures

Guest Editors: Konstantinos Samdanis, Steven Wright, Albert Banchs, Antonio Capone, Mehmet Ulema, and Kazuaki Obana

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

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

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

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

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

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

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

Telecommunication Standards Education

Guest Editors: Tarek S. El-Bawab, Periklis Chatzimisios, and David G. Michelson

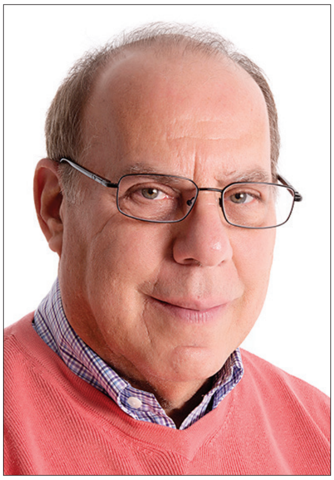
Green Communications and Computing Networks

Series Editors: Jinsong Wu, John Thompson, Honggang Zhang, RangaRao Venkatesha Prasad, and Song Guo

Sister Society Paper

Accepted from Open Call

WHAT'S NEW IN COMSOC CONFERENCES?



Harvey Freeman

Guoliang (Larry) Xue

Conferences are a very important part of ComSoc operations. We showcase cutting-edge technologies at our many sponsored global events. Every year, thousands of participants come to locations around the world to discover new ideas, explore innovative solutions, and network with the leaders who are changing the world of communications. Leading these efforts for the past year and a half is Larry Xue.

Guoliang (Larry) Xue is an IEEE Fellow and a Professor of Computer Science and Engineering at Arizona State University. He received the Ph.D. degree in computer science from the University of Minnesota in 1991, and the M.S. and B.S. degrees from Qufu Normal University (China) in 1984 and 1981, respectively. He has published extensively in top journals such as *IEEE/ACM Transactions on Networking* and *IEEE JSAC*, and at premier conferences such as INFOCOM, MobiCom, NDSS, CNS, and ICC/GLOBECOM. He received Best Paper Awards at ICC 2012, GLOBECOM 2011, ICC'2011, and MASS'2011. He is well cited, with an H-Index of 48 and over 10,000 citations. He has been a ComSoc Distinguished Lecturer. He was a keynote speaker at LCN'2011 and ICNC'2014. Larry has extensive experience in conference organizations and ComSoc services and currently serves as VP–Conferences for the IEEE Communications Society.

High-quality technical conferences are very important to both researchers and practitioners in the fast-changing field of communications and networks. People attend conferences to present new research results, exchange ideas, learn new developments, and network with other attendees. Maintaining high integrity and reputation, identifying important emerging research topics, ensuring a fair review process for paper selection, and keeping registration costs affordable are key factors for the success of ComSoc conferences. As a researcher and a long-time ComSoc volunteer, I have contributed to ComSoc conferences in various capacities such as an attendee, author, TPC member, TPC chair, general chair, and steering committee member. Serving as VP–Conferences gives me the opportunity to have a more comprehensive view of conferences and to work with the conference leadership team, ComSoc staff, and many dedicated volunteers to improve ComSoc conferences. In this column I will report what is new in ComSoc conferences with the hope you find these exciting, and continue to attend and contribute to ComSoc conferences. More importantly, I look forward to hearing from you about areas of concern and constructive suggestions for improvement.

IMPROVED STRATEGIC PLANNING FOR GLOBECOM/ICC

The IEEE Global Communications Conference (GLOBECOM) and the IEEE International Conference on Communications (ICC) are ComSoc's two flagship conferences. Each of these annual conferences attracts around 2,000 attendees from around the world. These flagship conferences feature Keynote Presentations from world-class scholars and executives of leading companies, industry forums, and technical sessions. For example, GLOBECOM 2016 attracted 2,288 scientists, researchers and industry professionals from around the world in Washington, DC last December to attend more than 1,500 technical and industry presentations and an exhibition with 21 exhibitors. In May of last year, ICC 2016 attracted nearly 2,000 scientists, researchers and industry professionals in Kuala Lumpur, Malaysia to attend over 1,500 technical and industry presentations. As there are many young professionals and students attending the conferences, special sessions for young professionals were organized.

GLOBECOM and ICC are under the leadership of GIMS (GLOBECOM/ICC Management & Strategy Standing Committee) and GITC (GLOBECOM/ICC Technical Content Standing Committee). GIMS is responsible for site selection, conference finances, strategic evolution, and operational policies of GLOBECOM and ICC. GITC is responsible for providing strategic vision and management of the technical content of GLOBECOM and ICC. GIMS and GITC have worked hard to further improve the planning and operations of GLOBECOM/ICC. In the past, each proposal for future GLOBECOM/ICC includes its own proposed TPC Chairs, in addition to conference site and organizing committee. In the future, proposals for GLOBECOM/ICC conferences will no longer include TPC chairs. Instead, GITC will work with the winning team to select the TPC chairs. We believe this will help optimize the selection of the TPC chairs for the winning team. Volunteers interested in organizing future GLOBECOM or ICC conferences should submit proposals to GIMS.

INNOVATIONS IN THE PAPER REVIEW PROCESS FOR INFOCOM

INFOCOM is one of the core conferences of ComSoc, with an acceptance ratio around 19%. Each year, INFOCOM receives around 1,600 submissions, of which around 300 papers are accepted for presentation at the conference and publication in the conference proceedings. Recent INFOCOM conferences have attracted around 900 researchers and professionals. In an innovative effort to enhance the fairness of the paper selection process, the INFOCOM Steering Committee decided to experiment with a double-blind review process with an automated system for optimal review assignment, together with a peer rating system to recognize top performing TPC members.

The motivation for INFOCOM's change of the paper review process is to enhance fairness and optimize the suitability of the review assignment. Previously, the review assignment was a paper-claim based assignment: each TPC member claims a set of papers to review on EDAS, based on the title/topic/abstract of the papers; EDAS does the review assignment based on this input and other constraints such as review load and potential conflict-of-interest (COI). However, the process of claiming papers is both error prone and time consuming. To ensure that the EDAS algorithm finds a reasonable assignment, TPC members need to claim a set of papers that is much larger than their maximum review load. This is very time consuming to the TPC members. As a result, some TPC members failed to find the time to enter the required number of claims. As fewer paper claims were received, the review assignment became more random. It was also noted that the paper claim process is not resilient to collusion, where a small group of TPC members attempt to claim papers based on authorship rather than review suitability. Also, TPC members tend to provide more objective reviews when the author names are not available so that the reviews are based solely on the quality of the paper. These considerations led the INFOCOM steering committee to experiment with an innovation in its review process.

INFOCOM's automated review assignment system, Erie, is described in an article published in the September/October 2016 issue of IEEE Network. It requires the following set of input data: (a) a list of reviewers, including their EDAS IDs as unique identifiers; (b) a list of submitted papers, including their paper IDs and PDF files; (c) a COI matrix between reviewers and the submitted papers; (d) for each reviewer, a list of his/her papers that best represent his/her expertise, in the form of PDF files. With the above input data, Erie can compute the suitability score between each submitted paper and each reviewer by comparing the PDF file of the submitted paper and the set of representative papers of the reviewer. This is accomplished by computing the similarity score between the submitted paper and each of the representative papers of the reviewer. The largest of these similarity scores is defined as the suitability score between the submitted paper and the reviewer. Erie finds the assignment that maximizes the total suitability for all submitted papers, subject to a number of constraints, including COI and review load, etc. It turns out that this integer linear programming can be solved in polynomial time, as the constraint matrix is totally unimodular. INFOCOM has experimented with double-blind review using Erie to make review assignment for the Area Chairs and all three TPC review assignments in the past two years. In order to incentivize the reviewers to provide high quality reviews, INFOCOM also experimented with a peer-rating system. For each paper, the three TPC members rate each other's review anonymously. The system aggregates peer-rating scores for all papers for each TPC member. Top TPC members based on peer-rating are recognized.

I wish to point out that each conference is unique in its scope, objective, and author base. Hence, each conference's steering committee needs to decide what is best for its conference.

Deep Industry Involvement and Startup City at WCNC

In recent years, attendance of most ComSoc conferences has been dominated by people from academia. ComSoc President Harvey Freeman mentioned multiple times that we need to strike a balance between academia and industry in our conferences. I am excited to report that the 2017 edition of the IEEE Wireless Communications and Networking Conference (WCNC), chaired by Andrea Goldsmith and Sarah Kate Wilson, achieved this goal. WCNC'2017 was a grand success, attracting over 1,000 researchers and professionals in San Francisco this March. Attendees packed the room to enjoy the enlightening plenary lectures, featuring top executives from leading wireless communications companies such as Qualcomm, Nokia Bell Labs, Assia, Huawei, China Mobile, Intel, Ericsson, and National Instruments. In addition to the high quality technical programs and informative panels, there were three industry forums and a Startup City track of presentations and demos by 15 startup companies in wireless communications. While there were a significant number of attendees from industry, the participation from academia remained strong. Many leading scholars, including several members of the National Academy of Engineering, attended the conference. It was great to see many leaders from both academia and industry at the conference.

Following are some of the tips the general chairs of WCNC 2017 shared:

•Deep engagement of industry starting early in the conference planning phase and through the event by forming an industrial advisory board 12 to 18 months before the conference, and appointing either a conference co-chair or an industry program co-chair who is well connected to industry

•Startup city as a platform to showcase the hottest wireless technologies in both an exhibition and a contest for the most innovative startups. This innovation made the conference more exciting. Connections between the conference organizers and the startups is key to this success

•Enlightening plenary talks by top executives. Connections between the conference organizers and high-level executives is key to this success

•A comprehensive set of industry-focused panels as well as workshops and tutorials strengthens the technical program

•A comprehensive program for students: a mentoring lunch with senior academics and industry folks; student poster/demo sessions with awards; student-industry networking session

•The use of professional conference planners provided by ComSoc, as the staff brings experience in conference planning/managing exhibits, which helped to ensure the success of the conference.

IEEE's Initiative in IoT

With the explosive growth in the number of smart devices connected via the Internet, Internet of Things (IoT) is entering our daily lives. Wireless communications and networking are fundamental to IoT. As a result, ComSoc is playing an active role in IEEE's IoT initiative, a multi-society undertaking with participation and sponsorship from 22 societies and councils. IoT is still in its early stages. It has experienced rapid growth, is multi-disciplinary, and is likely to have a profound effect on the world economy in the next few decades. At this stage, the identified verticals include: manufacturing; transportation; logistics; agriculture; aviation and defense; financial services; oil and gas; utilities; consumer electronics; health care; education; construction; and major movements such as smart cities, smart grid, and smart home. The identified topical areas include: big data; data processing; security and privacy; artificial intelligence and automation; cloud, fog and embedded computing; integration and design methods; virtual and augmented reality; connectivity; systems engineering; software and algorithms; sensors and sensor systems; integrated actuators; interfaces and displays; virtualization; and many more.

To create meaningful interaction and dialog among participants and to attract both practitioners and policy makers, the IEEE IoT is pioneering a new format for its meetings and conferences. The first is building events around a limited number of Verticals and a limited number of Topics important to selected Verticals. The second is balance between industrial, public sector, and researcher participation. The third is geographic balance, holding events in different regions of the world and selecting Verticals that are attractive for drawing local participation. The format for each event is to include overview material that is educational, exchange information on successes, difficulties, and challenges, and expose advanced ideas and concepts, but in the end use the events to drive recommendations that will result in future actions benefitting all three constituencies.

As an example, the University of Alaska, Anchorage, the State of Alaska, and IEEE IoT Activities are organizing the IEEE IoT Vertical and Topical Summit in Anchorage on September 18-20, 2017. The verticals were specifically picked because of local interest and the unique character of Alaska and other High Latitude Regions. These include the role of IoT in the following verticals and the unique aspects that go along with them: education; health care; oil, gas, and natural resources; aviation and unmanned aerial vehicles; and arctic region challenges. The topical areas this summit addresses include: security and privacy; communications and connectivity; big data and analytics; economic and societal impacts of IoT; and policy and regulations.

During the Summit the participants will be exposed to general aspects of IoT, and to local plans and challenges in Plenary and Keynote sessions. The Vertical and Topical Tracks will be used to present and discuss material at a much more granular level and to act as forums for conducting a dialog between experts in various aspects of IoT with practitioners and policy makers charged with implementation and deployment. Each Track will include a working group roundtable to identify courses of future action that will accelerate IoT adoption.

CONCLUSIONS

Delivering high quality conferences takes the collective efforts and collaboration from many people: conference organizers, ComSoc staff, authors, reviewers, patrons, and more. Above all, a successful conference should serve its attendees by offering high value for attendance. In this fast changing world, with new technologies emerging frequently, our conferences need to make corresponding changes as well in order to better serve our members. There are opportunities and challenges in front of us. We need to fix the problems and seize the opportunities. While the conference leadership teams and the steering committees of the conferences are working hard on strategic planning for the conferences, the wisdom of the crowd is more powerful than the wisdom of a few people. If you have areas of concerns or constructive suggestions to improve existing conferences or to propose new conferences, do not hesitate to contact the corresponding steering committee, any member of the conference organization team, or me. Let's work together to make our conferences better.

CANDIDATES ANNOUNCED FOR BOARD OF GOVERNORS ELECTION

Dear ComSoc Member,

In the following paragraphs you will find the position statements and biographies of an outstanding slate of candidates to lead the IEEE Communications Society. Your vote is very important to the individual candidates and to ComSoc as a whole.

Ballots will be emailed or mailed to all ComSoc members on 15 May 2017. We encourage your careful consideration as you cast your vote for the future success of the Society. The election begins 26 May 2017, and ends 21 July 2017.

In addition to the Vice Presidents' slate, each ballot will contain three slates for our Members-at-Large position:

a) one composed of six candidates from NA/LA (the Americas); b) one composed of three candidates from EMEA; and c) one composed of three candidates from AP regions. All voting members may select up to two from the NA/LA slate, up to one from the EMEA slate, and up to one from the AP slate.

The top two vote-getters from the NA/LA slate, the top vote-getter from the EMEA slate, and the top vote-getter from the AP slate will serve for a three-year term on the Board of Governors starting 1 January 2018.

If you do not receive a ballot email from ieee-comsocvote@ieee.org on 15 May 2017, but you feel your membership was valid before 1 May 2017, you may email ieee-comsocvote@ieee.org or call +1 732 562 3904 to check your member status and request a ballot. (You should provide your member number, full name, and address.)

Thank you.
Sergio Benedetto
Past President & Chair, Nominations & Elections

CANDIDATES FOR VICE PRESIDENT

TECHNICAL & EDUCATIONAL ACTIVITIES

Nelson L. S. da Fonseca

Candidate's Statement

Technical Activities not only represent the lifeblood of ComSoc, but also provides the foundation and manpower for building the ComSoc portfolio as well as for furnishing valuable services to our members. Educational Activities leverage the valuable knowledge produced in ComSoc activities to our members. I have had the privilege to serve ComSoc in different capacities, including in membership services, publications and conferences. It will be my honor to continue serving ComSoc members while sustaining its role as a major resource supporting members' needs throughout the world. If elected, I will:

- Enhance our technical activities to better address the needs of current and prospective members.
- Identify emerging technical areas, and facilitate ComSoc's entry into these field, promoting cooperation among technical committees and across Societies.
- Empower technical committee by increasing their participation and influence in ComSoc activities.
- Promote joint activities among technical committees and ComSoc chapters.
- Expand ComSoc's educational offering to members for professionals, practitioners and academics, and chapters by enlarging ComSoc's training program, free tutorials, and webinars.
- Expand activities for students such as student competitions.
- Strengthen ComSoc's globalization, promoting inclusion of volunteers around the world, young members and women.

Biography

Nelson L. S. da Fonseca received his Ph.D. degree from The University of Southern California. He is a full professor at the Institute of Computing of the State University of Campinas, Brazil. He has published 350+ papers and supervised 60+ graduate students. He is a ComSoc Distinguished Lecturer. Nelson has over 10 years of experience as a technical committee officer, and he served as chair of two technical committees: the Multimedia Communications TC and the Communications Systems Integration and Modeling TC. For his exemplary services, he received the Distinguished Service Award from these two TCs. Currently, Nelson serves as Vice President–Publications. He has also served as VP–Members Relations, Director of Conference Development, Director of the Latin America Region, and Director of On-line Services. He is a past EiC of *IEEE Communications Surveys and Tutorials*. He is a senior editor of *IEEE Communications Surveys and Tutorials* and a senior editor of *IEEE Communications Magazine*. He is the recipient of the 2012 IEEE Communications Society Joseph LoCicero Award for Exemplary Service to Publications, the Medal of the Chancellor of the University of Pisa, and the *Elsevier Computer Network Journal* Editor of the Year 2001 award. He created the ComSoc Student Competition Program, IEEE LATINCOM, the series of Multimedia Communications Symposia at Globecom/ICC, and the NFV-SDN conference.

Luigi Fratta

Candidate's Statement

The success of the IEEE Communication Society depends on its members' active participation. At present, I am serving ComSoc as VP–Technical and Educational Activities for 2016–2017. During this period, I have started a number of initiatives to re-energize technical and educational activities for our members. I have worked hard to make ComSoc an effective organization to promote communications research and developments by catalyzing cooperation among technical committees and emerging technical subcommittees. If re-elected, I will follow through on my new initiatives and expand ComSoc's role as a major resource to support our members' professional growth and career development.

Specifically, I will devote my second term to the following critical areas:

- Ensure that the Technical Committees keep abreast of the latest technology developments and evolution.
- Keep up with evolving technology advancements and nurture new ad hoc technical subcommittees.
- Encourage cooperation between academia and industry.
- Propose new interdisciplinary and emerging technical areas to engage younger members from both industry and academia.
- Support education and training programs in new initiatives that are beneficial to our members

Biography

Luigi Fratta received his doctorate degree in electrical engineering from Politecnico di Milano, where subsequently he was a professor for over 45 years. While at Politecnico di Milano, he led a number of national and international funded research projects in communications and networking. He also held several visiting positions around the world, including at the University of California at Los Angeles, the University of Hawaii, the University of Canterbury, New Zealand, Imperial College, United Kingdom, IBM T. J. Watson Research Center, IBM San Josè Research Laboratory, Bell Communications Research, and NEC Network Research Lab, Japan. He has been actively consulting with major telecom companies such as Siemens, Italtel, Alcatel, and Vodafone. Prof. Fratta is an IEEE Life Fellow (2009). He has been a dedicated volunteer of ComSoc for more than four decades and is currently serving as IEEE ComSoc Vice President–Technical and Educational Activities (2016–2017). His past ComSoc experience includes Technical Program Chair of IEEE INFOCOM '92 and a number of other conferences such as the IEEE LAN MAN Workshop, PIRMC '96, MMT '99, and NETWORKING 2009. He is a member of the Steering Committee of MEDHOCNET and IEEE INFOCOM. He has served on the editorial boards of several journals, including *Computer Networks*, *Wireless Networks*, and *Photonic Network Communications*, and he has been a co-guest editor for *IEEE JSAC* and the *Journal of Communications and Networking*. He has authored more than 150 refereed papers and holds five patents.

Hikmet Sari

Candidate's Statement

Technical and Educational Activities represent some of the most important activities of the IEEE Communication Society, which are essential to the professional growth of our members. As a long time volunteer of ComSoc, I have had the privilege of serving on various committees and leadership positions, and I would be honored to serve our society and membership as Vice President–Technical and Educational Activities for 2018–2019. My long experience in both industry and academic institutions, as well as my diverse activities in ComSoc, give me a very good understanding of the expectations of our members and ways to improve our services to them. If elected, I would be committed to:

- Work to strengthen the globalization effort of IEEE ComSoc and the representation of members from different regions in various committees.
- Help ComSoc address new and emerging topics and make it more relevant to diverse membership from academic institutions, research organizations, and industry.
- Improve our educational offerings to professionals and practicing engineers while preserving and further growing our value to academics and to the research community.
- Promote fairness, transparency, diversity, and rigor in Distinguished Lecturer selection, Fellow evaluation, and awards programs.

Biography

Hikmet Sari is a professor at Centrale Supelec and chief scientist of Sequans Communications. Previously, he held various research and managerial positions at Philips, SAT (SAGEM Group), Alcatel, Pacific Broadband Communications, and Juniper Networks. He received his engineering diploma and Ph.D. from ENST, Paris, and his habilitation degree from the University of Paris-Sud. His distinctions include the IEEE Fellow Grade and the Andre Blondel Medal in 1995, the Edwin H. Armstrong Achievement Award in 2003, and the Harold Sobol Award in 2012. He has served the IEEE Communications Society in numerous volunteer and leadership positions, including Vice President–Conferences, Distinguished Lecturer, member of the IEEE Fellow Evaluation Committee, member of the Awards Committee, member of several Technical Committees, Chair of the GITC, Chair of the Communication Theory Symposium of ICC 2002, Technical Program Chair of ICC 2004, Executive Chair of ICC 2006, General Chair of PIMRC 2010, General Chair of WCNC 2012, Executive Chair of WCNC 2014, Executive Co-Chair of ICC 2016, editor of *IEEE Transactions on Communications*, associate editor of *IEEE Communications Letters*, and guest editor of *IEEE JSAC*. Presently, he is serving as Executive Chair of ICC 2017.

MEMBER & GLOBAL ACTIVITIES

Nei Kato

Candidate's Statement

IEEE is truly a global organization with a vast membership. The global members of the IEEE share a critical bond that holds our community together. As VP–Member & Global Activities, I aim to reach out to our members by understanding and addressing their exact needs. Therefore, I will emphasize effective integration of member engagement across all the operating units of IEEE. I would like to organize meetings and events to link the local groups, sections, chapters, sister and related societies, and so forth, so as to bring the unique perspectives, relationships, expertise, and cultural perceptions to a broader level and make a bigger impact. Furthermore, my many years of experience in IEEE through various roles, particularly as Member-at-Large on the Board of Governors of ComSoc and Chair of the IEEE Sendai Chapter, will immensely help me to adequately focus on membership development by offering activities and networking possibilities in the local geographic units. I will ensure that members and volunteers have their voice heard so that they can flourish locally as well as globally. I would like to encourage the members to appreciate what IEEE has to offer to advance their professional careers and make their global presence truly felt by becoming active participants in various IEEE agendas.

Biography

Prof. Nei Kato received his bachelor degree from Polytechnic University, Japan, in 1986, and his M.S. and Ph.D. degrees in information engineering from Tohoku University in 1988 and 1991, respectively. He became a full professor at Tohoku University in 2003, and then the Director of Research Organization of Electrical Communication in 2015. He has been engaged in research in wireless mobile communications, satellite communications, ad hoc, sensor, and mesh networks, smart grid, IoT, big data, and pattern recognition. As a ComSoc volunteer, he has served as a Member-at-Large on the Board of Governors, a Distinguished Lecturer of IEEE, Chair of the Satellite and Space Communications TC, and Chair of the Ad Hoc & Sensor Networks TC. He has also served many times as General/TPC/Symposium Chair for IEEE conferences. His awards include the Outstanding Service and Leadership Recognition Award 2016 from the IEEE ComSoc Ad Hoc & Sensor Network TC, the Distinguished Contributions to Satellite Communications Award from the ComSoc Satellite and Space Communications TC, and so forth. He is the editor-in-chief of *IEEE Network*, Chair of the IEEE ComSoc Sendai Chapter, a member of the Awards Committee of ComSoc, and a Fellow of IEEE.

Rulei Ting

Candidate's Statement

Educated both in the United States and China, I have conducted research, engineering. and business efforts in both the United States and the Asia Pacific region. I value my personal experiences in state-of-the-art research and the engineering community as well as in a fast developing technical community. I have devoted my volunteering efforts to have both sides of the world benefit from each other. Working with ComSoc volunteers, I strive to increase services to our members at the local Chapter level, including professional education, publications, standards, student exchange via travel, conference participation, Chapter activity support, member services, and many more. As ComSoc's Director of Educational Services and IEEE's Co-Chair of the 5G Education WG, I have been working with many volunteers to bring distinguished speakers to our regional technical tutorial sessions, summer schools, and summits, and to special sessions of major conferences, in addition to the web-based education and training series. I would like to work with all of our ComSoc members to expand into regional activities. I have demonstrated passion and results in the telecom profession and volunteering for ComSoc. I hope that my background and experience will continue to be an asset to the IEEE Communications Society.

Biography

Rulei Ting has been with AT&T & Bell Labs, New Jersey, for 20+ years, with responsibilities from Distinguished Member of Technical Staff to senior technical director. He contributed to FT2000, which became an outstanding technology and industry success, delivering multi-billion dollars in revenue. He pioneered AT&T's business development in the AP region and has served as senior director in telecommunication equipment startups. He was awarded the AT&T Bell Labs President Award; the IEEE Millennium Award; the IEEE Region1 Award; and the President's Volunteer Service Gold Award of the United States. Over the past 15 years, he has volunteered as a ComSoc Chapter Chair, IEEE NJ Coast ExCom and Treasurer, and in 2003, his Chapter received ComSoc's Chapter Achievement Award. He led ComSoc's Engineering Certification and Education efforts. As WCET Committee Chair, he motivated a global team of volunteers developing and constructing the exam, setting up the strategies and launching partnership efforts. His team's efforts led to the financial turnaround of WCET. He has been Director of Educational Services of ComSoc since 2016, as well as Co-Chair of the IEEE 5G Education WG, establishing education and training programs all over the world. He earned his B.S. from Shanghai Jiao Tong University, China, and his Ph.D. from the City University of New York, New York. He earned an executive master's in technology management at Wharton and Penn Engineering of the University of Pennsylvania.

CONFERENCES

Stefano Bregni

Candidate's Statement

ComSoc is our global home. In 23 years as an enthusiastic volunteer, I have contributed concretely to our community. As a Distinguished Lecturer, I visited 14 countries and 29 Sections worldwide, preferring areas in Latin America and Asia where students have fewer opportunities to attend global conferences. As Vice-President–MR/MGA, I have a solid record of accomplishments, and I facilitated participation from all countries, also considering economic barriers. I fostered regional conferences. I am on the LATINCOM Steering Committee. I called four Regional Chapter Chair Congresses. I created and still lead the ComSoc Student Competition. I redesigned the *Global Communications Newsletter*. I strengthened the Women in Communications Engineering Committee, also launching the new Child Care Program for parents attending ICC/GLOBECOM. I strongly opposed the recent increase of ComSoc membership fees, including students. My contribution to ComSoc Conferences is solid. In GITC, I led the group defining the current ICC/GLOBECOM paper review process. As Conference TP Chair, I strictly ensured that paper selection always followed transparent procedures. I led the work to bring ICC 2016 to Kuala Lumpur, a magnificent but affordable location, which resulted in one of the best conferences ever. I have served on several Organizing Committees and consolidated remarkable practical experience. If elected, I commit to:

- Guarantee the quality and technical excellence of ComSoc Conferences, improving procedures for organizers, authors, and attendees, and ensuring fair paper review.
- Lower expenses for attending flagship conferences, by limiting registration fees and favoring affordable locations in all Regions.
- Foster regional conferences.
- Ensure balanced programs for academia and industry.
- Support and involve students.
- Promote women's engagement and ensure equal opportunities.

Quality and technical excellence are our best assets. I will defend them.

Biography

Prof. Stefano Bregni is with Politecnico di Milano, Italy. He graduated in electronics engineering (1990). After nine years in industry, he joined Politecnico di Milano in 1999. He has contributed to ETSI/ITU-T standards. He is an author of 90+ papers and of a recognized book on synchronization (Wiley, 2002). In ComSoc, he was a Distinguished Lecturer for seven years (2003–2009) and served as Vice President–Member and Global Activities (2016–2017), Vice President–Member Relations (2014–2015), Board of Governors Member-at-Large (2010–2012, 2013), Director of Education (2008–2011), GLOBECOM/ICC Technical Content (GITC) Committee member (2006–2009), TAOS TC Chair (2008–2009; Secretary/Vice-Chair 2002–2007). He has been the editor of the *Global Communications Newsletter* since 2007. He was ICC 2016 TP Co-Chair, GLOBECOM 2012 and GLOBECOM 2009 TP Vice-Chair, LATINCOM 2011 TP Co-Chair, and Symposium Co-Chair for nine other ICC/GLOBECOMs. He received the 2014 ComSoc Hal Sobol Award for Exemplary Service to Meetings & Conferences.

Robert S. Fish

Candidate's Statement

I am honored to be nominated for ComSoc VP–Conferences. If elected, I will emphasize a full innovation life cycle approach to conferences: high-quality research conferences, emerging technology conferences, and events targeted at our practitioner and industry communities. ComSoc's reputation in conferences is exemplary, and we must keep it that way. We will maintain our position as the premier venue for academic research, while creating in our portfolio events that specifically target practitioners.

As VP, I will look at expanding ComSoc's conference portfolio to fully reflect the breadth of ComSoc's technical scope, membership, and geographic diversity. Recently, IEEE launched a number of initiatives in the areas of 5G, IoT, green ICT, big data, and network softwarization. ComSoc should respond by expanding related conferences and events.

Finally, there are priorities that will make our events more affordable and more rewarding. As GIMS Chair and VP–Industry and Standards, I led efforts to create a global footprint for our conferences and Industry Summits. I will continue this trend by:

- Bringing events closer to our membership to minimize travel costs.
- Leveraging industry relationships to secure patronage and thus make our conferences more affordable.
- Attracting young professionals through offering training and career development opportunities.
- Facilitating industry-academia relationships.

Biography

Robert S. Fish [SM] received his Ph.D. from Stanford University. He is on the faculty of Princeton University and is president of NETovations, LLC. From 2007 to 2010, he was chief product officer and senior VP at Mformation, Inc. From 1997 to 2007, he was vice president and managing director of Panasonic U.S. R&D laboratories. Previously, he was executive director, Multimedia Communications Research at Bellcore after starting his career at Bell Laboratories. He has numerous publications as well as 17 patents. He is currently ComSoc's VP–Industry and Standards Activities. He has served on the ComSoc Board of Governors. He was chair of the GLOBECOM and ICC Management and Strategy Committee (GIMS) where he created the current structure of the committee and reformed the site selection process. He was co-founder and Steering Committee Chair of the Consumer Communications and Networking Conference and serves on the founding Steering Committee of the Conference on Standards for Communications and Networking. He served on the BoG of the IEEE Standards Association (IEEE-SA), and was a founding member of the IEEE-SA Corporate Advisory Group. His awards include the Multimedia Communications Technical Committee Distinguished Service Award and the IEEE SA's Standards Medallion.

Guoliang (Larry) Xue

Candidate's Statement

High-quality technical conferences are important to researchers and practitioners in the fast-changing field of communications and networks. People attend conferences to present new results, exchange ideas, and learn new developments. Maintaining high integrity and reputation, identifying new research topics, ensuring a fair review process for paper selection, and keeping registration fees affordable are the key factors for the success of ComSoc conferences. As the current VP-Conferences of IEEE ComSoc, I want to have the opportunity to sustain the work we have done in improving ComSoc conferences. If elected, I will strive to:

- Improve fairness in the paper review and selection process.
- Improve the author and attendee experience.
- Develop new revenue structures to keep conference registration fees affordable.
- Achieve a healthy balance of participation from various sectors.
- Cultivate promising new conferences and nurture young and active ComSoc volunteers to become leaders.

Biography

Guoliang (Larry) Xue is an IEEE Fellow and a professor of computer science and engineering at Arizona State University. He received his Ph.D. degree in computer science from the University of Minnesota in 1991, and his M.S. and B.S. degrees from Qufu Normal University, China, in 1984 and 1981, respectively. He has published extensively in top journals such as *IEEE/ACM Transactions on Networking* and *IEEE JSAC*, and premier conferences such as IEEE INFOCOM, MobiCom, and ICC/GLOBECOM. He received Best Paper Awards at ICC 2012, GLOBECOM 2011, ICC 2011, and MASS 2011. He has been a ComSoc Distinguished Lecturer. He was a keynote speaker at LCN 2011 and ICNC 2014. He has extensive experience in conference organizations and ComSoc services. He served as Secretary/Vice Chair/Chair of the Communications Switching and Routing Technical Committee. He served as a TPC Co-Chair of INFOCOM 2010 Workshops, Co-Chair of GLOBECOM 2012, and Symposium Chair for ICC/GLOBECOM. He served as a General Co-Chair of CNS 2014. He is an area editor (Wireless Networking) of *IEEE Transactions on Wireless Communications* and an editor of *IEEE Network*. He served as an editor of *IEEE/ACM Transactions on Networking* and an editor of *Computer Networks*. He currently serves as VP–Conferences of the IEEE Communications Society.

Publications

Xuemin (Sherman) Shen

Candidate's Statement

ComSoc publications have some of the highest reputations and impact in the technical community and are a key pillar to serve our members. With the rapidly changing technical fields and the global community, it is very important to undertake challenges to meet the diverse needs and to provide the most benefits to our members. If elected, I will work passionately, in collaboration with editors-in-chief and editors, to:

- Enhance IEEE ComSoc's leading role in telecommunications by promoting its high quality in journals, magazines, and online content.
- Inspire cutting edge research activities and high impact papers, especially among junior researchers (including graduate students).
- Promote industry participation in periodical special issues on new technologies, standards, and practices.
- Provide strong support and encouragement to junior researchers to get involved in ComSoc journal/magazine editorial boards and the review process to benefit their career progression and professional development.
- Improve the financial plan for the publications' page budgets.

My extensive volunteer service in ComSoc and dedication to member benefits will enable me to effectively serve and promote the excellence of our community.

Biography

Sherman Shen (http://bbcr.uwaterloo.ca/~xshen) received his Ph.D. degree in electrical engineering from Rutgers University, New Jersey, in 1990. He is a professor and University Research Chair at the University of Waterloo, Canada. He has supervised 40 postdoctoral fellows, 65 Ph.D. students, and 50 M.A.Sc. students, and received the 2006 Excellent Graduate Supervision Award from the university. He has co-authored 22 books and over 500 refereed journal papers, with h-index 81 and 35,000+ citations, and received 30 Best Paper Awards. He has served as the editor-in-chief for four publications, including the *IEEE Internet of Things Journal* and *IEEE Network*. His recent ComSoc leadership positions include Distinguished Lecturer, elected Member-at-Large on the Board of Governors, Chair of the Distinguished Lecturer Selection Committee, Chair of the Wireless Communications Technical Committee, Technical Program Co-Chair of GLOBECOM 2016 and INFOCOM 2014, Steering Committee Chair of ICC in China (ICCC) 2015 and 2014, and Technical Symposia Chair of ICC 2010. His distinctions include the 2015 ComSoc Joseph LoCicero Award and the 2013 Technical Recognition Award from the ComSoc Ad Hoc & Sensor Networks Technical Committee. He is a Fellow of IEEE, the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada.

Chengshan Xiao

Candidate's Statement

I have had the privilege to serve the IEEE Communications Society as a dedicated volunteer in a variety of technical programs. I have worked in industry, academia, and government, received my education from both the northern and southern hemispheres, and lived in both Eastern and Western cultures. If elected as Vice-President–Publications, I will rely on my diverse technical background and global work experience to focus on the following three issues.

First, I will enhance the quality of ComSoc's journal and conference publications. I will do my best to improve the review process, reward excellent editors and reviewers, and promote high-quality papers.

Second, I will maximize the diversity of volunteers serving on editorial boards, and conference organization and technical committees. I will engage people from different geographic, gender, ethnic, experience, and occupational backgrounds in the decision-making process.

Third, I will expand opportunities for the younger generation. I will encourage early-career professionals to get involved in ComSoc journal editorial boards, technical committees, and conference organizations.

My rich experience and commitment to ComSoc services will enable me to effectively address important issues in ComSoc.

Biography

Chengshan Xiao received a bachelor of science degree from the University of Electronic Science and Technology of China, Chengdu, a master of science degree from Tsinghua University, Beijing, China, and a Ph.D. in electrical engineering from the University of Sydney, Australia. He was a senior engineer with Nortel Networks, Ottawa, Canada. He is now a professor at Missouri University of Science and Technology, and a program director at the National Science Foundation, United States. He has published over 90 technical journal papers and holds three U.S. patents. Two of his invented algorithms have been implemented in wireless base station products. He has received several distinguished awards, including the 2014 Humboldt Research Award, the 2014 IEEE Communications Society Joseph LoCicero Award, and the 2015 IEEE Wireless Communications Technical Committee Recognition Award. Dr. Xiao has served as the editor-in-chief of *IEEE Transactions on Wireless Communications* (2010-2013), was elected to the ComSoc Board of Governors (2012–2014), Director of Conference Publications (2014-2015), Fellow Evaluation Committee member (2011–2014), and Distinguished Lecturer (2011–2013) of the IEEE Communications Society. He served as the Technical Program Committee (TPC) Chair for the IEEE International Conference on Communications in South Africa, May 2010, and the Founding Chair (2007–2008) of the IEEE Technical Committee on Wireless Communications. He has also served in many other positions for ComSoc's journals, technical committees, and conferences. He is currently a TPC Co-Chair for the IEEE Global Communication Conference in Singapore, December 2017.

INDUSTRY & STANDARDS ACTIVITIES

Stefano Galli

Candidate's Statement

ComSoc is experiencing increasing difficulty in recruiting, retaining, and engaging industry members, especially those not directly involved in R&D. Understanding this trend and working toward reversing it is essential because the vitality of our Society, the impact of publications and conferences in the research community, and the value of networking with peers rest on an active and balanced presence of academics, industry/government researchers, practitioners, managers, professionals, and students.

I believe we need to rethink ComSoc's membership value proposition while working not only toward differentiating our products/services, but also toward engaging our existing industry members, especially the practitioners. Standards activities are an excellent way to engage industry members, and we also need to reach out with targeted engagement to other industry members not involved in standardization.

I believe that my long career in the industry, standards savviness, and my experience in ComSoc/IEEE have prepared me well for the role of ComSoc VP–Industry and Standards Activities. If elected, I will work to:

- Identify/track metrics for measuring industry participation and engagement.
- Recruit more industry members as ComSoc volunteers to better understand and address industry needs.
- Share best practices with other Societies to better understand what works best in attracting industry membership and participation.
- Define new products/services focusing on industry membership without compromising on technical excellence, which is sought by both academia and industry.
- Expand our standards portfolio going beyond the PHY/MAC areas, broaden the use of ComSoc's rapid reaction standards workshops, and foster initiatives aimed at identifying early standardization needs.
- Leverage my experience in both IEEE and non-IEEE standards to revitalize and broaden ComSoc standardization activities, involve students, and bring standards into continuing education.

Biography

Stefano Galli received his Ph.D. in EE from the University of Rome, Italy. Currently, he is a lead scientist at Huawei USA. Previous positions include director at ASSIA, director at Panasonic, and senior scientist at Bellcore (now Ericsson). He is currently the rapporteur/chair of the ITU-T Communications for Smart Grid standardization group and has often contributed to many standards: IEEE, ITU-T, ETSI, BBF, ATIS, and NICC. He is an IEEE Fellow, and serves as Fellow Committee Vice-Chair, and a member of the Governance Committee and the Board of Directors Ad Hoc Committee on Industry Engagement. He served as an elected Member-at-Large of ComSoc's Board of Governors and first chair/founder of ComSoc's Committee on Powerline Communications. His awards include the 2014 Broadband Forum Outstanding Contributor Award, the 2013 IEEE Donald G. Fink Best Paper Award, and the 2011 ComSoc Donald W. McLellan Meritorious Service Award.

Mehmet Ulema

Candidate's Statement

IEEE ComSoc is a world class organization for communications professionals. However, attracting industry practitioners remains a challenge. If elected, my highest priority will be to address this challenge, and build a strong community where researchers and practitioners work together. I have unique experience in both industry and academia, and a track record proving my ability to achieve this goal. During my two terms as the Director of Standards Development, the number of ComSoc standards projects has increased three-fold thanks to the innovative Rapid Reaction Standardization Activity (RRSA) methodology. Through my effort, ComSoc has become a major player in IEEE standards and is on the global standardization map.

If elected, I will strive to:

- Create a full-service ecosystem for standardization professionals, industry practitioners, and academic researchers by initiating and supporting standards-related ComSoc publications, conferences, and educational products.
- Strengthen industry-academia partnerships in ComSoc standards activities by leveraging the ComSoc RRSA mechanism.
- Expand industry outreach and industry communities in emerging technology areas such as IoT, 5G, and SDN/NFV, especially to attract young professionals to IEEE Initiatives.
- Work with other VPs to create Industry Summits, practitioner-focused publications, meetings, and training.

I believe I am uniquely qualified for this role and would appreciate your vote.

Biography

Mehmet Ulema received his Ph.D. from Polytechnic Institute, Brooklyn, and his B.S. and M.S. from Istanbul Technical University. He is a professor at Manhattan College, New York, and has been a consultant to several companies. Previously, he held management and technical positions at AT&T Bell Labs and Bellcore, involved in numerous telecom projects. While working in industry and academia, he was actively involved in standardization in ITU, TIA, ATIS, and IEEE. More recently, he founded and has been leading the IEEE P1903 project on Next Generation Service Overlay Networks. He has been a prominent volunteer in ComSoc and the IEEE Standards Association. Currently he is the Director of Standards Development in ComSoc. He is also a member of the IEEE SA Standard Board, which helped ComSoc become a major player in IEEE standards governance. He has authored numerous industry reports and scholarly publications. He has been on the editorial boards of several major journals. Previously, he held leadership positions in ComSoc, including Director of Membership Programs. He also served in leadership positions in major IEEE conferences, including GLOBECOM and ICC. He has received several awards, including the IEEE SA Millennium Award and the ComSoc Harold Sobol Award for Exemplary Service to Meetings and Conferences.

Candidates for Members-at-Large

Asia/Pacific Region (one will be elected)

Sumei Sun

Candidate's Statement

I have been an active volunteer in ComSoc and IEEE over the last decade. I have served in various roles, including ComSoc Asia Pacific Board Vice Director (2016–2017), Information Services Committee Chair (2014–2015), ComSoc Singapore Chapter Vice Chair (2015–2016), Executive Vice Chair of GLOBECOM 2017, Symposium Co-Chair of ICC 2015 and 2016, Publicity Co-Chair of PIMRC 2015, editor of *IEEE Wireless Communication Letters* (2011–2016), and editor of *IEEE Communications Surveys & Tutorials* since 2015. My voluntary services have helped me develop a good understanding of the fundamental needs of our members, and the importance as well as challenges of supporting our community. If elected as a Member-at-Large, I will focus on promoting the excellence of our community and commit myself to advocate for the best interests of our members. In particular, I will make my best efforts to:

- Strengthen industry involvement in ComSoc activities.
- Promote the values of ComSoc to its existing and potential members, from both academia and industry, through ComSoc programs and services that are made easily accessible and affordable.
- Engage and support female researchers and young professionals to participate in ComSoc activities, increase their representation, and promote inclusive recognition of their achievements.

Biography

Sumei Sun [F] is head of the Communications and Networks Cluster, Institute for Infocomm Research, Agency for Science, Technology, and Research, Singapore. Working in this publicly funded research institute for 20 years, she has a good understanding of both academic and industry research, and she contributed to bridge theoretical research to industry applications. She has been actively contributing as a volunteer to ComSoc and IEEE, organizing conferences in different roles, including Executive Vice Chair of GLOBECOM 2017, Symposium Co-Chair of ICC 2015 and 2016, Track Co-Chair of IEEE VTC-Spring 2014 and VTC-Fall 2016, Publicity Co-Chair of PIMRC 2015, and so on. She has been an editor of *IEEE Communications Surveys & Tutorials* since 2015, and was an editor for *IEEE Wireless Communication Letters* during 2011–2016. She is also serving as the Vice Director of the ComSoc Asia Pacific Board (2016–2017).

Li-Chun Wang

Candidate's Statement

ComSoc needs a bridge to cross technological and international borders. Bridging the gap between academia and industry is increasingly important for advancing state-of-the-art research toward marketable products and services. Bridging the gap between communications and other research areas is crucial for the future of communications researchers and engineers in the era of big data. Bridging the gap among all the regions in the world is a vision of IEEE. As a Member-at-Large, I will be your bridge to cross the aforementioned gaps of various regions, research areas, and societies based on the following IEEE strategies:

- Identify new opportunities for interdisciplinary research.
- Engage more friends around the world.
- Enhance global cooperation by sharing information and initiating activities.
- Expand ComSoc's impact on other IEEE Societies.

I will commit myself to listening to your voices and being an advocate for your best interests. I sincerely hope to receive your support and have the honor to serve you.

Biography

Li-Chun Wang [F'11] received M.Sc. and Ph.D. degrees from Georgia Institute of Technology in 1995 and 1996, respectively. He was affiliated with Nortel, Texas (1995), and AT&T Laboratories-Research, New Jersey (1996–2000). Currently, he is jointly appointed by the ECE and CS Departments of National Chiao Tung University, Taiwan, where he chaired the ECE Department (2012–2016). He was names an IEEE Fellow for his contributions in cellular architectures and radio resource management. He won the IEEE Jack Neubauer Best Paper Award (1997), the Taiwan NSC Distinguished Research Award (2012), and the ComSoc APB Best Paper Award (2015). He served as Treasurer of the IEEE Taipei Section (2009–2010), and Chapter Chair for the Vehicular Technology (2003–2005) and Information Theory (2007–2009) Societies. He represented the Taipei Chapter in accepting ComSoc's and ITSOC's Chapter of the Year Award (2007/2009), and helped the Taipei Section win the IEEE MGA Outstanding Large Section Award (2009).

Wei Zhang

Candidate's Statement

I am honored to be nominated for a Member-at-Large from the AP (Asia Pacific) region. If elected, I will work passionately to serve ComSoc by:

- Attracting new members by developing and promoting new programs, especially for students, young professionals, and those in industry.
- Maintaining the excellence of ComSoc publications while maximizing the impact of the research outputs to our society by leveraging my experience as Editor-in-Chief.
- Providing support for pursuing and nurturing new technology directions through establishing new technical committees.
- Promoting industry participation by broadening and strengthening our linkages to industry through new collaborative programs.

Biography

Wei Zhang (F'15) has been heavily involved with IEEE ComSoc by serving in numerous volunteer and leadership positions, including Vice Director for IEEE ComSoc Asia Pacific Board, Vice Chair for IEEE ComSoc Wireless Communications Technical Committee, a Distinguished Lecturer of IEEE ComSoc, Symposium Co-Chair of ICC 2011, Workshop Co-Chair of GLOBECOM 2017, and TPC Co-Chair of APCC 2017. He is the editor-in-chief of *IEEE Wireless Communications Letter*, and an editor for *IEEE Transactions on Communications* and *IEEE Transactions on Cognitive Communications and Networking*. He also served as an editor for *IEEE Transactions on Wireless Communications* and *IEEE Journal on Selected Areas in Communications – Cognitive Radio Series*. He is an IEEE Fellow. He is the recipient of four IEEE Best Paper Awards and the 2009 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award. He received the Ph.D. degree from the Chinese University of Hong Kong in 2005. Currently, he is an associate professor at the University of New South Wales, Sydney, Australia.

Europe/Middle East/Africa Region (one will be elected)

Gerhard Fettweis

Candidate's Statement

ComSoc should continue to be the world's most important Society for R&D communications engineers. We need to take responsibility for defining technology roadmaps, providing leadership to industry, funding agencies, and venture capitalists. We determine the economic future of countries worldwide. We must continue driving the excellence of our publications and conferences, ensuring that we attract the most important innovations for publication within ComSoc's media, being a catalyst for innovations and continued-education.

Biography

Gerhard Fettweis earned his Ph.D. under H. Meyr's supervision from RWTH Aachen in 1990. After one year at IBM Research, San Jose, California, he moved to TCSI, Berkeley, developing digital cellular chipsets. Since 1994 he has been the Vodafone Chair Professor at TU Dresden, Germany, with 20 companies from Asia/Europe/the United States sponsoring his research on wireless transmission and chip design. He coordinates two DFG centers, cfaed and HAEC. His team spun out 16 startups, and set up funded projects of e0.5 billion. He co-heads the 5GLabGermany. He co-chairs the IEEE 5G Initiative. He is an IEEE Fellow, a member of the German Academy of Sciences ("Leopoldina") and the German Academy of Science and Technology ("acatech"), has been a Distinguished Speaker of IEEE Societies, and has received awards including a Dr.h.c. from TU Tampere, ComSoc's Harold Sobol Award, and the VDE Medal of Honor. He was TPC–2012 and 2014. He ran a TwinLab with Masdar Institute, Abu Dhabi. He was elected Member-at-Large of IEEE SSCS (1999–2004) and ComSoc (1998–2000, 2011–2013, 2015–2017). He was associate editor for *IEEE JSAC* (1998–2000) and *IEEE Transactions CAS-II* (1993–1996). From 1991 to 1998 he was ComSoc's delegate within the IEEE Solid State Circuits Council. He has served on the IEEE Corporate Innovation Award Committee, GIMS (2012 chair), the IEEE Fellow Committee, and the Steering Committees of *IEEE Transactions on Communications* and *IEEE Wireless Communications Letters*. During 2008–2009 he chaired IEEE IT Society's Germany Chapter.

Andrzej Jajszczyk

Candidate's Statement

The IEEE Communications Society operates in a challenging global environment. If elected, I will work to improve and enhance ComSoc's activities and adapt them to the rapidly changing needs. I will also work to diversify and improve the quality of services offered by our Society to its members, including dissemination of technical information, services via the web, the Distinguished Lecturer program, awards, as well as publications, conferences, and workshops. As I have been active in ComSoc over many years, I well understand its role in serving its members. Since I have spent several years as a visiting scientist in North America, Australia, and France, and have been a technology consultant in Australia, Canada, China, France, Germany, India, Poland, and the United States, I am in a good position to know the needs of communications professionals in different geographic areas.

Biography

Andrzej Jajszczyk [M'91, SM'95, F'99] is a professor at AGH University of Science and Technology, Krakow, Poland. He graduated from Poznan University of Technology. He is the author or co-author of 12 books and over 300 papers, as well as 19 patents in the areas of telecommunications switching, high-speed networking, and network management. He was the founding editor of the *IEEE Global Communications Newsletter*, an editor for *IEEE Transactions on Communications*, and editor-in-chief of *IEEE Communications Magazine*. He has held important positions in IEEE Communications Society, such as Director of Magazines, Director of the Europe, Africa, and Middle East Region, and Vice President–Technical Activities. He is currently Director of the Europe, Middle East, and Africa Region. He has served on Technical Program and Steering Committees of numerous conferences and was a ComSoc Distinguished Lecturer. He is a member of the Polish Academy of Sciences and a member of the Scientific Committee of the European Research Council. He is Vice-President of the Kyoto-Krakow Foundation, fostering cultural and scientific relations between Asia and Poland.

George K. Karagiannidis

Candidate's Statement

I have been active in the IEEE Communications Society for over 20 years, serving as editor or guest editor of several ComSoc journals and magazines, as editor-in-chief of *IEEE Communications Letters*, and as Technical Program Committee member and Program Chair of major ComSoc conferences. Also, I served the community as Chair for four years of the IEEE ComSoc Greek Chapter. If I am elected Member-at-Large, I will do my best to promote the leading role of ComSoc in both academia and industry, especially to support junior and female researchers and engineers. Furthermore, my objective is to exploit my long experience in publications in order to improve ComSoc's journals and conference publications. Finally, I will try to contribute in bridging the digital divide in the EMEA Region, mainly through free and low-cost online education services for members.

Biography

George K. Karagiannidis joined the Aristotle University of Thessaloniki, Greece, in 2004, where he is currently a professor of digital communications systems in the Electrical & Computer Engineering Department and director of the Digital Communications Systems and Networks Laboratory. He has been involved as General Chair, Technical Program Chair, and member of Technical Program Committees in several IEEE and non-IEEE conferences. In the past, he was an editor of *IEEE Transactions on Communications*, senior editor of *IEEE Communications Letters*, and a frequent guest editor of *IEEE Journal on Selected Areas in Communications*. From 2012 to 2015 he was the editor-in-chief of *IEEE Communications Letters*. From 2007 to 2011 he was Chair of the IEEE ComSoc Greek Chapter. He is a highly cited author across all areas of electrical engineering, recognized as a 2015 and 2016 Thomson Reuters highly cited researcher.

America (LA/NA) Regions
(one will be elected)

Tarek El-Bawab

Candidate's Statement

As a Member-at-Large, I will represent the interests and aspirations of ComSoc membership. Our field has evolved and expanded rapidly. ComSoc should:

- Attract more youth/students and encourage them to volunteer and lead.
- Motivate the telecom industry to be further involved in our activities.
- Adapt products and services optimized for our community's needs, and ensure affordable, member-friendly delivery to all.
- Collaborate with other IEEE Societies and other organizations to better serve our global communities.

For nearly 25 years, I have acquired in-depth knowledge of ComSoc's technical committees, education, conferences, publications, and standards. I have the experience and academic/industrial background to be instrumental, if elected, in serving our membership in all areas of our Society's activities.

Biography

Tarek El-Bawab is a professor with Jackson State University, West Virginia. Before this, he was with Alcatel-Lucent, Colorado State University, and the University of Essex, United Kingdom. Earlier, he led international industrial projects for 10 years. He is a member of IEEE and Eta Kappa Nu. He is an IEEE Distinguished Lecturer, the editor of Springer's Series: Textbooks in Telecommunication Engineering, and associate editor-in-chief of *IEEE Communications Magazine*. His research interests include networks and their enabling technologies. He has 70+ publications and a reference book. He led a movement that resulted in ABET's recognition of telecommunication engineering as a distinct education discipline. He received ComSoc's first ever education award for this work and for making changes to our education system that benefit our profession and society. He is a member of the IEEE Educational Activities Board and ComSoc's Educational Services Board. He served ComSoc as a Board of Governors member, Director of Conference Operations, TAOS' Chair (two terms), and Chair/Organizer in several ICC/GLOBECOM conferences. He is also a member of the IEEE Computer, Electron Devices, Photonics, and Education Societies. He has B.Sc. and B.A. degrees from Ain Shams University, an M.Sc. from the American University in Cairo, an M.Sc. from the University of Essex, and a Ph.D. from Colorado State University.

Gustavo Giannattasio

Candidate's Statement

As a leading organization, ComSoc faces increasing challenges that need our attention in the areas of membership, volunteer base, WIE, YP, practitioners, and industry participation in emerging technologies such as 5G and IoT.

- Services, tools and programs need to be updated to address emerging technologies, and strong leadership to attract volunteers, professionals, students, WIE, and practitioners is needed.
- Worldwide education summer schools, hackathons, excellence in publications in emerging technologies, attracting industry patrons, and a larger volunteer base will give us the boost needed to continue being a strong organization.

Over 28 years in industry and academia, and as Region Director, a member of the IoT and Smart Cities Steering Committees, head of the MOOC on emerging Smart Cities technologies, a Program Chair of Conferences, editor-in-chief of the *IEEE Wireless Engineering WEBOK*, and an invited editor of the *IEEE Systems Journal*, I believe that together we will support ComSoc's growth in several directions.

Thank you for your support and vote.

Biography

Gustavo Giannattasio [SM] graduated from the University of Uruguay as a telecommunications engineer, received an M.B.A. from Catholic University Uruguay, and a postgraduate degree in electronic engineering from Philips Institute Holland. For 10 years he has been a Project Manager Professional certified PMI USA. For 20 years he has been a professor of communications and networking at Catholic University, ORT, and UDELAR Uruguay. He is an online course professor, NGN, for ITU Switzerland. He is head of the Smart Cities Department, ACCSA Uruguay, Solutions for Municipalities: Smart Waste, Urban Noise, Electronic Meters, Smart Traffic, Smart Parking, Air, Water and Quality Monitoring. He has been a ComSoc Chapter Chair twice, Section Chair of IEEE Uruguay twice, IEEE Region Latin America and Caribbean Director, a member of the IEEE Board of Directors, MGA, Committees: GOLD, Awards, Public Visibility, A&A, Smart Cities Steering, and IoT Standardization Committee. He was IEEE I2SCC Italy 2016 and ComSoc GIIS Mexico 2015 Program Chair. He spent four years as Program Director PMI Chapter, and a member of the Education Group, IEEE Project Management Council. He was editor-in-chief of the *ComSoc WEBOK WCET*. He has been a keynote speaker in IEEE and ITU conferences in the Bahamas, Mexico, and Colombia on SDN/NFV, protocol communications for smart grids, smart cities, and the Internet of Things.

Ekram Hossain

Candidate's Statement

I have been volunteering for IEEE and ComSoc for more than 16 years in different capacities (journal editor, guest editor for special issues, editor-in-chief, conference Technical Program Committee member/Co-Chair/Chair, ComSoc Technical Committee Executive, IEEE local section Chair). If elected, by leveraging this rich experience, I will work toward serving our community in both academia and industry in a better way. In particular, I will emphasize improving the effectiveness and affordability of ComSoc services and activities, including conferences, especially to student members and young professionals, improving the portfolio of ComSoc publications, maintaining the high standards/reputations of our publications, developing new services and activities, and reaching out to/engaging our colleagues in developing regions.

Biography

Ekram Hossain [F] is currently a professor of electrical and computer engineering at the University of Manitoba, Canada (home.cc.umanitoba.ca/~hossaina). He is a member of the College of the Royal Society of Canada (https://rsc-src.ca/en/college-new-scholars-artists-and-scientists). He received his Ph.D. in electrical engineering from the University of Victoria, Canada, in 2001. He has more than 16,000 citations and has won several research awards, including the IEEE ComSoc Fred W. Ellersick Prize Paper Award, the IEEE WCNC '12 Best Paper Award, and the IEEE VTC-Fall '16 Best Paper Award. He serves as an editor for *IEEE Wireless Communications* and IEEE Press. Previously, he served as the editor-in-chief of *IEEE Communications Surveys & Tutorials*, an area editor for *IEEE Transactions on Wireless Communications*, an editor for *IEEE Transactions on Mobile Computing*, and an editor for the *IEEE Journal on Selected Areas in Communications–Cognitive Radio Series*. He was a Distinguished Lecturer of the IEEE Communications Society (2012–2015). He serves/has served as the Technical Program Chair/Co-Chair for Symposiums/Tracks in VTC-Fall '17, WCNC '16, ICC '12, WCNC '12, PIMRC '11, VTC-Spring '10, and VTC-Fall '10. He has served as an IEEE ComSoc Chapter Chair and currently serves as the IEEE Winnipeg Section Chair.

Urbashi Mitra

Candidate's Statement

Communications systems have changed dramatically over the last 20 years from a former focus on point-to-point systems to today's need for complex networks with heterogeneous nodes to the future biological communication networks of my current research. My goals are to ensure that:

- The IEEE Communications Society broadens its scope to be the main architect of future communication systems, whether they be at terahertz frequencies, the glue to the wireless Internet of Things, or the basis of engineering solutions to biological problems.
- We continue our efforts to increase the number of excellent communications engineers through diversity efforts and outreach.
- The IEEE Communications Society publications continue to be the flagship purveyors of knowledge while supporting the society and facilitating publication for its authors.

Biography

Urbashi Mitra received her B.S. and M.S. degrees from the University of California, Berkeley, and her Ph.D. from Princeton University. She is currently a Dean's Professor of Electrical Engineering at the University of Southern California, and previously at the Ohio State University and Bellcore. Specific to IEEE ComSoc, she is the inaugural editor-in-chief of *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, a Distinguished Lecturer for the IEEE Communications Society, 2017

Vice Chair of the Communication Theory Technical Committee, 2017 Chair of the *Transactions on Wireless Communications* Steering Committee, a former associate editor of *IEEE Transactions on Communications*, and a Technical Co-Chair for the ICC 2003 Communication Theory Symposium. She received the 2016 ComSoc Women in Communications Engineering Mentoring Award. She is an IEEE Fellow. Other awards include: being a U.S. Fulbright Scholar, a U.S. National Academy of Engineering Lillian Gilbreth Lectureship, a GLOBECOM Signal Processing for Communications Symposium Best Paper Award, Texas Instruments Visiting Professor, the Okawa Foundation Award, the OSU College of Engineering Research Award, the OSU College of Engineering Teaching Award, and a NSF CAREER Award.

Aria Nosratinia

Candidate's Statement

I have been an active ComSoc member and volunteer for the majority of my 29 years as an IEEE member. If elected, I will bring to bear my extensive experience with IEEE publications and conferences as a volunteer, organizer, and editor on one hand, and an active consumer of these products on the other. I aim to facilitate the engagement of junior and mid-career ComSoc members in local Chapters as well as in ComSoc Technical Committees. I am interested in returning value to our members and others by exploring the reduction of page charges for journal papers and conference registration fees. Being a registered professional engineer as well as an academic, I am eager and deeply committed to serve the interests of our members regardless of their status as student, engineer, researcher, or academic. If elected I will continue to actively listen to our members for what they want from their representatives on the BoG.

Biography

Aria Nosratinia is the Erik Jonsson Distinguished Professor and associate head of the Electrical Engineering Department at the University of Texas at Dallas. He received his Ph.D. in electrical and computer engineering from the University of Illinois at Urbana-Champaign in 1996. He has held visiting appointments at Princeton University, Rice University, and the University of California, Los Angeles. He is an area editor of *IEEE Transactions on Wireless Communications*, and was formerly an editor of *IEEE Transactions on Information Theory*, *IEEE Transactions on Image Processing*, *IEEE Signal Processing Letters*, and *IEEE Wireless Communications*. He has received the National Science Foundation Career award, as well as IEEE Chapter Outstanding Service awards. He has served on the Technical Program Committees of numerous ComSoc conferences. He is a registered professional engineer in the state of Texas. In 2016 he was named a Thompson-Reuters highly cited researcher. He is a fellow of IEEE.

Kazem Sohraby

Candidate's Statement

ComSoc is a major professional organization for faculty, researchers in industry, and in particular for students in all areas of communications. However, this Society needs to work harder and make space for the voice and desire of all its membership, in particular students. I believe we need to find avenues for more and effective student and membership participation in various decisions that come to the Society's Board of Governors (BoG). For example, it would be effective to bring to the attention of its membership, through the ComSoc website, major issues prior to voting, and have comments and discussions by membership reflected in the BoG decisions. I would also like to create unconstrained/unlimited student access, at a low flat fee, to a broad range of resources such as IEEE Xplore products/services. I believe these steps would help increase student membership through their participation and give all members a more effective voice in running the Society.

Biography

Kazem Sohraby is a professor of electrical and computer engineering at the South Dakota School of Mines and Technology. Previously, he was a Distinguished Member of Technical Staff at Bell Labs, a professor of electrical engineering and director of the Wireless Networks and Systems Laboratory, and a professor and head of the Department of Computer Science and Computer Engineering at the University of Arkansas. He earned a Ph.D. in electrical engineering (New York University's Polytechnic Engineering Division), an M.S. in electrical engineering (Worcester Polytechnic Institute), and a B.S. with highest honors in electrical engineering (Amir Kabir University). He serves on the editorial and advisory boards of IEEE and non-IEEE publications, chaired IEEE INFOCOM 2012 and 1996, and the IEEE Electro Information Technology Conference in 2013. He has served on several IEEE ComSoc Technical Committees and as the Director of Online Content. He is currently ComSoc's Director of Conference Development.

Jinsong Wu

Candidate's Statement

If I am elected, I would like to solidly work with other ComSoc members on sustainability:

- Supporting analyses of and actions on the existing (and emerging) challenges of ComSoc.
- Attracting new members and retaining existing ones by improving the supporting environments, enhancing benefits, and developing value-added programs.
- Promoting new activities and benefits for engaging students and young professionals as ComSoc members.
- Promoting and accelerating ComSoc's responses to new interdisciplinary and emerging topics.
- Encouraging and extending our ComSoc activities and efforts toward environmental, social, and economical sustainability goals.
- Growing ComSoc values to academics while intensifying and encouraging the participation of industry.
- Increasing our global impact via supporting emerging technology areas, enabling collaboration with various IEEE initiatives, IEEE Societies and Councils, as well as other relevant non-IEEE international organizations

Please find more details at http://www.cec.uchile.cl/~jinsongwu/Statement.pdf

Biography

Jinsong Wu was the leading editor and co-author of the book *Green Communications: Theoretical Fundamentals, Algorithms, and Applications* (2012), the first comprehensive book covering the broad topics of green wireless communications, green wireline communications, and relevant general topics. He is elected Vice Chair-Technical Activities, IEEE Environmental Engineering Initiative under IEEE Technical Activities Board (TAB). He is the founder and founding Chair of the ComSoc Technical Committee on Green Communications and Computing, and the primary co-founder and Vice-Chair of the ComSoc Technical Committee on Big Data (TCBD). He was the first proposer (2012) and long-term promoter of IEEE green journals and transactions. He is the founder of the *IEEE Communications Magazine* Series on Green Communication and Computing Networks. He was one of the key proposers and a long-term promoter (2012–2015) of going from the Green Track to a dedicated Green Symposium within ComSoc flagship conferences. He has opened the Big Data Track with a very general topic coverage in flagship conferences of IEEE Communications Society, starting at GLOBECOM 2016. Please find more details at http://www.cec.uchile.cl/~jinsongwu/Biography.pdf

5G Mobile and Wireless Communications Technology

By Afif Osseiran, Jose F. Monserrat, and Patrick Marsch (eds.), Cambridge University Press, 2016, ISBN 978-1-107-13009-8, hardcover, 406 pages

Reviewer: Rafal Krenz

4G wireless communication systems are well established on the market, and currently, the next generation is among the hottest research topics, therefore a related book is welcome. Here, we have a position written by leading experts in 5G research, and it is a comprehensive overview of the current state of the 5G landscape. The contents are based on several deliverables of the EU project, Mobile and Wireless Communication Enablers for the Twenty-Twenty Information Society (METIS), which was accomplished between 2012 and 2015 by 25 companies and institutions from all over Europe.

After a short introduction in Chapter 1, the following 13 chapters reflect the results of the work packages of the METIS project. Chapter 2 provides the most relevant 5G use cases along with the challenges and requirements to be met in order to fulfill the expected needs of end users. Next, the 5G system concept is described, which generalizes the key characteristics of the use cases and specifies three generic communication services: Extreme Mobile Broadband (xMBB), Massive Machine-Type Communication (mMTC), and Ultra-reliable Machine-Type Communication (uMTC). In Chapter 3, the proposed architecture of future networks is presented, with emphasis on the key enablers, such as Network Function Virtualization (NFV) and Software Defined Networking (SDN), which are needed to provide the required flexibility, especially for the core network. Chapter 4 is devoted to Machine-Type Communication and addresses its requirements, fundamental techniques (e.g., transmission of short packets and non-orthogonal protocols for distributed access), as well as corresponding design principles and technology components. Chapter 5 focuses on direct Device-to-Device (D2D) communication, which is one of the important cornerstones to improve system performance and to support new services beyond 2020 in 5G systems by local management of short-distance communication links and local traffic offloading. The 10-fold increase of data rates and 1000-fold increase of traffic volume predicted in 5G systems requires access to a new piece of spectrum, improved spectral efficiency, and network densification enabled by the use of small cells. As explained in Chapter 6, this will be possible through the use of centimeter (above 6 GHz) and millimeter waves, and the adoption of beamforming, both very challenging for the hardware. Chapter 7 presents novel radio-access technologies that fulfill a number of diverse requirements of various services foreseen for 5G systems, demanding more flexibility and scalability, e.g., filtered multi-carrier and non-orthogonal multiple access schemes. The aforementioned increased spectral efficiency per cell will be possible with the application of massive MIMO, explained in Chapter 8, and adds many spatial degrees of freedom provided by a large number of antennas, allowing multiple users to communicate with use of the same time-frequency resource. Chapter 9 discusses Coordinated Multi-Point (CoMP) techniques (e.g., joint transmission, dynamic point selection, and coordinated scheduling/beamforming), where multiple nodes in the network cooperate to limit the impact of interference, and consequently increase the overall system capacity. Chapter 10 is devoted to relaying and network coding that will be used to improve the performance of 5G networks. Three new key techniques, i.e., multi-flow relaying, non-orthogonal multiple access, and buffer-aided relaying, which help to regain the spectral efficiency loss caused by in-band half-duplex relays, are described. Chapter 11 deals with network-level solutions aiming at enhancing the end-user experience, such as new forms of 5G deployments (ultra-dense networks and moving networks), interference and mobility management, as well as dynamic network reconfiguration for energy-aware RANs. The possible choices for 5G spectrum and new spectrum access modes are examined in Chapter 12. The assessment of the overall performance and comparison of different 5G technology proposals would not be possible without new wireless propagation channel models. Two different approaches, i.e., stochastic and map-based, addressing specific features of 5G systems, are introduced in Chapter 13. Finally, in Chapter 14, simulation methodology, which ensures the consistency of results and allows for a direct comparison of different 5G technology components, is proposed.

The book may be a good starting point to become acquainted with novel transmission techniques and networking solutions proposed for 5G system. Nearly 600 references will help the interested reader dive deeper into a specific topic. It may also serve as a valuable source of information for graduate students specializing in mobile communications. I can truly recommend this book to the entire wireless communication community.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

May 2017
ISSN 2374-1082

MEMBERSHIP SERVICES

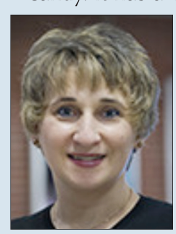
WICE: Promoting the Role of Women in Communications Engineering

Interview with Octavia Dobre, Chair of the WICE Standing Committee

By Stefano Bregni, Vice-President for Member and Global Activities, and Octavia A. Dobre, Chair of WICE Standing Committee

This is the fifth article in the series started in November 2016 and published monthly in the IEEE ComSoc Global Communications Newsletter, which covers all areas of IEEE ComSoc Member and Global Activities. In this series of articles, I introduce the six MGA Directors (Sister and Related Societies; Membership Services; AP, NA, LA, EMEA Regions) and the two Chairs of the Women in Communications Engineering (WICE) and Young Professionals (YP) Standing Committees. In each article, one by one they present their sector activities and plans.


Stefano Bregni


Octavia A. Dobre

In this issue, I interview Octavia A. Dobre, Chair of the IEEE ComSoc Standing Committee on Women in Communications Engineering (WICE). Octavia is a Professor and Research Chair with Memorial University, Canada. She is the Editor-in-Chief of *IEEE Communications Letters*, and has served as editor and senior editor for other prestigious journals. She also chairs the ComSoc Signal Processing for Communications and Electronics Technical Committee and has served as an officer for other technical committees, as well as technical co-chair of symposia at various conferences.

It is my pleasure to interview Octavia and offer her the opportunity to outline her current activities and plans for ComSoc WICE.

Bregni: Hello Octavia! This is the second time I have the opportunity to interview you for the GCN, after our first article in 2015, to present the activities of the IEEE ComSoc WICE Standing Committee. I am particularly glad I again have this chance. True equal opportunities for women in our field of engineering is one of the five strategic directions I have set in ComSoc Member and Global Activities, as I indicated since the beginning of my first term as Vice-President. Therefore, your Committee is key to this goal. Would you recall how it developed in the last years?

Dobre: I have been the WICE Chair since 2014, when it became a Standing Committee of ComSoc. In almost four years, the committee has been growing and its activities have been increasing significantly. It has a very active website (http://wice.committees.comsoc.org/) and Facebook page, as well as a very large number of subscribers to the mailing list.

It is a vivid committee, which organizes a significant number of activities to promote the visibility and roles of women communications engineers and to provide a venue for their professional growth. Certainly, any ComSoc member is welcome to the WICE events.

Bregni: What about the Board working with you on organizing the various WICE activities?

Dobre: WICE has seven positions on the Board, e.g., industry liaison, student activity coordinator, publicity chair, WIE committee society coordinator, as well as five members-at-large and a representative to the ComSoc Student Competition. The WICE members in these positions are very active, organizing various events and supporting other WICE initiatives.

(Continued on Newsletter page 4)


Photos from the Third Women's Workshop on Communications and Signal Processing at IEEE GLOBECOM 2016, Washington DC, US.

CONFERENCE REPORT

IEEE ComSoc Panama Chapter: Winner of the 2016 Chapter Achievement Award

Activities and Accomplishments in 2015-2016

By Maytee Zambrano N., Panama Chapter Chair

Communications Chapter Board 2015-1016. From the left Eng. Alkin Saucedo, Eng. Hector Saavedra, Dr. Maytee Zambrano, Eng. Jorge Ruiz and Dr. Carlos Medina.

During the visit of Dr. Tillman Wolf the chapter presented a certification for his distinguished conference.

Participants of the 5G half Day Conferences with José Otero from 5G Americas and Hector Marin from Qualquom Latin America.

One key element of current society and its development are the telecommunication systems. For instance, technologies such as 5G, wireless networks, and IoT are crucial for the development of smart cities. Therefore, it is important as an organization related to telecommunications to have a presence and participation in the community. This motivated our interest, in 2015, in reactivating the IEEE Panama Communications Chapter with a board made up of professionals with different backgrounds including industry, academia and government.

Our main objective during 2015-16 was to offer a variety of technical and non-technical activities that contribute to the professional enrichment of our local members and the local professional and student community in the field of telecommunications.

To reach our goal, we organized several technical activities during 2015. They included: two IEEE Distinguished Lecture Tours, one about "Applications of Field-Programmable Gate Arrays in Scientific Research" by Dr. Jinyuan Wu, and the other on "Attacks and Hardware Defenses for Network Infrastructure" by Dr. Tillman Wolf, and a technical conference "Mobile Communications: Evolution and Challenges" by Eng. Orlando Calderon, who is a technical account manager for Xceed-Anite Technologies.

Considering the fundamental role of telecommunications in almost all fields, we developed a half-day seminar about the fundamentals of telecommunications for non-engineering professionals. The audience targeted were professionals in marketing, sales, customer service and media IT industries, lawyers, managers and executives who needed a basic knowledge in telecommunications for their activities.

During 2015, we met with members of the board of the Panamanian Secretariat of Science, Technology and Innovation (Senacyt) and The Panamanian Chamber of Professional and Companies in IT Sector (CAPATEC) to promote the ComSoc chapter and attract new members, with a view to carrying out future joint activities and achieving synergy with other scientific and industry professionals.

In 2016, following the current trends in telecommunications, several technical activities were organized on hot topics, such as 5G, spectrum regulation, and smart cities. The distinguished lecturers for the 5G half-day conference were Jose Otero, Director of Latin America and the Caribbean for 5G Americas, and Hector Marin, Director for Business Development for Qualcomm in Latin-America and the Caribbean. The scope of this conference was the Evolution of 5G and the policies for its implementation, the roadmap for standardization, and the challenges and perspective for Latin America from a technical and commercial point of view.

Participants of the conference "Challenges of Panama city to become a Smart city" with the Schneider's representative and the Panama City Vice Major Raisa Banfield after the event.

To reach the undergraduate student community, the benefits of being a member of the IEEE Communications Society were explained to them during the 1st Congress of Electrical Engineering Students of Panama. In addition, a technical activity in spectrum management and local regulations was given for undergraduate students by Eng. Alkin Saucedo, Deputy Director of Radio, Television and Spectrum Management of the Panamanian Regulatory Agency.

Considering that the city of Panama growing rapidly, the Panamanian chapters of the Communications Society and the Power and Energy Society organized together a technical conference on smart cities in 2016. The objective was to review the challenges that the city of Panama faces in order to become a smart city. A Schneider's representative showed the different technologies they offer for smart cities and explained some of their local projects. Also, the Vice Major of Panama City, Ms. Raisa Banfield, presented the scope of projects of the major's office related to automation and data collection in the city. Attendees were students and professionals from academia and industry.

In 2016, the IEEE Communication Society recognized our work by granting us the "Latin-America Chapter Achievement Award" based on our 2015 accomplishments.

Our motivation behind each activity is to highlight the benefits of "being part of the top of the world: a member of the IEEE Communication Society."

CHAPTER REPORT

RNDM 2016 Workshop and COST CA15127-RECODIS 2nd Meeting

Highlights from the Resilience Week in Halmstad, Sweden

By Magnus Jonsson, Sweden, Jacek Rak, Poland, Dimitri Papadimitriou, Belgium, and Arun Somani, USA

Prof. Jacek Rak, RNDM 2016 General Chair, Sahel Sahhaf and Francesco Musumeci, recipients of RNDM 2016 Best Paper Awards, and Prof. Magnus Jonsson, RNDM 2016 Co-chair.

Leading network resilience researchers took part in the Resilience Week on Sept. 12-15, 2016 at Halmstad University, SE by Prof. Magnus Jonsson from the Centre for Research on Embedded Systems (CERES), Halmstad University, SE, and Prof. Jacek Rak from Gdansk University of Technology, PL. It included two major events:

•The 2nd Meeting of COST CA15127–RECODIS Action (Resilient Communication Services Protecting End-user Applications from Disaster-based Failures, http://www.cost-recodis.eu) held on Sept. 12-13, 2016.

•RNDM 2016–8th International Workshop on Resilient Networks Design and Modeling on Sept. 13-15, 2016.

The meeting of COST RECODIS was attended by approximately 45 researchers from 24 European countries and the US. Its agenda included sessions scheduled for meetings of the Management Committee (MC) and Working Groups (WGs). Special focus was on communication resilience issues related to disaster-based failure scenarios including large-scale natural disasters, weather-based disruptions, technology-related disasters and malicious human activities. The main objective was to summarize the results of the first stage of the Action (i.e., analysis of state-of-the-art solutions) as well as to discuss the respective research issues to be solved in the near future including tasks to be completed in the framework of Short Term Scientific Missions (STSMs).

The meeting of COST RECODIS was enriched by the invited talk by Prof. Biswanath Mukherjee (IEEE Fellow from University of California, Davis, US) on network resilience for massive failures and attacks.

The meeting of COST RECODIS was followed by the 8th edition of RNDM workshop co-organized on Sept. 13-15, 2016 at Halmstad University, SE by Prof. Magnus Jonsson, SE, Prof. Jacek Rak, PL and Prof. Arun Somani (Iowa State University, US). RNDM 2016 followed the success of its seven former editions gathering over 40 leading researchers from both academia and industry.

After a detailed review process (each paper submitted to RNDM 2016 received at least four reviews), accepted papers were organized into the following technical sessions:

•Theory of network resilience
•Optical networks survivability
•Models and algorithms of resilient networks design
•Disaster resilience
•Methods of evaluation and improvement of network resilience
•Resilience of virtualized and cloud architectures
•Special session of COST CA15127-RECODIS

Apart from regular presentations, RNDM 2016 offered three invited talks, a panel discussion on disaster resilience–research challenges and perspectives, and two keynote talks, by Prof. Biswanath Mukherjee from the University of California, Davis, US entitled “Disaster resilience of telecom infrastructure,” and by Prof. Yan Zhang from University of Oslo, NO on “Reliable smart energy networks.”

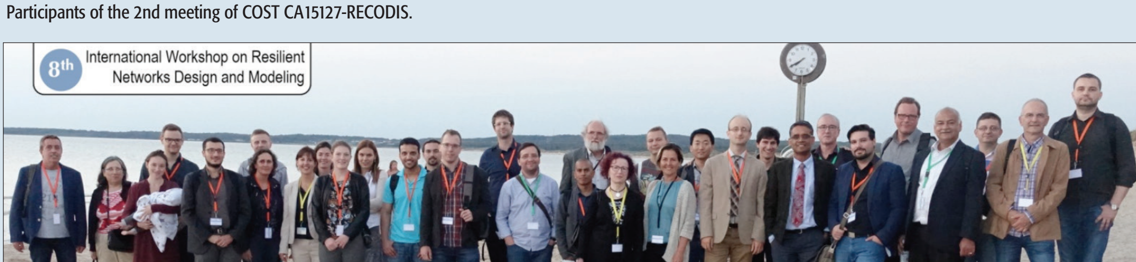
The Best Paper Award was given to presenters of two papers: Sahel Sahhaf from Ghent University-iMinds, BE for the paper entitled “Resilient availability and bandwidth-aware multipath provisioning for media transfer over the Internet,” and Francesco Musumeci from Politecnico di Milano, IT (related to the paper on “Virtual network function placement for resilient service chain provisioning”).

A special issue of *Networks*, published by Wiley, is being prepared for extended versions of selected RNDM 2016 papers.

RNDM 2016 was organized in conjunction with the 4th International Workshop on Understanding the Inter-play between Sustainability, Resilience and Rubustness in Networks (USRR 2016), and the 11th International Workshop on Communication Technologies for Vehicles (Nets4Cars 2016 Fall).

The next edition of RNDM will be in Alghero, Sardinia, Italy on September 4-6, 2017. More information can be found at http://www.rndm.pl/2017/

Participants of the 2nd meeting of COST CA15127-RECODIS.



Participants of the RNDM 2016 Workshop.

REGIONAL REPORT

Inauguration of the Singapore International Cyber Week: GovernmentWare (GovWare) 2016

By Leo Hwa Chiang, Director of IEEE Asia Business Development, Singapore

The foundation event for the inaugural Singapore International Cyber Week, GovernmentWare (GovWare), was held at the Suntec Singapore Convention & Exhibition Centre, 10–12 October 2016. The Prime Minister of Singapore, Lee Hsien Loong, was the Distinguished Guest-of-Honor. He gave an opening speech at GovWare and highlighted the theme for this year, "Building a secure and resilient digital future through partnership," reflecting Singapore's desire to strengthen the digital future through building robust local and international partnerships. This conference was organized by the Cyber Security Agency of Singapore (CSA), with a speaker faculty of over 100 Government officials.

In collaboration with the IEEE Communications Society, the Director of IEEE Asia Business Development, Mr. Leo Hwa Chiang, organized two workshops at this premier conference. The first workshop, Cyber Forensics, was held on 11 October 2016. There were 15 attendees at this workshop, with an objective to inspire and boost interest in cyber security. Participants in this workshop had the opportunity to learn the fundamental principles of cyber forensics and explore the specialized software tools, techniques and processes employed to preserve and recover digital evidence.


Workshop organized by Leo Leo Hwa Chiang.

On 12 October 2016, Leo organized a Cyber Risk Management Workshop. This workshop attracted 16 attendees. Participants were enthusiastic in joining the discussion to share their organizations' experience with cyber risk management, and also discuss the challenges they have encountered in improving and maintaining their cyber risk management capability.

WICE/*Continued from page 1*

Bregni: What is the most significant event organized by WICE in the last year?

Dobre: In December 2016, we organized the Third Women's Workshop on Communications and Signal Processing in conjunction with GLOBECOM 2016 in Washington, DC, USA. It was a fabulous event, with well known speakers from industry and academia, such as Prof. Andrea Goldsmith (Stanford University), Monique Morrow (CISCO), Dr. Antonia Tulino (Nokia Bell Labs), Dr. Peying Zhu (Huawei), and Dr. Grace Wang (National Science Foundation).

In addition to technical talks, there were discussions focusing on the role of women in the IEEE and IEEE ComSoc, as well as panel discussions to address questions regarding career development in both industry and academia. The panel was led by Prof. Katie Wilson (Santa Clara U.), and the invited panelists were Dr. Sheila Hemami (Draper Labs and IEEE VP Publications), Prof. Muriel Medard (MIT), Prof. Elza Erkip (NYU), and Dr. Antonia Tulino. The workshop attracted 54 attendees, from Australia, Canada, China, France, Germany, Ireland, Italy, India, New Zealand, Spain, Switzerland, Sweden, Turkey, UK, and US.

Bregni: What other recent initiatives run by WICE can you briefly highlight?

Dobre: WICE presents three awards annually. In 2016, they were presented at the WICE workshop to Prof. Elza Erkip (Outstanding Achievement Award), Prof. Shalinee Kishore (Outstanding Service Award), and Prof. Urbashi Mitra (Mentorship Award).

In 2017, starting with WCNC, and continuing with ICC and GLOBECOM, we are organizing various panels led by WICE members. An example of such a panel is "How to encourage young women in Communications Engineering – best practices and experiences," which was led by Profs. Ana Garcia Armada and Meryem Simsek at WCNC, and supported by Intel.

ComSoc student members also participate in the organization of these events, being provided with travel grants. The WICE organizers serve as mentors for the students as well.

In 2016, WICE also presented the first Child Care Grants to the IEEE GLOBECOM 2016 attendees, who were IEEE ComSoc members and brought small children to the conference or incurred extra expenses in leaving their children at home. We hope this program will continue at the IEEE ICC and GLOBECOM conferences, and extend to others, such as IEEE WCNC.

Bregni: Oh, be sure that the Child Care Grant program is among my priorities as VP-MGA. As you remember, we had to push strongly and spend a considerable effort to make it real, solving non-trivial practical and legal issues. I will do my best to make it a permanent program in ComSoc MGA also in the future.

This is the fourth and therefore last year of your two terms as WICE Chair. Is there any special remark you wish to make?

Dobre: As my two terms as WICE chair come to an end in 2017, I would like to take this opportunity to express my gratitude and appreciation to the entire WICE Board, the organizers of the WICE Workshop, as well as to other colleagues and friends whom I worked with over the past four years. Without their contributions and support, WICE would not be the amazing committee that is today.

I would also like to thank you, Stefano, the ComSoc President, as well as the ComSoc staff for your steady support. My wish is that WICE will continue to thrive, and thus, contribute to the growth of our Society.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

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

IEEE ComSoc IEEE Communications Society
www.comsoc.org/gcn
ISSN 2374-1082

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

2017

MAY

IEEE INFOCOM 2017 — IEEE Int'l. Conference on Computer Communications, 1–4 May
Atlanta, GA
http://infocom2017.ieee-infocom.org/

ICT 2017 — Int'l. Conference on Telecommunications, 3–5 May
Limassol, Cyprus
http://ict-2017.org/

IFIP/IEEE IM 2017 — IFIP/IEEE Int'l. Symposium on Integrated Network Management, 8–12 May
Lisbon, Portugal
http://im2017.ieee-im.org/

IEEE EIT 2017 — IEEE Int'l. Conference on Electro Information Technology, 14–17 May
Lincoln, NE
http://engineering.unl.edu/eit2017/

ISNCC 2017 — Int'l. Symposium on Networks, Computers and Communications, 17–19 May
Marrakesh, Morocco
http://www.isncc-conf.org/

IEEE ICC 2017 — 2017 IEEE Int'l. Conference on Communications, 21–25 May
Paris, France
http://icc2017.ieee-icc.org/

JUNE

IEEE BlackSeaCom 2017 — IEEE Int'l. Black Sea Conference on Communications and Networking, 5–9 June
Istanbul, Turkey
http://blackseacom2017.ieee-blackseacom.org/

GIoTS 2017 — Global Internet of Things Summit, 6–9 June
Geneva, Switzerland
http://iot.committees.comsoc.org/global-iot-summit-2017/

IEEE CTW 2017 — IEEE Communciation Theory Workshop, 11–14 June
Natatola Bay, Fiji
http://ctw2017.ieee-ctw.org/

IEEE LANMAN 2017 — IEEE Workshop on Local & Metropolitan Area Networks, 12–15 June
Osaka, Japan
http://lanman2017.ieee-lanman.org/

IEEE SECON 2017 — IEEE Int'l. Conference on Sensing, Communication and Networking, 12–14 June
San Diego, CA
http://secon2017.ieee-secon.org/

EuCNC 2017 — European Conference on Networks and Communications, 12–15 June
Oulu, Finland
http://eucnc.eu/?q=node/156

IEEE/ACM IWQOS 2017 — IEEE/ACM Int'l. Symposium on Quality of Service, 14–16 June
Vilanova i la Geltrú, Spain
http://iwqos2017.ieee-iwqos.org/

IEEE CAMAD 2017 — IEEE Int'l. Workshop on Computer Aided Modeling and Design of Communication Links and Networks, 19–21 June
Lund, Sweden
http://weber.itn.liu.se/~vanan11/CAMAD17/

TMA 2017 — Network Traffic Measurement and Analysis Conference, 21–23 June
Dublin, Ireland
http://tma.ifip.org/

NETGAMES 2017 — Annual Workshop on Network and Systems Support for Games, 22–23 June
Taipei, Taiwan
http://netgames2017.web.nitech.ac.jp/

CLEEN 2017 — Int'l. Workshop on Cloud Technologies and Energy Efficiency in Mobile Communication Networks, 22 June
Turin, Italy
http://www.flex5gware.eu/cleen2017

IEEE HPSR 2017 — IEEE Int'l. Conference on High Performance Switching and Routing, 27–30 June
Campinas, Brazil
http://hpsr2017.ieee-hpsr.org/

JULY

IEEE ISCC 2017 — IEEE Symposium on Computers and Communications, 3–6 July
Heraklion, Greece
http://www.ics.forth.gr/iscc2017/index.html

IEEE NETSOFT 2017 — IEEE Conference on Network Softwarization, 3–7 July
Bologna, Italy
http://sites.ieee.org/netsoft/

ICUFN 2017 — Int'l. Conference on Ubiquitous and Future Networks, 4–7 July
Milan, Italy
http://icufn.org/

IEEE ICME 2017 — IEEE Int'l. Conference on Multimedia and Expo, 10–14 July
Hong Kong, China
http://www.icme2017.org/

SPLITECH 2017 — Int'l. Multidisciplinary Conference on Computer and Energy Science, 12–14 July
Split, Croatia
http://splitech2017.fesb.unist.hr/

CITS 2017 — Int'l. Conference on Computer, Information and Telecommunication Systems, 21–23 July
Dalian, China
http://atc.udg.edu/CITS2017/

ICCCN 2017 — Int'l. Conference on Computer Communication and Networks, 31 July–3 Aug.
Vancouver, Canada
http://icccn.org/icccn17/

AUGUST

ISWCS 2017 — Int'l. Symposium on Wireless Communication Systems, 28–31 Aug.
Bologna, Italy
http://iswcs2017.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.

Enabling Mobile and Wireless Technologies for Smart Cities: Part 3


Ejaz Ahmed


Muhammad Imran


Mohsen Guizani


Ammar Rayes


Jaime Lloret


Guangjie Han


Wael Guibene

Due to advancements in communication and computing technologies, smart cities have become the main innovation agenda of research organizations, technology vendors, and governments. To make a city smart, a strong communications infrastructure is required for connecting smart objects, people, and sensors. Smart cities rely on wireless and mobile technologies for providing services such as healthcare assistance, security, and safety, real-time traffic monitoring, and managing the environment, to name a few. Such applications have been a main driving force in the development of smart cities. Without the appropriate communication networks, it is really difficult for a city to facilitate its citizens in a sustainable, efficient, and safer manner/environment. Considering the significance of mobile and wireless technologies for realizing the vision of smart cities, there is a need to conduct research to further investigate the standardization efforts and explore different issues/challenges in wireless technologies, mobile computing, and smart environments.

In this *IEEE Communications Magazine* Feature Topic (FT), we invited researchers from academia, industry, and government to discuss challenging ideas, novel research contributions, demonstration results, and standardization efforts on enabling mobile and wireless technologies for smart cities. After a rigorous review process, 18 papers were selected to be published in this FT of *IEEE Communications Magazine*. Five of the 18 are published here in Part 3 of the FT.

To investigate and verify the data validity in an urban transportation system, the authors of "Exploring Data Validity in Transportation Systems for Smart Cities" used the records of vehicle location and transaction logs of smart cards. The authors proposed a mechanism to find time discrepancies by investigating the signal patterns of origin inference success rate. An investigation is made by using the results of data mining to find that a small amount of missing records causes severe bias on passenger flow data mining.

A broad impact of various applications on operational flow of market is studied in "Smarter Markets for Smarter Life: Applications, Challenges, and Deployment Experiences." The authors also identify challenges of embedded wireless systems that should be addressed when targeting market deployments. Further, the impact of daily market activities on wireless networks is also investigated. Based on the findings, a prototype IoT system has been designed to enable the realization of a smart market environment.

WiFi networks deployment and utilization in smart cities are key concerns. The authors of "WiFi Networks in Metropolises: From Access Point and User Perspectives" try to answer the questions related to WiFi deployment in smart cities. Based on the study, four key observations are made: a) WiFi networks have dense deployment in urban areas; b) public and business WiFi networks are more concentrated; 3) the access pattern for a majority of the WiFi users is regular with either ultra short connections or quite long connections; and 4) WiFi users form a well structured network.

Transmission reliability is crucial in vehicular social networks for realizing the vision of smart cities. The authors in "Vehicular Social Networks: Enabling Smart Mobility" emphasize the significance of reliable transmission in vehicular social networks of smart cities. The case of traffic anomaly detection for vehicular social networks is studied by trajectory data analysis.

Energy efficiency is a key concern in smart cities. The authors in "Toward Eco-Friendly Smart Mobile Devices for Smart Cities" propose a solution that is based on the distributed shared time access algorithm. The algorithm ameliorates the energy efficiency of the mobile users in terms of battery life. It forms eco-friendly mobile networks for the future smart cities. The simulation results validate the solution and show the supremacy of the proposed solution compared to other access mechanisms.

In a smart connected city, the backhaul wireless network is a key infrastructure to forward aggregated data from various sources such as sensors and machines. The authors in "Connecting a City by Wireless Backhaul: 3D Spatial Channel Characterization and Modeling Perspectives" have introduced the concepts and methodology of spatial channel modeling and reviewed the recent state-of-the-art models developed by the standardization bodies. Moreover, characteristics of the urban wireless backhaul channel are also revealed due to their merits in the design and deployment of wireless backhaul for a smart city.

The Guest Editors would like to thank all the involved people, including the contributing authors for their high-quality submissions, the anonymous reviewers for their timely and insightful comments, and the *IEEE Communications Magazine* staff for their continuous support. We believe that the presented contributions in this FT will captivate and spark novel research directions for mobile and wireless technologies for smart cities.

BIOGRAPHIES

EJAZ AHMED has worked as a researcher at C4MCCR, University of Malaya, Malaysia, CogNet Lab, NUST, and CoReNet, Maju, Pakistan. He is an Associate Technical Editor of *IEEE Communications Magazine*, *IEEE Access*, *Springer MJCS*, and *Elsevier JNCA*. He has also served as a Lead Guest Editor for the *Elsevier FGCS Journal*, *IEEE Access*, *Elsevier Computers & Electrical Engineering*, *IEEE Communications Magazine*, *Elsevier Information Systems*, and *Transactions on Emerging Telecommunications Technologies*.

MUHAMMAD IMRAN is currently working at King Saud University and is a visiting scientist at Iowa State University. His research interests include MANETs, WSNs, WBANs, M2M/IoT, SDN, and security and privacy. He has published a number of research papers in refereed international conferences and journals. He serves as a Co-Editor-in-Chief for *EAI Transactions* and Associate/Guest Editor for *IEEE Access*, *IEEE Communications Magazine*, *Computer Networks*, *Sensors*, *IJDSN*, *JIT*, *WCMC*, *AHSWN*, *IET WSS*, *IJAACS*, and *IJITEE*.

MOHSEN GUIZANI [S'85, M'89, SM'99, F'09] received his B.S., M.S., and Ph.D. from Syracuse University. He is currently a professor and the ECE Department Chair at the University of Idaho. His research interests include wireless communications/mobile cloud computing, computer networks, security, and smart grid. He is the author of nine books and 400+ publications He was the Chair of the IEEE Communications Society Wireless Technical Committee. He served as an IEEE Computer Society Distinguished Speaker.

AMMAR RAYES [S'85, M'91, SM'15] is a Distinguished Engineer focusing on the technology strategy for Cisco Services. His research interests include IoT, network management NMS/OSS, machine learning, analytics, and security. He has authored three books, over 100 publications in refereed journals and conferences on advances in software and networking related technologies, and over 25 patents. He received B.S. and M.S. degrees from the University of Illinois at Urbana and his D.Sc. degree from Washington University, all in electrical engineering.

JAIME LLORET [M'07, SM'10] received his M.Sc. in physics in 1997, his M.Sc. in electronic engineering in 2003, and his Ph.D. in telecommunication engineering in 2006. He is the head of the Communications and Networks research group of the Research Institute IGIC. He is Editor-in-Chief of *Ad Hoc and Sensor Wireless Networks* and *Network Protocols and Algorithms*. He has been General Chair of 36 international workshops and conferences. He is an IARIA Fellow.

GUANGJIE HAN [S'01, M'05] is currently a professor with the Department of Information and Communication Systems, Hohai University, China. His current research interests include sensor networks, computer communications, mobile cloud computing, and multimedia communication and security. He has served on the Editorial Boards of 14 international journals, including *IEEE Access* and *Telecommunication Systems*. He has guest edited a number of Special Issues in IEEE journals and magazines. He is a member of ACM.

WAEL GUIBENE has been a research scientist at Intel Labs since June 2015. He was awarded his Ph.D. from Telecom ParisTech in July 2013. He also holds an M.Eng. and a Master's degree in telecommunications obtained in 2009 and 2010, respectively. He worked at Eurecom as a research engineer from 2010 to November 2013, and then joined Semtech to work on LoRa systems from 2013 to June 2015. His research activities include IoT, 5G, and wireless communications.

Exploring Data Validity in Transportation Systems for Smart Cities

Yongxin Liu, Xiaoxiong Weng, Jiafu Wan, Xuejun Yue, Houbing Song, and Athanasios V. Vasilakos

In transportation systems, data-driven management can dramatically enhance the operating efficiency by providing a clear and insightful image of passengers' transportation behavior. The authors focus on the data validity problem in a cellular network based transportation data collection system from two aspects: internal time discrepancy and data loss.

Abstract

Efficient urban transportation systems are widely accepted as essential infrastructure for smart cities, and they can highly increase a city's vitality and convenience for residents. The three core pillars of smart cities can be considered to be data mining technology, IoT, and mobile wireless networks. Enormous data from IoT is stimulating our cities to become smarter than ever before. In transportation systems, data-driven management can dramatically enhance the operating efficiency by providing a clear and insightful image of passengers' transportation behavior. In this article, we focus on the data validity problem in a cellular network based transportation data collection system from two aspects: internal time discrepancy and data loss. First, the essence of time discrepancy was analyzed for both automated fare collection (AFC) and automated vehicular location (AVL) systems, and it was found that time discrepancies can be identified and rectified by analyzing passenger origin inference success rate using different time shift values and evolutionary algorithms. Second, the algorithmic framework to handle location data loss and time discrepancy was provided. Third, the spatial distribution characteristics of location data loss events were analyzed, and we discovered that they have a strong and positive relationship with both high passenger volume and shadowing effects in urbanized areas, which can cause severe biases on passenger traffic analysis. Our research has proposed some data-driven methodologies to increase data validity and provided some insights into the influence of IoT level data loss on public transportation systems for smart cities.

Introduction

The goal of a smart city is to provide inhabitants with a promising quality of life by using technology to improve the efficiency of services and meet inhabitants' demands [1]. A key symbol of a smart city is an efficient public transportation system with an intelligent management strategy. This is widely accepted as a powerful tool for environment-friendly urbanization and an essential part of citizens' daily lives [2].

To plan and schedule an efficient urban transportation system, there is a high reliance on the Internet of Things (IoT), mobile wireless networks, and data mining technologies [3, 4]. IoT uses real-time systems and sensors to collect data from citizens and objects, while big data mining technologies use data analytics and fusion techniques to identify the travel patterns of residents and provide various levels of insightful images on citizens' ridership. Both IoT and data mining can be used by transportation operators to enable intelligent decisions to be made, while traditional methods of achieving this would require huge volumes of manual surveys. Due to the high cost involved, these studies are typically based on data from a tiny sample of residents.

Currently, with advancements in communication and the ubiquity of wireless connectivity, transportation operators can collect transactional records and vehicular locations using cellular networks, and this data may also potentially incorporate the travel pattern of passengers. Recently, researchers have been focusing on using low-cost data mining techniques to derive bus riders' commuting patterns from transportation operators' automated fare collection (AFC) and automated vehicular location (AVL) systems. For example, transaction records can be correlated with vehicle location records using timestamps to identify an individual's boarding or alighting point. However, although a huge amount of AFC and AVL records can be collected for analysis in real time, the accuracy of data mining conclusions still relies on the quality of data [5]. Two of the most common errors that reduce data quality are location record loss and time discrepancy. Time synchronization discrepancy occurs when internal clock sources within AVL and AFC devices are not consistent, while location record loss happens when vehicles' location information are transmitted in cellular networks with low quality of service (QoS).

In this article, we use smart card transaction records and vehicle location records along with manual investigation results to explore and verify data validity in an urban transportation system. We evaluate the impact of location data loss and time discrepancies in extracting passengers' origin information. We propose a method to identify time discrepancies by analyzing the output signal patterns of passenger origin inference success rate using the continuous time shift values as input. We discovered that the location data loss events have a strong and positive relationship with shadow effects in urbanized areas where

Digital Object Identifier:
10.1109/MCOM.2017.1600240

Yongxin Liu, Xiaoxiong Weng, and Jiafu Wan (corresponding author) are with South China University of Technology; Xuejun Yue is with South China Agricultural University; Houbing Song is with West Virginia University; Athanasios V. Vasilakos is with Luleå University of Technology.

high passenger volume exists. And we prove by using our data mining results that even a small amount of missing records can cause severe bias on passenger flow data mining.

The remainder of the article is organized as follows. After an overview of related works, we introduce our dataset and methodology for data analysis and time discrepancy identification. We show our algorithm performance along with the spatial distribution of boarding and data loss events. Finally, we summarize our work with a short discussion of our future plans.

Related Work

In this section, we review relevant works on transportation data mining for intelligent decision making in smart cities and also discuss recent efforts to increase data validity.

For transportation smart card data mining, Ma *et al.* [6] proposed a data-driven platform for online transportation performance monitoring. The origin and destination of individual transportation riders can be estimated by utilizing a series of data mining techniques, which are then incorporated into a geographic information system (GIS)-based platform to calculate transportation performance measurements. Kieu *et al.* [7] provided a methodology for passenger segmentation using smart card data to understand passengers' behavior to provide passengers relevant information and services. Wang *et al.* [8] used smart card transaction records of metro stations to analyze metro trip patterns at an aggregate level. Sui *et al.*, in [9], applied a geo-visualization-based method to a large smart card database to understand spatial-temporal dynamics of a bus rapid transportation (BRT) system and ridership variation according to calendar events.

Efforts have been made to increase data validity for various applications. Marcela *et al.* [10] provided a series of methods to enhance the robustness of transportation riders' origin-destination estimation. Zhang *et al.* [11] provided a probabilistic-based methodology to increase the accuracy of boarding and alighting inference, and their methodology was verified by estimating the residents' home and work locations. Ma *et al.* [12] provided a Bayesian-based method to derive bus riders' boarding points from smart card transaction records and travel time, and their method can reduce the requirement for vehicular location records.

In summary, smart card data mining can provide insights that can be used for mass transportation planning and to provide understanding of transportation behavior in urban areas, although data mining and other processes are required to decode useful information from raw datasets. However, two problems that reduce data validity still exist in practical applications and have not been thoroughly studied: internal time discrepancy and location data loss. These problems are the motivation for our research.

Methodology

In this section, we first introduce our dataset used this research, and then briefly provide an overview of the architecture of an urban transportation data collection system in China. We then describe the algorithm for time discrepancy error identification together with the location loss evaluation algorithm.

Datasets

We collected both AFC and AVL data, which included 1506 transportation buses and 614,397 smart card transaction records, for the target city (with a population of 1.6 million, 162 public bus routes) on June 2, 2015 from the local transportation agency. Each AFC transaction record contains: Card ID, Transaction Time Stamp, Route ID, and Bus ID. The AFC system is designed for billing only, and therefore it does not store geographical location information. For AVL records, the corresponding fields are: Bus ID, Route ID, Station ID, and Time Stamp. On the same day, manual surveys were conducted on 10 randomly chosen buses. In these manual surveys, the exact moments when passengers boarded and left the bus were recorded. Additionally, we collected the public bus route table, geographical points of interest (POIs), and bus stop coordinates of the city using application programming interfaces (APIs) from Autonavi.com (http://lbs.amap.com/api/javascript-api/summary/).

Architecture of the Urban Transportation Data Collection System in China

A simplified architectural overview of a typical urban transportation data collection system in China is provided in Fig. 1. In this figure, we can separate the data collection events into two different scenarios: buses in the range of a bus station, depicted by "station area" in Fig. 1 and buses out of range of a station, depicted by "road area" in Fig. 1. In the road area scenario, the onboard AVL system periodically transmits the bus's location records (containing a timestamp and coordinates from a positioning system such as GPS) into the public cloud servers via a cellular network. In our target city, the location update period is 30 s. In the station area scenario, the onboard AVL system keeps working while the AFC system handles the smart card ticketing events. In this scenario, transaction records (containing billing information and individual timestamps) are transmitted back to the transportation companies' private cloud servers and stored in non-volatile storage devices for immediate verification. Since the transaction records are verified, we assume that data loss does not exist in AFC transactional records. However, it may be possible to use the continuous AVL trajectories to mine the event using the information that buses enter and allow boarding in the station area, which is known as stop event detection.

Apart from the system architecture, there are two drawbacks that should be highlighted within the data collection system. First, in Fig. 1, it can be noticed that there are two different clock sources employed by AVL and AFC equipment, depicted as the AVL time system in blue and the AFC time system in green. In the target city, these two systems are independent, and the AFC time source requires manual calibration. Therefore, time discrepancies may occur due to the accumulation of daily errors. Second, data is transmitted using the cellular network, which may not be a good choice for data transmission with a high QoS demand [13], since AVL location records may encounter

Smart card data mining can provide insights that can be used for mass transportation planning and to provide understanding of transportation behavior in urban areas. However, two problems that reduce data validity still exist in practical applications and have not been thoroughly studied: internal time discrepancy and location data loss.

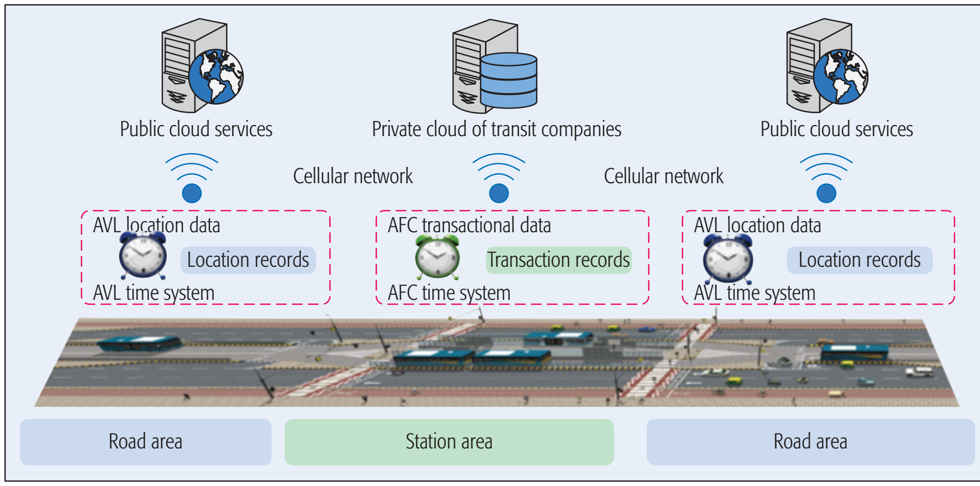


Figure 1. Data flow of an urban transportation data collection system in China.

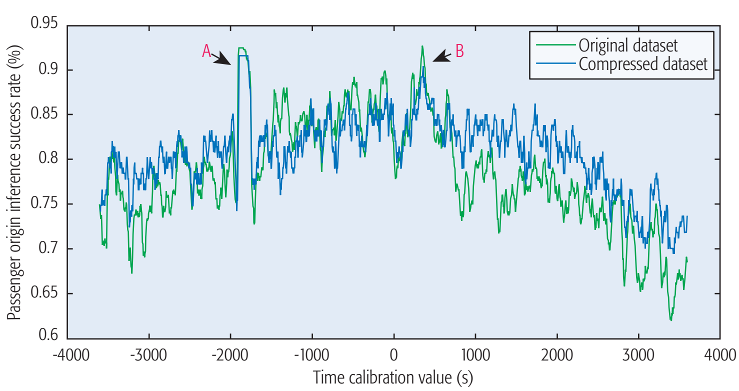


Figure 2. The correlation success rate with different time shift values.

packet loss during transmission. Conventional passenger origin inference algorithms correlate transaction timestamps of smart card records with the timestamps of vehicular location records in order to assign a geographical tag to each transaction. Therefore, in summary, incorrect timestamps lead to incorrect judgments on the boarding events of passengers, while location loss results in the failure of the boarding inference.

Time Discrepancy Identification

In this subsection, the effect of time discrepancies on passenger origin inference success rate is first evaluated, and then a theoretical analysis for our algorithm is provided. Finally, the procedure of our method is summarized.

Time Discrepancy Affect: We evaluate the influence of time discrepancies on the passenger origin inference success rate. A prior data processing experiment using AFC and AVL records from several buses has been conducted, which contains two steps:

- Smart card transactional records from the original dataset are clustered with a maximum interval of 30 s; then the transaction record with the earliest timestamp is extracted to form a new dataset, which is defined as a compressed dataset.
- Passenger origin inference of both the original dataset and the compressed dataset are executed using different time calibration values () added to the AFC data tuples for multiple times. The calibration values range from 3600 s to +3600 s with an interval of 1 s. The success rate (r_{match}) was observed as shown in Fig. 2.

A typical result obtained using both the original and the clustered AFC datasets is provided in Fig. 2. The actual time gap of the AFC and AVL system on this bus is –1800 s according to manual investigation.

In Fig. 2, the first observation is that the peak value of the passenger origin inference success rate occurs at points A and B. The time calibration value at point A corresponds to the actual time discrepancy, while point B can be regarded as an interference point that might lead to bias identification of the time synchronization discrepancy. However, from the point of view of signal processing, the waveform patterns of points A and B are different, and the waveform pattern at point A is denoted after this as Signal-A. The second observation is that the passenger origin inference results on both the original and the compressed datasets demonstrate that the clustering compression algorithm does not have a significant impact on the time discrepancy identification.

Intuitively, this prior test reveals that by using AFC and AVL datasets from each bus, it can be attempted to discover the time discrepancy using different adjustment values and observing the output passenger origin inference success rate.

Time Discrepancy Identification Theoretical Analysis: To provide a theoretical foundation and illustrate this algorithm, several assumptions should be made:

1. The time interval between any two consecutive bus stops ($t_{station}$) should be longer than the two time margins for the passenger origin inference (t_{margin}), for example, the time margin could be set at 60 s in most cases, and in China, the time interval between two consecutive bus stops is usually longer than three minutes [6].
2. When there is no significant loss of AVL records and the time discrepancy ($|t_{err}|$) is

less than the passenger origin inference time margin (t_{margin}), all ticketing records can be correlated, and the passenger origin inference rate (r_{match}) therefore reaches its maximum (R_{max}).
3. The time discrepancy between the AVL and AFC systems on each bus does not vary between each individual day.

These assumptions are easily satisfied in actual urban transportation systems. Based on these assumptions, we can draw the following statements:
- Based on assumption 2, we can summarize the pattern of Signal-A: The passenger origin inference success rate reaches its maximum; the duration of this peak value is longer than two time margins.
- According to the Nyquist sampling theorem, Signal-A can be sampled if we have a sampling interval (t_{sample}) shorter than the time margin.

Based on these assumptions and statements, we can further summarize the procedure of searching for an appropriate t_{cal} as a mathematical optimization process where we want to maximize the passenger origin inference success rate while satisfying the pattern of Signal-A. To avoid a brute force search, two evolutionary algorithms (the genetic algorithm and the differential evolution algorithm) [14] were employed and discussed. In order to avoid being trapped (also known as convergence at a local optima) at interference points with a high correlation rate (e.g., point B in Fig. 2), our algorithm judges that if its virtual population is trapped at a local optimum, this local optimum should be eliminated by multiplying its corresponding correlation success rate by a scaling factor ($\lambda < 1$), which will renew the population with the unvisited time shift values.

The second statement can be helpful for compressing the search space, for example, if we know that the time margin is 60 s, an original search space from –3600 s to 3600 s that contains 7200 points (where each point represents 1 s) can be resampled and compressed into a subset from –60 min to 60 min with 120 points (where each point represents 1 min).

Evolutionary Time Discrepancy Identification Algorithm: The subsection above can be summarized to obtain a general procedure for the time discrepancy identification algorithm:
1. Generate a compressed AFC dataset by clustering smart card transactional records with a maximum time interval of 30 s.
2. Choose six (or more) initial time shift values to be the first generation of the population.
3. Perform passenger origin inference and judge the signal pattern, return a time shift value if Signal-A is found.
4. Use the genetic algorithm or the differential evolution mechanism to generate a new population.
5. Go to 3 or return a failure if the maximum iteration count is reached.

The essence of this time discrepancy identification algorithm can be summarized by three aspects:
- Find the corresponding time shift value at Signal-A iteratively or to solve an optimization problem iteratively.

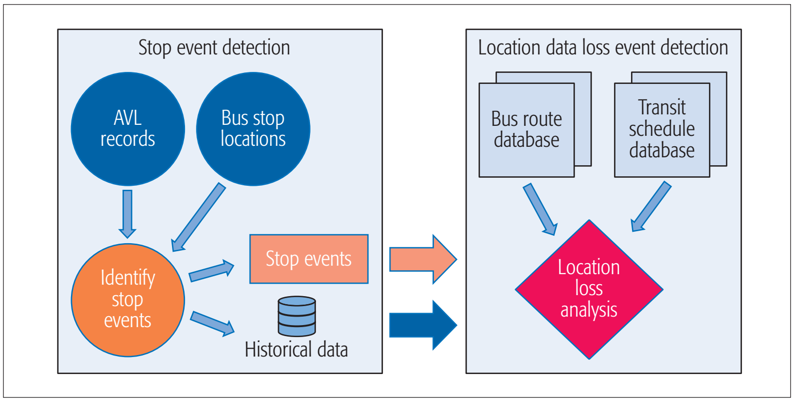


Figure 3. The location data loss analysis procedure.

- To reduce the computational density, data compression techniques, and search space resample techniques are employed.
- Evolutionary algorithms are employed to provide a faster and more directional search strategy.

Location Data Loss Analysis

In this subsection, we use our method to mine location data loss events that result in failures of stop event detection regardless of time discrepancies. The block diagram for our procedure is provided in Fig. 3.

Our process (Fig. 3) can be separated into two phases: bus stop event detection and location data loss event detection. In the stop event detection, AVL trajectories for buses are analyzed by measuring the speed and location of each vehicle alongside the bus stop coordinates to judge whether a bus has stopped within a station area. In location data loss event detection, our algorithm refers to the city's bus route table, the transportation operators' vehicle schedule table, and the historical stop events of each bus in order to determine missing bus stops that result from location data loss events. Additionally, for each missing bus stop, the vehicle's historical bus stop events are used to estimate the timestamp for the missing bus stop, using similar methods to [12].

Finally, given the AFC and AVL datasets, the passenger origin inference algorithm including time discrepancy identification and location data loss analysis contains the following steps:
- Perform location loss analysis and complement the AVL dataset.
- Identify and rectify time discrepancies between each vehicle's AVL and AFC records.
- Perform passenger origin inference on the rectified AFC and complemented AVL dataset.

Algorithm Evaluation and Analysis

In this section, the performance of the time discrepancy identification using both a genetic algorithm and a differential evolution algorithm is first evaluated by computational complexities, average successful identification probabilities, and convergence probabilities. The spatial distribution of location data loss with spatial correlation analysis are then provided, which leads to our final conclusion.

The average convergence probability denotes the probability that the evolutionary algorithm will be trapped by inference points. In this research, the average convergence probability is defined as a count of the visited time shift points divided by the summation of the convergence points.

Performance Evaluation of Time Discrepancy Identification

In this subsection, we focus on the performance of time discrepancy identification based on the genetic algorithm and the differential evolution algorithm. For the genetic algorithm, we define the range of the mutation probability and the cross-over probability from 0.1 to 0.9 with a step of 0.1. The same range is applied to the cross-over probability and scaling factor in the differential evolution algorithm. For both algorithms, the initial population was set to six with a maximum iteration of 1500; the same dataset that was used to generate Fig. 2 was utilized for the performance evaluation. To attain reliable results, 1000 tests were performed for each combination of parameters; thus, over 81,000 results were finally obtained, which can be summarized as follows.

Average Computational Complexity and Success Rate: Two indicators, the average time discrepancy identification success rate and the average number of passenger origin inferences (most computational expensive), were selected to test the performance of our algorithm. This comparison was provided in Fig. 4.

As can be seen in Figs. 4c and 4d, the differential evolution algorithm requires fewer passenger origin inferences than the genetic algorithm in most cases but with a lower success rate (Figs.4a and 4b). Referring to Fig. 4c, we can see that the optimal configuration for the genetic algorithm has a mutation probability of 0.5 and a cross-over probability of 0. 5 (the lowest center point). Referring to Fig. 4a, we can see that at this point, we have a high success rate (100 percent) with an acceptable computational cost (less than 60 times that of passenger origin inference).

Average Convergence Probability: The average convergence probability denotes the probability that the evolutionary algorithm will be trapped by inference points. In this research, the average convergence probability is defined as a count of the visited time shift points divided by the summation of the convergence points. A comparison of average convergence probabilities is provided in Figs. 4e and 4f.

A comparison of the average converge probabilities has demonstrated that the differential evolution algorithm has a higher risk of convergence than the genetic algorithm, which may cause a lower success rate for time discrepancy identification. This phenomenon may result from the fact that the differential evolution algorithm employs a greedy strategy for reproduction.

Further Analysis: A generalized comparison of the performance of the two employed evolutionary algorithms, including all tested parameters, is provided in Table 1. The best case (least boarding correlation) for the genetic algorithm can be parameterized with both a cross-over probability and a mutation probability equal to 0.5, while a cross-over probability equal to 0.6 and a scaling factor equal to 0.9 represent the best case for the differential evolution algorithm.

From Table 1, we can learn that:

- The genetic algorithm can be employed for practical applications.
- Further enhancement could be made to increase the success rate of the differential evolution algorithm.
- The computational cost can be dramatically reduced by introducing the differential evolutionary algorithm, compared to a brute force search (in the worst case, this method may have to perform passenger origin inferences more than 100 times).

The genetic algorithm driven time discrepancy identification algorithm was selected and tested on the whole dataset of AFC and AVL from the target city. The results are quite promising, and with our method, time discrepancies of 1329 (88.2 percent) public buses were detected. The time discrepancies were not identified for vehicles with a data pattern that does not satisfy assumptions 1 and 2 above.

Spatial Distribution Characteristic of Location Data Loss

In this subsection, we use geographical spatial visualization and correlation analysis to provide some insights into the distribution and impacts of location data loss. The spatial heat map of location data loss, passengers' boarding events, and bus route density are shown in Figs. 5a, 5b, and 5c, respectively. The spatial distributions of POIs, cellular base station, and public bus stops are depicted in Fig. 5d, 5e, and 5f, respectively. In the map of the city, Area A is the downtown area of the target city, areas B and C are two satellite towns and area D is the main bridge connecting areas A and C.

According to the cities' POI in Fig. 5d and cellular base stations in Fig. 5e, we can identify the city's urbanized areas by assessing the density of the sparks. Along with Figs. 5a and 5b, it is obvious that there is a higher possibility of location data loss in bustling areas and conjunctions of arterial roads, where the POI and cellular base station density is also high; therefore, the cause of such distribution of location data loss is not a result of a poorly distributed cellular base station.

The Pearson correlation coefficient between location data loss, passengers' boarding events, and the city's POI density was then calculated. The correlation coefficients showed that the spatial distribution of the location record loss events are highly correlated to passengers' boarding events (R = 0.8011), while location loss events have a positive but low correlation with the POI density (R = 0.4950). At the same time, the POI density was discovered to have a low but positive correlation with the residents' boarding event (R = 0.4912). Apparently, in this city, the POI density map does not have a high contribution when analyzing the AVL system's location loss, but the smart card transaction heat map can be utilized instead to predict the location record loss events. A possible explanation for this location record loss distribution is that in this city, the public transportation operator uses the general packet radio service (GPRS) to transmit the AVL records to a remote server, and both GPS malfunctions and GPRS packet loss can result in location data loss. In the downtown area where shadowing effect can be strong, both GPRS packet loss and GPS malfunction are possible, while in suburban areas with sparsely distributed cellular base stations, GPRS packet loss becomes the cause.

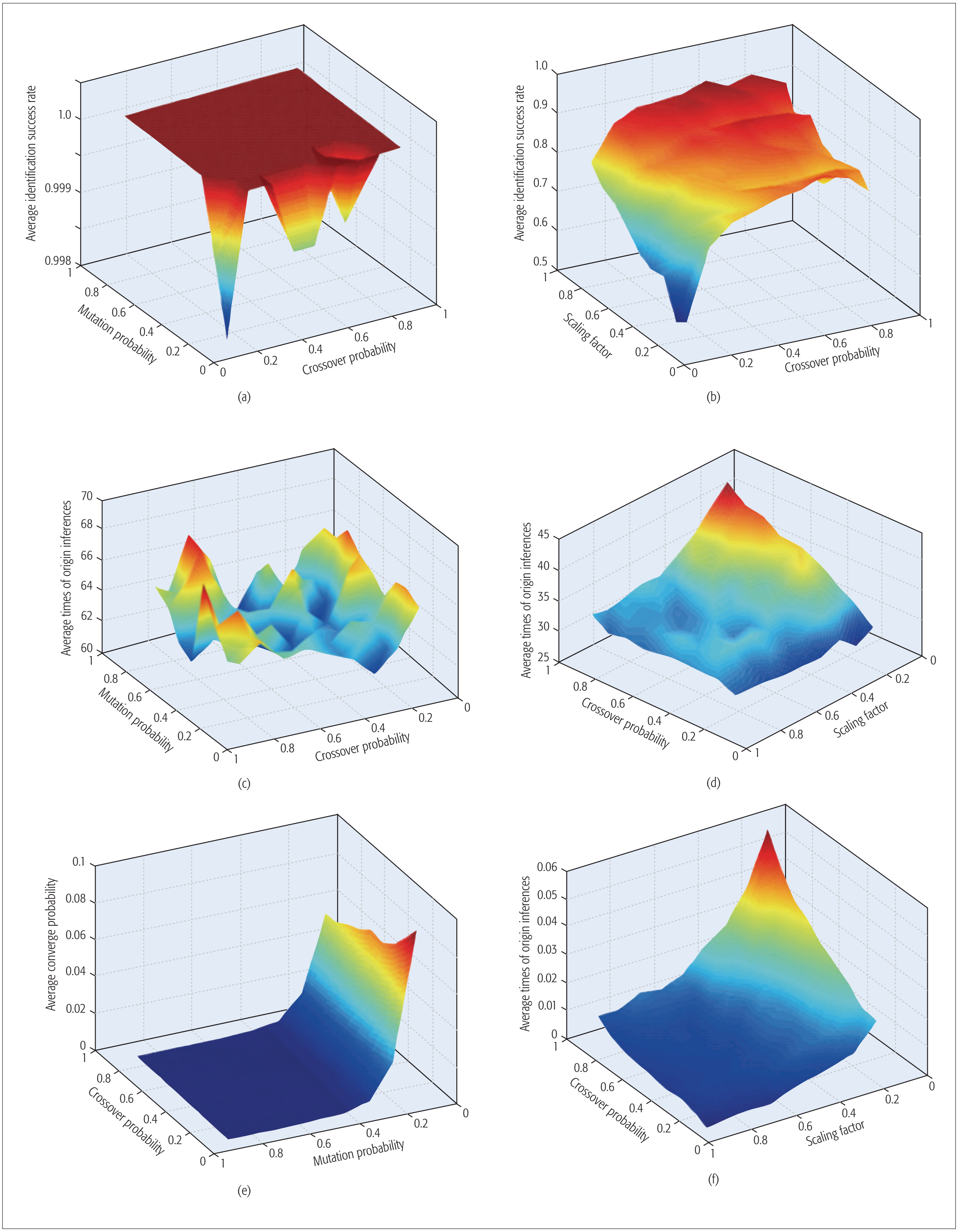


Figure 4. Performance evaluation of the genetic algorithm and the differential evolution algorithm: a) average time discrepancy identification success rate of the genetic algorithm; b) average time discrepancy identification success rate of the differential evolution algorithm; c) average passenger origin inferences of the genetic algorithm; d) average inferences of the differential evolution algorithm; e) average convergence probability of the genetic algorithm; f) average convergence probability of the differential evolution algorithm.

Algorithm	Average (best) success ratio	Average (best) convergence ratio	Average (best) times of passenger origin inferences
GA	99.99 % (100 %)	0.98 % (0.01 %)	64.44 (61.66)
DE	80.76 % (91.2 %)	1.68 % (0.76 %)	34.42 (32.36)

Table 1. Generalized comparison of genetic algorithm and differential algorithm.

It should be noted that in this research, the amount of missing location data was not significant (less than 30,000 per day) compared to the whole city's AVL dataset (more than 600,000 records per day), the missing location records resulted in more than 110,000 (19.67 percent) failures of passenger origin inference in AFC records. Therefore, we believe that the occurrences of location data loss in bustling areas have severe side effects on passenger traffic pattern mining. On the other hand, IoT devices and mobile communication subsystems in urban public transportation systems need further enhancements to support the ongoing demands of intelligent public transportation systems in smart cities [15].

Conclusions

In this study, we have leveraged the huge amount of smart card transactional records and vehicle location records along with manual investigation results to explore and verify data validity in an urban transportation system. In this context, we have provided an algorithmic framework to increase the data validity

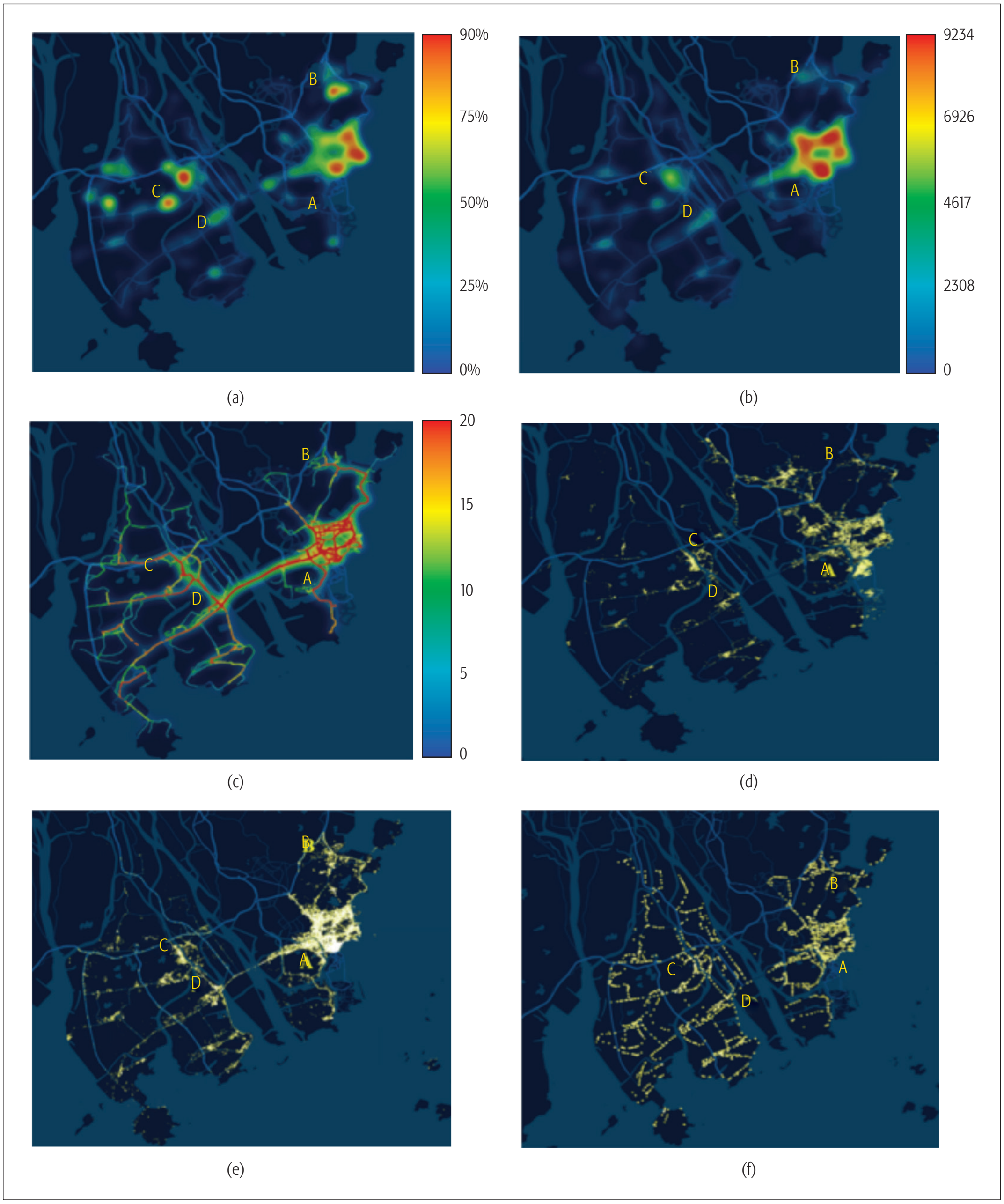


Figure 5. The location loss event heat map compared to other source of maps: a) heat map of the location loss event; b) heat map of the passengers' boarding event; c) heat map of the city's transportation network; d) the spatial distribution of POI; e) the spatial distribution of the cellular base station; f) the spatial distribution of public bus stops.

using pattern recognition and statistics inference. Moreover, we have thoroughly discussed the spatial distribution of location data loss as well as its impact on passenger flow analysis, which put forward the requirement for better IoT infrastructure.

Acknowledgment

Jiafu Wan is the corresponding author. This research is sponsored by the 2015 Science and Technology Project of the Guangdong Provincial Department of Transportation (no. 2015-02-076), the Fundamental Research Funds for the Central Universities (no. 2015ZZ079), the Natural Science Foundation of Guangdong Province, China (Nos. 2016A030313734, 2016A030313735), and the SCUT Grid high performance computation platform.

References

[1] I. Vilajosana *et al.*, "Bootstrapping Smart Cities through a Self-Sustainable Model Based on Big Data Flows," *IEEE Commun. Mag.*, vol. 51, no. 6, June 2013, pp. 128–34.

[2] K. Farkas *et al.*, "Crowdsending Based Public Transport Information Service in Smart Cities," *IEEE Commun. Mag.*, vol. 53, no. 8, Aug. 2015, pp. 158–65.

[3] J. Wan *et al.*, "Context-Aware Vehicular Cyber-Physical Systems with Cloud Support: Architecture, Challenges and Solutions, " *IEEE Commun. Mag.*, vol. 52, no. 8, Aug. 2014, pp. 106–13.

[4] M. Imran *et al.*, "A Novel Wireless Sensor and Actor Network Framework for Autonomous Monitoring and Maintenance of Lifeline Infrastructures," *IEEE ICC*, June 2012, pp. 6484–88.

[5] F. Chen *et al.*, "Data Mining for the Internet of Things: Literature Review and Challenges, " *Int'l. J. Distrib. Sens. N.*, vol. 2015, no. 8, Aug. 2015, Article ID 431047.

[6] X. Ma and Y. Wang, "Development of a Data-Driven Platform for Transit Performance Measures Using Smart Card and GPS Data," *J. Transp. Eng.*, vol. 140, no. 12, Dec. 2014, pp. 04014063-1–12.

[7] L. M. Kieu, A. Bhaskar, and E. Chung, "Passenger Segmentation Using Smart Card Data," *IEEE Trans. Intell. Transp.*, vol. 16, no. 3, June 2015, pp. 1537–48.

[8] W. L. Wang, S. M. Lo, and S. B. Liu, "Aggregated Metro Trip Patterns in Urban Areas of Hong Kong: Evidence from Automatic Fare Collection Records," *J. Urban Plann. Dev.*, vol. 141, no. 3, Sept. 2015, pp. 0514018-1–10.

[9] T. Sui, C. Jonathan, and I. Mateo-Babiano, "Exploring Bus Rapid Transit Passenger Travel Behaviour Using Big Data," *Appl. Geogr.*, vol. 93, Sept. 2014, pp. 90–104.

[10] M. Marcela, D. Flavio and N. Claudio, "Validating Travel Behavior Estimated from Smartcard Data," *Transport Res. C-Emer.*, vol. 44, July 2014, pp. 70–79.

[11] F. Zhang, N. J. Yuan, and Y. Wang, "Reconstructing Individual Mobility from Smart Card Transactions: A Collaborative Space Alignment Approach," *Knowl. Info. Syst.*, vol. 44, no. 2, Aug. 2015, pp. 299–323.

[12] X. Ma, Y. Wang, and F. Chen, "Transit Smart Card Data Mining for Passenger Origin Information Extraction," *J. Zhejiang Univ-Sci. C.*, vol. 13, no. 10, Oct. 2012, pp. 750–60.

[13] M. Guizani *et al.*, "Large-Scale Cognitive Cellular Systems: Resource Management Overview," *IEEE Commun. Mag.*, vol. 53, no. 5, May 2015, pp. 44–51.

[14] X. Ning, M. Daniel, and O. M. Leon, "A Walk into Metaheuristics for Engineering Optimization: Principles, Methods and Recent Trends," *Int'l. J. Comp. Int. Sys.*, vol. 8, no. 4, July 2015, pp. 606–36.

[15] G. Han et al., "Mobility Support for Next-Generation Wireless Sensor Networks," *Int'l. J. Distrib Sens N.*, vol. 2016, Article ID 2462754.

Biographies

Yongxin Liu (pcwhy@163.com) received his B.S. and M.S. from South China Agricultural University (SCAU) in 2011 and 2014, respectively, and he is currently a Ph.D. candidate in the School of Civil Engineering and Transportation, South China University of Technology (SCUT). He has authored and co-authored more than 10 peer-reviewed papers in academic journals and conferences. His major research interests include data mining, wireless network, the Internet of Things, and unmanned aerial vehicles.

Xiaoxiong Weng (ctxxweng@scut.edu.cn) is a professor in the School of Civil Engineering and Transportation, SCUT. She has authored/ co-authored more than 20 peer-reviewed papers. Her research interests include intelligent transportation systems, dynamic modeling of urban traffic flow, data mining on transit systems, traffic signal control systems, and transit commuter behavior analysis.

Jiafu Wan [M] (jiafuwan_76@163.com) is a professor in the School of Mechanical and Automotive Engineering, SCUT. Thus far, he has authored/co-authored more than 70 journal papers (with 60+ indexed by ISI SCIE) and 30 international conference papers. His research interests include cyber-physical systems, Industry 4.0, smart factory, industrial big data, industrial robots, and the Internet of Vehicles. He is a member of ACM.

Xuejun Yue (yuexuejun@scau.edu.cn) is a professor at SCAU, where she has been working on precision agriculture technology, wireless sensor networks, and UAV engineering. She has been teaching in areas including advanced information perception, WSN, cloud computing, and UAV engineering. Her research includes crowdsourcing information perception, wireless communication, and UAV application systems.

Houbing Song (h.song@ieee.org) received his Ph.D. degree in electrical engineering from the University of Virginia in 2012. His research interests are focused on optical communications, wireless communications and networking, vehicular networking, smart grid communications, and cyber-physical systems. He worked with Shandong Academy of Sciences as an assistant research scientist from 2004 and 2005 and with Texas Transportation Institute (TTI), Texas A&M University as an engineering research associate in 2007.

Athanasios V. Vasilakos (vasilako@ath.forthnet.gr) is currently a professor at Luleå University of Technology, Sweden. He has authored or coauthored over 200 technical papers in major international journals and conferences. He is author/coauthor of five books and 20 book chapters in the areas of communications. He has served as General Chair and Technical Program Committee Chair for many international conferences. He has served or is serving as an Editor for many technical journals and is General Chair of the Council of Computing of the European Alliances for Innovation.

We believe that the occurrences of location data loss in bustling areas have severe side effects on passenger traffic pattern mining. On the other hand, IoT devices and mobile communication sub-systems in urban public transportation systems, needs further enhancements to support the ongoing demands of intelligent public transportation systems in smart cities.

Smarter Markets for Smarter Life: Applications, Challenges, and Deployment Experiences

Hyung-Sin Kim, JeongGil Ko, and Saewoong Bahk

The authors introduce application requirements and technical challenges for smart market applications and pilot deployment experiences in an urban large-scale market. Their pilot deployment experiences of wireless systems in markets provides quantified knowledge on the impact of various in-market activities on the wireless link performance and led us to design MarketNet, a wireless networking architecture suitable for IoT-based smart market systems.

ABSTRACT

While a number of studies reveal the performance and effectiveness of applying wireless systems to various smart city applications, surprisingly, a market environment, in which we rely on a daily or weekly basis for purchasing essential goods, is still understudied. A wireless system in a market, along with rapidly growing IoT technology, can enable interesting applications such as automated electronic price tag updates, shopping-cart-based advertisements and information display, and automated inventory/stock management. These applications can first benefit market staff by automating routine tasks or opening possibilities for additional profits. Not only that: for everyday customers, these applications provide more reliable services by eliminating many man-made errors and offer additional services that can be enabled by newly gained connectivities. In this work, we introduce application requirements and technical challenges for such smart market applications and pilot deployment experiences in an urban large-scale market. Our pilot deployment experiences of wireless systems in markets provide quantified knowledge on the impact of various in-market activities on the wireless link performance and led us to design MarketNet, a wireless networking architecture suitable for IoT-based smart market systems. We see this work as a first step in bringing the attention of many researchers to a new application area, where wireless systems have the potential to benefit the quality of life of many smart city residents.

INTRODUCTION

Markets, by definition, are places where people gather to purchase and sell goods that are essential to their everyday lives. As consumers we visit markets on a weekly (if not daily) basis. Historically, well-operated marketplaces have catalyzed population growth and formed the basis of a city. Therefore, a critical but until now less explored application for realizing smart cities is designing a smarter market environment. It is true that the introduction of novel technologies such as barcodes, computers, RFID, and smartphones have dramatically changed the operational flow and shopping cycle in markets [1]. Nevertheless, with recent enhancements in low-power embedded computing technologies coupled with the explosion of interest in the Internet of Things (IoT), we see another chance to revolutionize today's market environments by connecting various in-store items to the Internet.

This article presents an overview of how different applications change a market's operational flow, ease internal management, and benefit shopping experiences for customers. Using recent surveys,[1] we identify unique application challenges that an embedded wireless system would face and should overcome when targeting market deployments. Furthermore, using empirical field studies performed in an urban market environment, we provide answers to how daily market activities impact the wireless link and network performance. With the findings from these studies, we design a prototype system to realize a smart market environment, MarketNet, and validate its usefulness with pilot studies in a real market. Finally, we leverage our experiences from this pilot study to outline remaining issues and discuss the steps necessary to bring smarter markets to reality.

WIRELESS SYSTEMS FOR SMARTER MARKETS

The development of low-power and low-cost computing and communication technologies has introduced a number of smart *anything* applications over the last decade. These research efforts are now being commercialized with the IoT hype to make traditionally manual environments smarter than ever before. The successes in various application domains stimulate the need to design smarter market environments that are tightly coupled with our everyday lives. The subsequent sections outline a few potential applications of smart markets and identify novel challenges based on a series of interviews with urban market managers. Specifically, in this work we use the term *smart market* for market environments that enable IoT connectivity for supporting applications such as the following. With the infrastructure for such applications, we envision that personal devices such as smartphones and tablets can be an effective interface for interacting with customers (e.g., combination of online and in-store shopping experiences) and staff members (e.g., less operational burden).

SMART MARKET APPLICATIONS

Labor-Free Price Tag Reconfiguration: The market managers with whom we worked reported that item prices should change dynamically

[1] We used three sets of interviews (i.e., EY, Accenture, and a local Korean interview). EY's interview was undertaken with 69 retailers and suppliers across 11 countries over the period December 2014 – July 2015. Accenture's one was conducted with 162 companies and 10,096 consumers across more than 10 countries in November 2015. To double check this, we conducted a local interview with several market managers in Korea.

Digital Object Identifier: 10.1109/MCOM.2017.1600260

Hyung-Sin Kim and Saewoong Bahk (corresponding author) are with Seoul National University; JeongGil Ko is with Ajou University.

according to many external factors such as customers' purchase patterns, inventory, competing markets' responses, and goods freshness [2]. In addition to this, the status of each customer, such as member status, or discount information can further complicate the pricing policy. Moreover, price tags present other details of products. such as the origin, precaution information, and detailed usage instructions. However, in most markets, a large number of these price tags are still manually managed, incurring labor and printing costs, man-made errors, and delayed updates. With recent development of low-power displays (e.g., e-ink) and wireless networking technologies, automatically reconfiguring price tags is becoming a feasible option. Such application systems allow markets to not only reduce labor and printing costs for price tag updates, but also update prices more dynamically with reduced human error to better meet customer needs.

Automated Inventory and Stock Management: Currently, the process of checking the rack status and maintaining the inventory is mostly performed manually, with staff members hovering over the entire market to identify missing items.

As a result, 27 percent of consumers reported that solving the out-of-stock problem is the most critical factor to improve their in-store shopping experiences. With energy-efficient sensing techniques (e.g., infrared and/or weight sensors) combined with low-power wireless networks, we can design a system where the status of each rack is reported in real time, and the data or pattern of purchases can be autonomously analyzed for predictive ordering.

Shopping-Cart-Based Applications: Shopping carts in markets are one of the most widely accessed "human interaction interfaces" but are still underutilized. Some shopping carts have small plastic holders for simple advertisements, but since they cannot be updated in real time, a majority of these advertisements focus on prevalent topics such as benefits of enrolling in the market's membership program. Low-power screen technologies and minimal power consuming radios combined with energy harvesting modules on cart wheels open interesting applications. For example, given that only 7 percent of retailers enable real-time promotions to customers through their smartphones, these shopping carts can be good media to inform customers of deals in which they may be interested. Furthermore, cart location tracking and virtual fencing can ensure that carts are not taken off market premises.

Customer Pattern Analysis: An important piece of information that market managers can benefit from is the purchase and movement patterns of customers, which allows the managers to plan their product layouts over the ideal shopping route. To obtain this information without cameras, which incur a security threat, we may consider a wireless network (especially a low-power network) given that its positive side-effect is in variations of signal strength with the population intensity in a target region. In such applications, the locations of individual customers are not needed. Rather, the trend of movement, say to answer questions such as "When do customers tend to move more toward items near the refrigerators?" can provide hints on how the items can be displayed.

Challenges for Designing Smart Market Systems

While low-power wireless sensing systems have the potential to enable a number of exciting applications, the current state-of-the-art technology still faces a number of technical challenges, especially at the system design level, when we start to design practically meaningful systems for market environments. We illustrate a few of these major challenges below.

Low-Power Operations and Minimal Costs: System designers should consider the fact that nodes will be deployed in large quantities, which incurs the need for low device and operational costs. From the perspective of system designers, such constraints raise many difficulties. First, the software should be lightweight and carefully engineered. Second, given that low-cost hardware modules themselves are prone to hardware faults, applications requiring high reliability (e.g., price tag updates) should consider fault tolerance as an additional factor. Third, since wireless modules need to operate within "cost-free" frequency band, there should be considerations for external radio interferences depending on operating frequency (e.g., 2.4 GHz) and deployment area.

Finally, battery operated nodes (tagged to items, carts, or shelves) may face significant challenges in maintaining a practical long enough lifetime. The selection of hardware components becomes critically important, and low-power operation schemes such as radio duty-cycling are a must.

Diversified Traffic Patterns: From a wireless traffic perspective, market applications can introduce complicated network traffic, unlike "collection-oriented" patterns for most wireless sensing systems. For example, price tag updating and real-time advertising applications generate a large amount of downward (from the server to individual nodes) traffic, rack status updates generate upward (from individual nodes to the server) traffic, and applications such as cart tracking require active bidirectional packet exchange. Compared to the simplified world of collection-oriented data traffic patterns, such diversity in wireless traffic brings dramatic changes in system design. Most importantly, the wireless networking protocols used with the applications of today may not be suitable for these environments. Furthermore, traffic diversity also impacts node lifetime given that complex node wake-up patterns are possible in a full multihop network.

Data Reliability over Mass-Scale and Real-Time Data Delivery: Applications in markets are very sensitive to reliability since the performance of the wireless system directly affects the market profit and complaint ratio. In fact, data delivery reliability is the highest system-level priority in a list of requirements provided by market managers. Therefore, it is important that our systems show high reliability with low latency for a large number of nodes. This is especially true for applications such as remote price updating and rack status updates.

Usability and Manageability: Finally, once our system is deployed, the system should provide a user interface (UI) for market managers or employees to effectively utilize without the help of system developers. This requirement suggests

With recent development of low-power displays and wireless networking technologies, automatically reconfiguring price tags is becoming a feasible option. Such application systems allow markets to not only reduce labor and printing costs for price tag updates, but also update prices more dynamically with reduced human errors to better meet customer needs.

While the wireless research community has revealed the link characteristics of various environments [3–5], market environments are still unknown. Nevertheless, like many smart city applications, a market environment holds characteristics such as active human mobility, unique environment-specific activities, and various factors for RF interference.

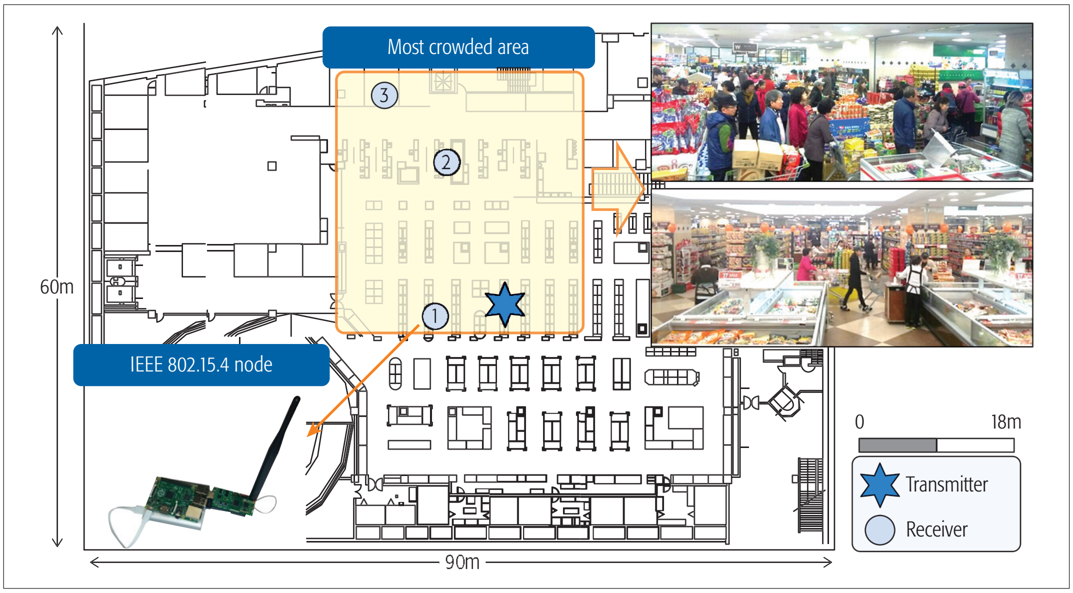


Figure 1. Link testing topology in a crowded urban market. We deployed a transmitter and three receivers in the most crowded area of the market for link quality testing.

building an easily controllable network for typical IT personnel within the market, and also asks for a further study from a human-computer interaction perspective. For designing an IT-personnel-friendly system, we believe that IPv6 protocols for low-power and lossy networks (LLNs) help staff by allowing them to utilize IP-level management services. This IP connectivity can also be used to support IP-level security protocols such as IPSec to the market system.

Wireless Environments in Markets

In order to deploy wireless IoT systems for markets, we found the need to better understand wireless channel characteristics in such environments. While the wireless research community has revealed the link characteristics of various environments [3–5], market environments are still unknown. Nevertheless, like many smart city applications, a market environment holds characteristics such as active human mobility, unique environment-specific activities, and various factors for RF interference. To quantify their impacts, we present a set of results collected from a study performed at an urban market located in Seongnam, South Korea. Specifically, this market holds more than 10,000 types of items and is visited by more than 5000 customers every day.

Within this market, we start our studies by deploying four nodes in the most crowded area (i.e., near the checkout counter). To maintain a low-power profile of the nodes, we use IEEE 802.15.4 radios (CC2420, data rate of 250 kb/s) for the nodes. One of the four nodes, shown as the star in Fig. 1, was set as a periodic data sending unit with a payload size of 72 bytes, an inter-packet interval (IPI) of 50 ms, and transmission power of 0 dBm. We perform our studies on the IEEE 802.15.4 channel 26 (free from WiFi in the United States, but not in Korea).[2] We use TinyOS as an embedded software for our experiments. With this setup, we measured link characteristics of the market for 10 hours (12:00 to 22:00) of a weekday. The goal of this preliminary study is to understand how low-power wireless signals would perform, or be distracted, within a busy market environment.

Among various results that reveal wireless channel characteristics of our target market, Fig. 2 shows the results of per-minute packet reception ratio (PRR) and per-minute average received signal strength indicator (RSSI) for the three links that we tested (i.e., one link for each transmitter-receiver pair). From these results, we obtained some very interesting observations. First, while analyzing the data, we noticed that the RSSI performance of all three links dropped very sharply at ~14:30. Following this event, we noticed that the PRR dropped for links 2 and 3. This was an unexpected observation, and we were surprised to see similar phenomena occurring on a daily basis throughout our experiments. This question led us to observe activities taking place near our deployments throughout the day, which eventually revealed that a shelf-filling activity took place on a daily basis at the location where the transmitter was located. This activity introduced an additional layer of obstacles on the wireless links, causing links with longer lengths (e.g., links 2 and 3) to quickly suffer from PRR degradation. We also noticed that in our market of interest, the shelves were not filled entirely in a single round. Rather, the employees took multiple rounds in filling up the shelves. As a result, we can see the RSSI for link 1 degrading in steps, eventually leading to PRR loss. These observations from our pilot studies suggest that market environments introduce a new level of unexpected performance where the quality of wireless links can be affected heavily by routine market-specific activities.

In addition to the impact of typical market activities, the daily interaction of customers with shelves (i.e., strolling around the shelves and picking products up) may introduce high and frequent fluctuations in the channel between the transmitter and the receiver. To make a more detailed

[2] Wireless interference conditions can be changed in different frequency band and/or different locations.

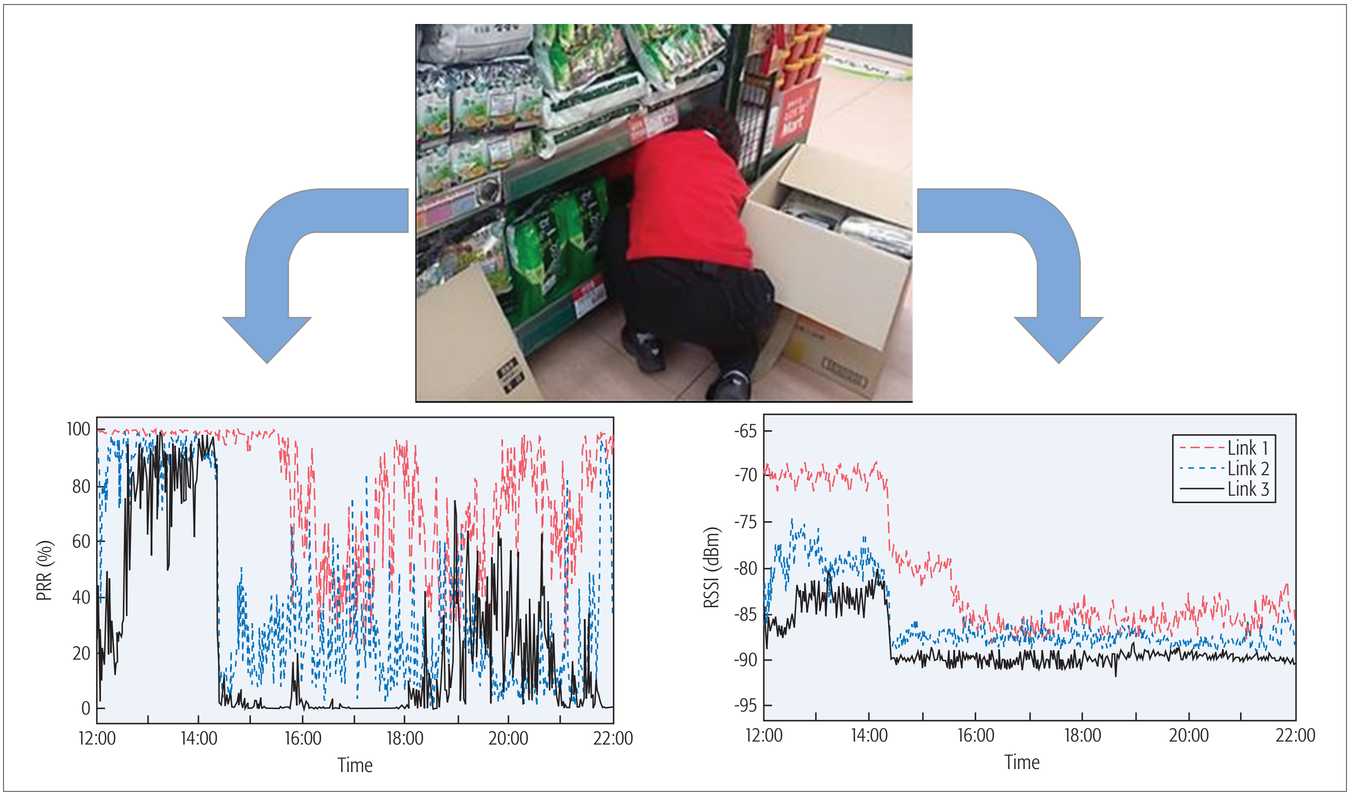


Figure 2. Packet reception ratio (PRR) and received signal strength indicator (RSSI) of links 1, 2, and 3 for 10 hours. The results reveal that a crowded urban marketplace has very dynamic wireless links which fluctuate in both short- and long-term manners.

observation and quantify this, we manually counted the number of active customers within the market and also measured the conditional probability density function (CPDF) [6] of the links corresponding to the same time of day. CPDF corresponds to the probability of a packet being successfully received after *n* consecutive failures or successes. Negative numbers represent consecutive successes, while positive numbers represent consecutive failures. For example, CPDF(20)/CPDF(–20) is the probability of a successful delivery after 20 consecutive failures/successes on the link. Therefore, the CPDF is a good measure of link burstiness (the channel coherence time, when combined with IPI).

In Fig. 3 we plot per-hour CPDF for link 2 and the per-hour number of customers from 16:00 to 22:00. First notice that the CPDF values are very small when $n > 0$ and large when $n < 0$. This non-uniformity of the CPDF plots shows that links in the market environments are heavily bursty. Furthermore, we can observe that the length of the CPDF's left tail varies over time. Given that the left tail length represents the maximum number of consecutive successes, a long left tail length implies that the wireless link shows bursty performance. Here, we notice that the number of customers, which sharply decreases from 20:00, shows an inverse correlation with the link's burstiness.

Finally, we also report that links which were closely located to the booths where finger food was offered for free tasting experienced a high level of fluctuation due to the frequent use of the microwave oven. The impact of microwave ovens on 2.4 GHz radios is well known, but can be a practical issue for market deployments.

MarketNet: A Wireless System for Smart Markets

MarketNet Design

There have been a number of studies that apply wireless technologies for smart market systems, and the industry shows interest in these systems (e.g., M²Communication, Pricer, and LG Innotek). To this end, various wireless technologies have been investigated, from low-power/cost radio such as IEEE 802.15.4 and RFID to infrared/visible light [7–9]. From the perspective of network architecture, all employed systems for smart markets allow each low-power node (e.g., electronic price tag) to communicate with a gateway via a single hop. This single-hop architecture is simple, but requires many gateways to provide wireless connectivity for all nodes deployed in a large urban marketplace under highly fluctuating and bursty wireless links. This incurs device cost, installation, and management issues in practice.

Low-power multihop network architectures could be an alternative due to the reduced number of gateways. However, the link characteristics in a market environment, along with the requirement and technical challenges introduced earlier, suggest that the traditional form of a data-collection-oriented multihop network of low-power nodes may not be suitable for market applications. Specifically, the increased interest in network traffic diversification provides us with a reason to abandon the concept of optimizing the routing "tree" toward a single direction and consider routing efficiency in both directions. Furthermore, active link quality variations in a market environment introduce the need for a very robust networking architecture tolerant against a large

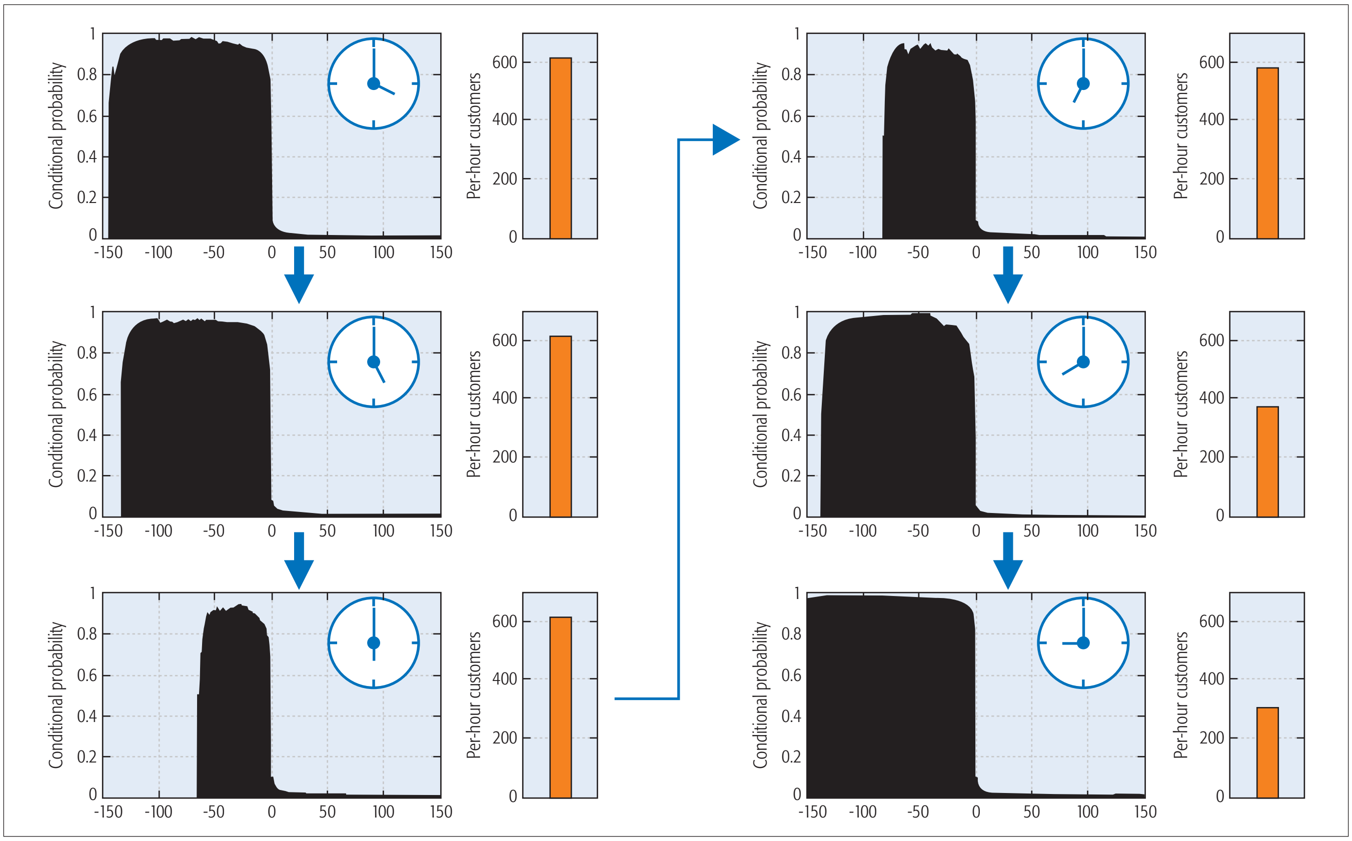


Figure 3. Per-hour conditional probability density function (CPDF) of link 2 and per-hour number of customers for six hours (from 4 p.m. to 10 p.m.).

amount of link quality fluctuation. Our preliminary experimental results with collection-oriented protocols such as RPL (IPv6 routing protocol for low power and lossy network) [10, 11] in a market environment show that these protocols are unable to adjust link selections robustly enough to provide bidirectional reliability (especially for downward packet delivery) [12]. LOADng [13], another recent Internet Engineering Task Force (IETF) standard routing protocol that is not collection-oriented but ad hoc, even underperforms RPL in terms of latency, memory overhead, and control packet overhead [14].

After analyzing the benefits and disadvantages of traditional single- and multihop network architectures, an aspect that we found inefficient was the homogeneous network transmission powers for all nodes. It is true that, for many low-power network deployments, transmission power homogeneity is considered to be natural and the most efficient way to operate a network. However, under diverse traffic patterns and the fact that market deployments are sure to have always-powered nodes like a network gateway, we consider an option of utilizing a heterogeneous network architecture. Specifically, we imagine a network where the gateway node holds a radio with higher transmission power (using a signal amplifier) compared to other low-power nodes. This option potentially simplifies network management by reducing the number of transmission hops (or the number of gateways) in downward traffic-oriented applications.

The aforementioned issues identified for market deployments encourage us to rethink a novel wireless networking architecture suitable for market environments, which takes potential application traffic patterns, unique link characteristics, and practical system-deployment-related factors into consideration. MarketNet is our solution to designing more efficient low-power wireless systems for market environments. Specifically, MarketNet utilizes asymmetric transmission power between nodes, where the gateway sends packets at a much higher transmission power to reach nodes via a single hop, while individual low-power nodes use a multihop architecture to deliver packets to the gateway. The single-hop downward transmission eliminates the need to use unreliable downward routing in fluctuating link environments requiring memory and control overhead. Lastly, MarketNet combines this network architecture with the IEEE 802.15.4 link layer since it provides IPv6 6LoWPAN connectivity [10] and RPL [11], improving usability and manageability.

We illustrate the core operational functionalities of MarketNet using Fig. 4. MarketNet utilizes the gateway's high transmission power capability. First, given that the gateway transmits its control signal to all nodes in a single hop, we achieve global time synchronization and allow nodes to share a single superframe architecture. In MarketNet, a superframe period consists of a beacon period, a downlink period, an uplink period, and an inactive period.[3] Specifically, the gateway broadcasts beacons in the beacon period and transmits downward traffic (e.g., price tag update information and shopping cart advertisements) within the downlink period. Each node wakes up (e.g., activates its radio after being in inactive mode to conserve energy) at the start of a beacon

[3] All time periods need to be configured according to system requirements such as latency and energy consumption.

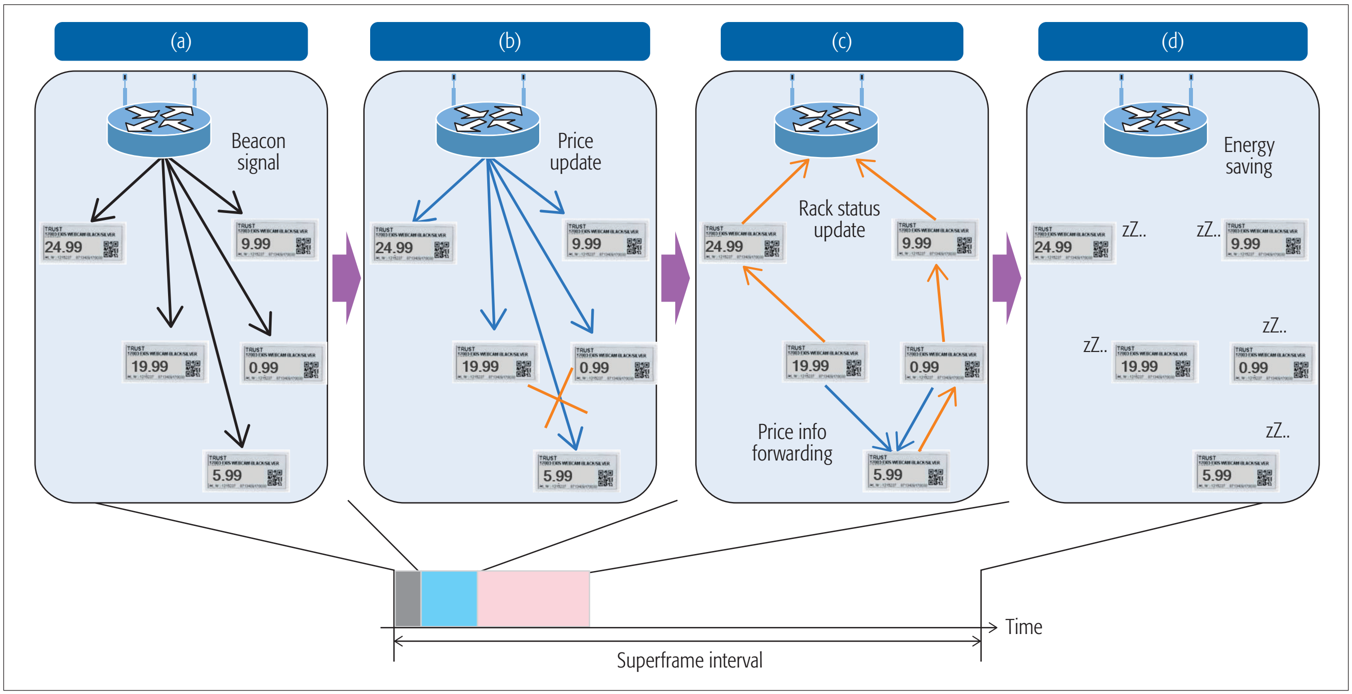


Figure 4. MarketNet architecture: a) beacon period; b) downlink period; c) uplink period; d) inactive period. Using gateways that exploit higher transmission power raather than battery-powered nodes as proposed provides network-wide time synchronization, and single-hop downward with multihop upward packet delivery.

period to receive a beacon, transmits upward traffic (e.g., rack status information or cart locations) in the uplink period, and turns off its radio once again in the inactive period for energy-efficient operations. This separation of low-power node transmissions from the gateway's transmissions (e.g., separation of uplink and downlink) ensures that high-power transmissions from the gateway do not interfere with transmissions from low-power nodes. Note that for the multihop upward traffic delivery, existing data collection protocols such as RPL [11] can be used to provide reliable packet delivery as well as IPv6-based addressing for simplifying the network management operations.

Nevertheless, the asymmetric transmission power-based network architecture holds some fundamental issues in operation. One of the main issues is that the asymmetry in transmission power restricts acknowledgment frames from low-power nodes from reaching the gateway node in a single hop. Even if the gateway uses a high transmission power to send packets to make single-hop transmissions, the acknowledgment frames should be transmitted over a multihop path, resulting in additional transmission overhead. To alleviate this issue, we allow neighbor nodes of a low-power destination node to overhear both downward packets from the gateway and the acknowledgments from the destination node. If a neighbor node receives a downward packet but not an acknowledgment, it detects a failed transmission and forwards the overheard packet to the destination node. If both data and acknowledgment are received, the neighbor node discards the overheard data with the conclusion that the packet was successfully delivered. This forwarding scheme allows MarketNet to support reliable downward packet delivery over fluctuating wireless links in an asymmetric transmission power network, without multihop transmissions of acknowledgment frames. We refer interested readers to [12] for previous work and further details on the MarketNet architecture.[4]

Performance Evaluation

Given that MarketNet was designed for real market environments, we performed preliminary evaluations of our proposed network architecture in the same market where we performed our link studies. In this environment, we deployed a single gateway node with a 10 dB amplifier so that it covers the entire market. As a test deployment, we distributed 30 nodes on the item racks, which were used to emulate a price tag and rack status update application scenario. While the downward link connectivity was covered via a single hop, we noticed that the low-power nodes used multihop upward routes with a maximum of three hops to the gateway node by using the RPL protocol.

Figure 5 plots the mean PRR performance (and the standard deviation for all nodes' performance) for standalone RPL (multihop upward and downward routes) and MarketNet under various configurations (i.e., MarketNet with and without time synchronization). We can see that RPL's PRR degrades in downward-focused traffic scenarios. Given that traffic such as price updates or advertisement distribution rely on downward traffic, we see this as a major issue in system reliability. Nevertheless, this is surprising since RPL was designed to perform well in lossy environments. We see such results occurring because, while RPL was designed for links with active losses, it was designed to tolerate losses on upward routes. In other words, with upward path link losses, RPL would actively reconstruct its routes. However, since RPL uses the reverse of upward routes as downward routes, unless active losses occur

[4] This work can be differentiated from the previous work as follows. First, we describe smart market applications, requirements, and remaining challenges in more detail. Second, we add some link measurement results which are not part of the prior work. Third, we describe differences between MarketNet and previous approaches (both of industry and research community) in a more practical point of view.

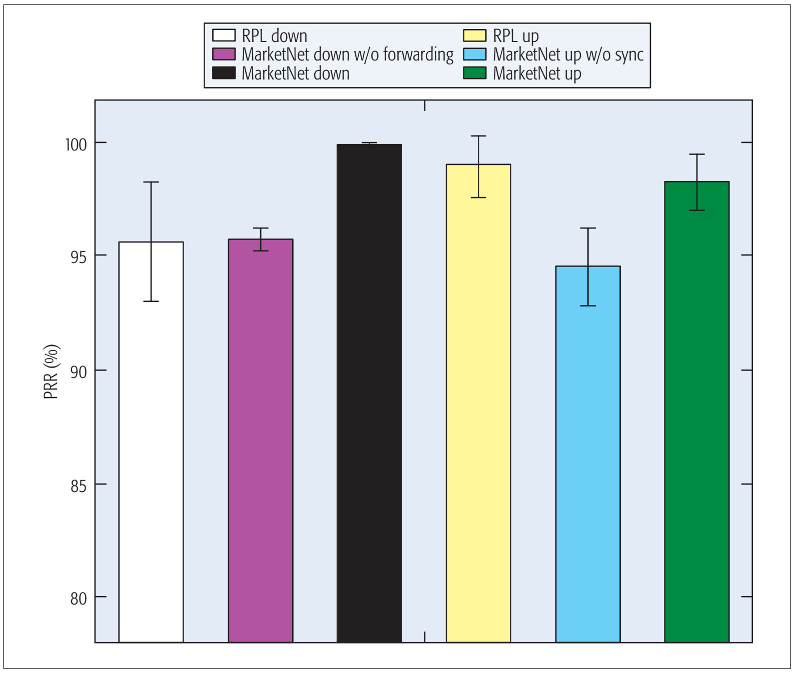


Figure 5. Bidirectional packet reception ratio performance of competitive network protocols in a crowded urban market environment.

upward, downward link losses will not trigger a route change. Therefore, for scenarios with more downward traffic than upward, RPL fails to maintain a high PRR.

On the other hand, compared to RPL, MarketNet reduces its PRR variance among nodes even without the neighbor forwarding scheme. This is due to proper provisioning of the deployment topology and careful deployment of nodes in the field. When MarketNet includes neighbor forwarding, it achieves ~100 percent PRR for downward packets since an effective packet retransmission scheme is added. This result reveals that the combination of asymmetric transmission power and neighbor forwarding creates a meaningful synergistic effect, making it suitable for market deployments.

Since MarketNet utilizes RPL as the underlying protocol for multihop upward packet delivery, the best case scenario is to perform as well as RPL. Nevertheless, without the network-wide time synchronization, downward high-power transmissions cause a significant amount of interference for upward packets, thus resulting in degraded uplink PRR performance. However, MarketNet's network-wide superframe allows for a separation between upward and downward packets, allowing a similar uplink PRR compared to RPL.

Lastly, Fig. 6 plots the radio duty cycle of the low-power nodes over time. Here, we can observe that RPL not only provides the (relatively) worst duty cycle performance, but also shows significantly unfair duty cycle performance with some nodes experiencing very high duty cycles (e.g., red-crossed outliers). As reported in some previous work, this is an issue with RPL's load balancing performance [15]. Given that human intervention is needed when at least one battery-powered device becomes inactive, this severely unfair duty cycle performance can lead to a significant amount of manual labor.

MarketNet, without the network-wide time synchronization, shows that an asymmetric transmission power-based network improves duty cycle performance by eliminating transmissions of routing control packets (e.g., destination advertisement objects [DAOs] in RPL) and multihop downward data packets. Furthermore, the radio duty cycle performance of MarketNet with all features enabled suggests that a network-wide superframe architecture using high-power beacon transmission significantly improves nodes' radio duty cycles with near perfect fairness.

Summary and Remaining Challenges

This work started with a surprise that while markets are actively accessed within our everyday lives, IoT systems technology has not yet reached its gates, despite the potential for introducing interesting applications. In fact, market managers were very enthusiastic about applying wireless IoT technologies for opening new opportunities for profit making, and also as a way to reduce the frequency of customer complaints caused by today's manual operations.

Our studies show, however, that these market environments, due to their diversity in activities, are not wireless friendly. To mitigate various market-oriented obstacles, we designed and tested a prototype system, MarketNet, as a low-power wireless IoT network infrastructure for smart market systems. We envision that smartphones and tablets can be added on top of this infrastructure for more advanced services (e.g., price inconsistency alarms).

As systems scale up for higher density and larger quantities, there is a chance that even a system like MarketNet can show unexpected performance. Furthermore, another practical challenge based on our pilot study with MarketNet was tampering with our devices. This calls for a series of extensive studies on the performance of wireless systems in markets and also suggests the need for a durable enclosure as we reach commercialization of wireless systems for enabling smart market applications. We foresee this work as a meaningful stepping stone in asking the research community to actively start research in this surprisingly less explored application domain, full of interesting challenges waiting for us to address.

Acknowledgment

For H. -S. Kim and S. Bahk, this work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2015R1A2A2A01008240), and in part by the ICT R&D program of MSIP/IITP [B0717-16-0026].

For J. Ko, this work was supported by the Ministry of Science, ICT, and Future Planning (MSIP) Korea under the Information Technology Research Center (ITRC) support program (IITP-2016-R2718-16-0015), supervised by the National IT Industry Promotion Agency (IITP).

References

[1] V. Shanker *et al.*, "Innovations in Shopper Marketing: Current Insights and Future Research Issues," *J. Retailing*, vol. 87, no. 1, July 2011, pp. S29–42.
[2] D. S. Kung *et al.*, "Business Analytics: IT-Based System with Dynamic Pricing Algorithm," *Business J. Entrepreneurs*, vol. 2013, no. 2, 2013, pp. 117–27.

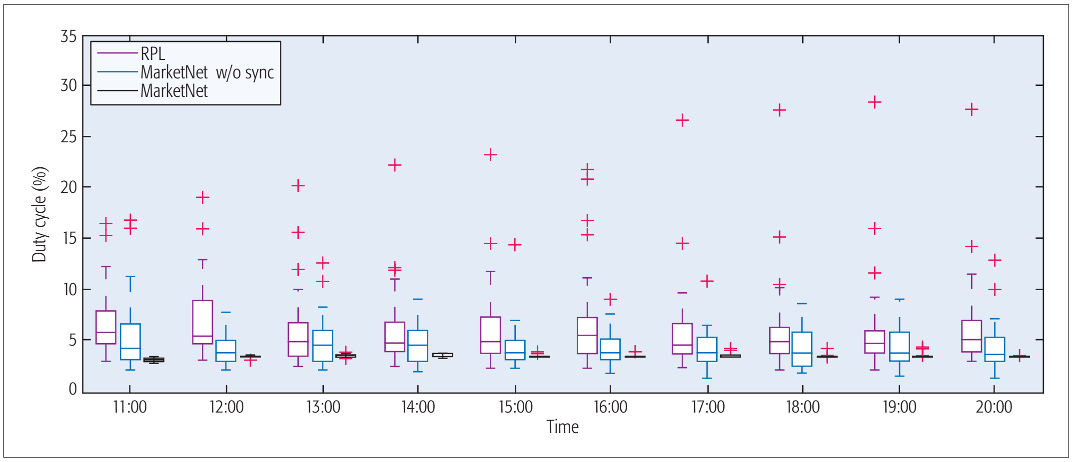


Figure 6. Radio duty cycle performance of various network protocols in an urban market environment over time.

Our studies show that these market environments, due to their diversity in activities, are not wireless friendly. To mitigate various market-oriented obstacles, we designed and tested a prototype system, MarketNet, as a low-power wireless IoT network infrastructure for smart market systems.

[3] J. Ko, T. Gao, and A. Terzis, "Empirical Study on a Medical Sensor Application in an Urban Emergency Department," *Proc. Int'l. Conf. Body Area Networks*, 2009.

[4] P. Huang *et al.*, "Energy-Efficient Computing for Wildlife Tracking: Design Tradeoffs and Early Experiences with ZebraNet," *Proc. Int'l. Conf. Architectural Support for Programming Languages and Operating Systems*, 2002.

[5] G. Werner-Allen *et al.*, "Deploying a Wireless Sensor Network on an Active Volcano," *IEEE Internet Computing*, Special Issue on Data- Driven Applications in Sensor Networks, Mar. 2006.

[6] H. Lee, A. Cerpa, and P. Levis, "Improving Wireless Simulation Through Noise Modeling," *Proc. Int'l. Conf. Information Processing in Wireless Sensor Networks*, Apr. 2007, pp. 21–30.

[7] J. Sung, "End of Paper Labels: Emerging Smart Labels toward Internet of Things," *Proc. IEEE WF-IoT'15*, Dec. 2015, pp. 216–21.

[8] G. Roussos, "Enabling RFID in Retail," *Computer*, vol. 39, no. 3, Mar. 2006, pp. 25–30.

[9] J. S. Park and B. J. Jang, "Electronic Shelf Label System Employing a Visible Light Identification Link," *Proc. IEEE RWS '16*, Jan. 2016, pp. 219–22.

[10] J. Ko, S. Dawson-Haggerty *et al.*, "Connecting Low-Power and Lossy Networks to the Internet," *IEEE Commun. Mag.*, vol. 49, no. 4, Apr. 2011, pp. 96–101.

[11] T. W. Ed. *et al.*, "RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks," IETF RFC 6550 , Mar. 2012.

[12] H.-S. Kim *et al.*, "MarketNet: An Asymmetric Transmission Power-Based Wireless System for Managing e-Price Tags in Markets," *Proc. ACM Conf. Embedded Networked Sensor Systems*, Nov. 2015, pp. 281–94.

[13] T. Clausen *et al.*, "The Lightweight On-Demand Ad Hoc Distance-Vector Routing Protocol — Next Generation (LOADng)," IETF Internet Draft , Jan. 2016.

[14] M. Vuini, B. Tourancheau, and A. Duda, "Performance Comparison of the RPL and LOADng Routing Protocols in a Home Automation Scenario," *Proc. IEEE WCNC '13*, Apr. 2013, pp. 1974–79.

[15] H.-S. Kim, J. Paek, and S. Bahk, "QU-RPL: Queue Utilization Based RPL for Load Balancing in Large Scale Industrial Applications," *Proc. IEEE Int'l. Conf. Sensing, Communication, and Networking*, June 2015, pp. 265–73.

Biography

Hyung-Sin Kim (hskim@netlab.snu.ac.kr) received his B.S. degree in electrical engineering from Seoul National University (SNU), Korea, in 2009. He received his M.S. degree in 2011 and his Ph.D. degree in 2016, both in electrical engineering and computer science, from SNU. From March 2016 to August 2016, he was a postdoctoral researcher at the Network Laboratory (NETLAB) at SNU led by Dr. Saewoong Bahk. He is currently a postdoctoral researcher of the Software-Defined Building (SDB) group at the University of California, Berkeley led by Dr. David E. Culler. His research interests include development of IoT systems for urban marketplaces, smart grid and SDB, and network and communication protocols for low-power wireless systems.

JeongGil Ko (jgko@ajou.ac.kr) received his B.Eng. degree in computer science and engineering from Korea University in 2007. He received his M.Sc. degree in engineering and Ph.D degree in computer science from Johns Hopkins University in 2009 and 2012, respectively. At Johns Hopkins, he was a member of the Hopkins interNetworking Research Group (HiNRG) led by Dr. Andreas Terzis. In 2010, he was at the Stanford Information Networking Group with Dr. Philip Levis at Stanford University as a visiting researcher. From June 2012 to August 2015, he was a senior researcher at the Electronics and Telecommunications Research Institute. He is a recipient of the Abel Wolman Fellowship awarded by the Whiting School of Engineering at Johns Hopkins University in 2007. His research interests are in the general area of developing web and cloud-based sensing systems with ambient intelligence for the Internet of Things and cyber physical systems. He is currently with the Department of Software and Computer Engineering, College of Information Technology, and also jointly with the Department of Biomedical Informatics, School of Medicine at Ajou University as an assistant professor, and leads the Ajou Embedded Intelligent Systems Laboratory.

Saewoong Bahk [SM] (sbahk@snu.ac.kr) is a professor at SNU. He served as director of the Institute of New Media and Communications during 2009–2011. Prior to joining SNU, he was with AT&T Bell Laboratories as a member of technical staff from 1991 to 1994, where he worked on network management. He received the KICS Haedong Scholar Award in 2012. He has led many industrial projects on 3G/4G/5G and IoT connectivity supported by Samsung Electronics, LG Electronics, SK Telecom, and others, and published more than 200 technical papers and holds 72 patents. He was TPC Chair for IEEE VTC-Spring 2014 and General Chair of JCCI 2015. He is co-Editor-in-Chief of the *IEEE/KICS Journal of Communications and Networks*, and was on the Editorial Boards of the *Computer Networks Journal* and *IEEE Transactions on Wireless Communications*. He is a member of Whos Who Professionals in Science and Engineering. He received B.S. and M.S. degrees in electrical engineering from SNU in 1984 and 1986, respectively, and a Ph.D. degree from the University of Pennsylvania in 1991.

WiFi Networks in Metropolises: From Access Point and User Perspectives

Lei Zhang, Liting Zhao, Zhi Wang, and Jiangchuan Liu

The authors have closely collaborated with a leading network service provider to collect massive information about wireless APs and their users in four metropolises. In this article, based on the large-scale dataset, they attempt to answer the critical questions on how the APs are deployed and how they are utilized in urban areas.

Abstract

Although WiFi was initially designed for providing local access to the Internet, today's expansive and explosive deployment of WiFi networks has enabled nearly ubiquitous access for mobile users in many urban areas. However, these WiFi networks have been woefully undermeasured and underinstrumented during the wild expansion. We have closely collaborated with a leading network service provider to collect massive information about wireless APs and their users in four metropolises. In this article, based on the large-scale dataset, we attempt to answer the critical questions on how the APs are deployed and how they are utilized in urban areas. At the macro level, we depict the coverage of WiFi networks and the usage patterns, and by carefully classifying the APs, we unveil rich geographical features of today's urban WiFi networks. At the micro level, we identify the implicit social relationships among WiFi users, and uncover the underlying social communities that have great potential for network optimization.

Introduction

Modern cities are ever increasingly expanding in terms of geographical coverage, residential population, and social and economic functionalities. With such speed of urbanization, today's metropolises are facing significant challenges in informational, intelligent, and integrated management. To build appreciable and measurable *smart cities* with seamless interconnection and interoperability, ubiquitous Internet access has become one of the underlying fundamentals. While *wireless local area network* (WLAN) technology was initially designed to provide local access for a limited number of users, the expansive and explosive deployment of 802.11 WiFi networks and the universal availability of WiFi interfaces in state-of-the-art mobile terminals has led to great Internet coverage nowadays, especially in urban areas. Besides business-oriented restaurants and hotels, such public service sectors as universities and community centers have deployed WiFi networks on tremendous scales as well. Many Internet service providers, together with regional governments, have also initiated plans toward city-wide WiFi coverage. Urban users reportedly have registered WiFi accesses 70 percent of the time [1], and are often exposed to multiple *access points* (APs) in one location.

Given that nearly 80 percent of mobile data usage is more nomadic than highly mobile [2], the massive urban WiFi networks have become major carriers, offloading much of the traffic from the conventional wired networks and the more expensive cellular mobile networks [3]. Considerable research efforts have been made on leveraging WiFi networks to balance traffic load, improve energy efficiency, and assist content delivery for mobile users. However, during this era of wild expansion, it remains unclear how the WiFi APs are deployed in modern cities, not to mention how they are connected and utilized by the massive number of users, and their dynamics over time. Pioneering studies on the deployed WiFi networks and the corresponding user access patterns have quite limited network scale [4], geographical area (e.g., campus only) [5], target application [6], and hardware platform [7]. Even though the state-of-the-art WiFi measurements [7] attempt to provide relatively comprehensive understanding from different layers and aspects, the social implications of WiFi users are seldom addressed.

In this article, based on a large-scale dataset (about *8 million WiFi APs and 27 million connection records from 6.4 million active users per day*) collected during a one-month period (March 12 to April 12, 2015) in four metropolises, we attempt to answer the questions of how WiFi networks are deployed and how they are utilized in urban areas. At the macro level, we depict the coverage of WiFi networks and the usage patterns, and by carefully classifying the wireless APs, we unveil rich geographical features of today's urban WiFi networks. At the micro level, we identify the implicit social relationships among WiFi users, revealing the underlying social communities that have great potential for network optimization. In particular, we find that:
- Today's WiFi networks are densely deployed in urban areas, which are driven by and naturally reflect the intensity of human social and economic activities.
- Business and public WiFi networks' deployments are more concentrated, whereas residential WiFi networks are more evenly distributed in urban areas.
- A majority of WiFi users have regular access patterns during workdays, and most user-AP connections have durations that are either ultra-short (less than 5 minutes) or quite long (longer than 1 hour).

Digital Object Identifier: 10.1109/MCOM.2017.1600262

Lei Zhang and Jiangchuan Liu are with Simon Fraser University; Liting Zhao and Zhi Wang are with Tsinghua University.

 0163-6804/17/$25.00 © 2017 IEEE

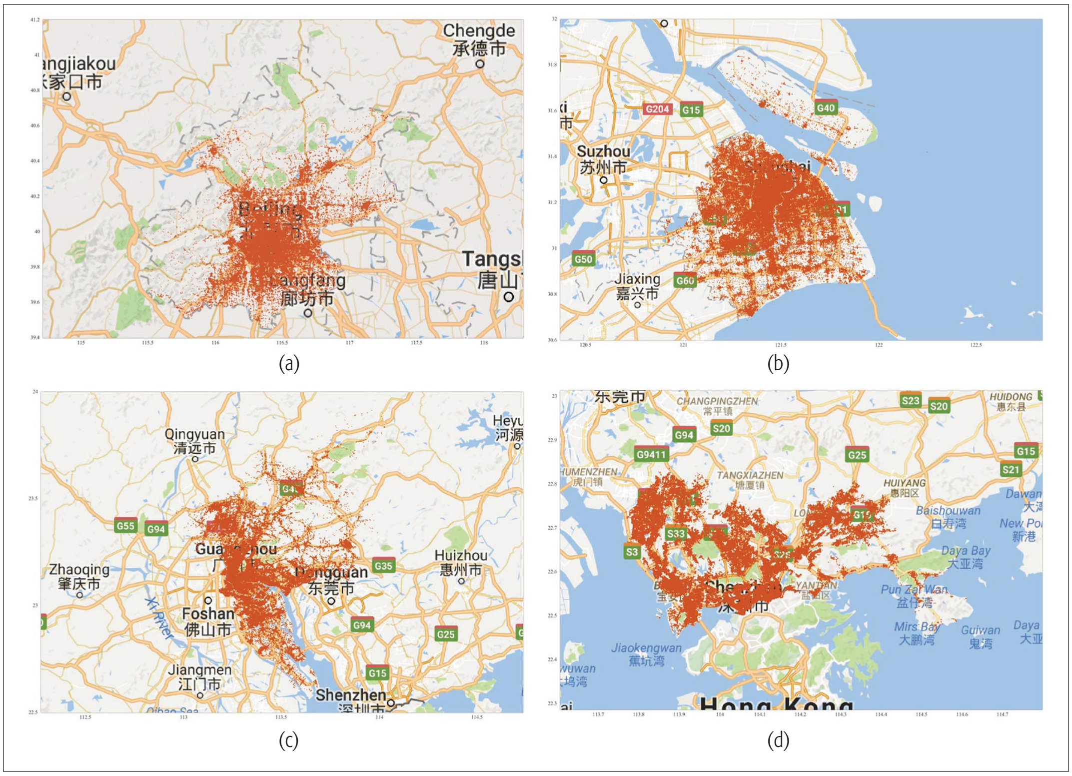


Figure 1. The deployments of wireless APs in four cities: a) Beijing; b) Shanghai; c) Guangzhou; d) Shenzhen.

To ensure secure access to the shared WiFi networks, the users are willing to report their connections to Tencent and have the traffic monitored. It is worth noting that only the related meta data is collected, so that none of the users' personal information was traced during the collection, nor were the conversation data intercepted or stored.

- The WiFi users geographically form a well structured small-world network, which has never been clearly identified by researchers and WiFi users before.

All these observations suggest that WLANs are shifting from a complementary technology for wide area access to a primary accessing technology for today's metropolises, although challenges remain to be addressed.

Data Collection and Description

We have closely collaborated with *Tencent*, one of the largest Internet companies in China, to conduct a nation-wide measurement. Tecent's major products, including QQ, Tencent Weibo, and WeChat, have over 1 billion active subscribers. In a one-month duration (from March 12 to Aril 12, 2015), we collected the location and ownership information of the reported wireless APs, as well as the connection records between WiFi users and those APs. The data were contributed by the users and business partners of *Tencent Mobile Manager* (http://m.qq.com/), a widely used utility software on mobile platforms with over 230 million downloads. A key functionality of the manager is to facilitate a user smartly selecting the best WiFi network among the currently available ones, and for the business partners to share networks with individual users. A considerable portion of the users also volunteer to share known WiFi accesses by providing the passwords to the manager and hence other users of the manager. To ensure secure access to the shared WiFi networks, the users are willing to report their connections to Tencent and have the traffic monitored. It is worth noting that only the related meta data is collected, so none of the users' personal information was traced during the collection, nor were the conversation data intercepted or stored.

Our collected dataset consists of two parts, the data of *8 million wireless APs* and the data of user-AP connection records from *6.4 million daily active users* (around *27 million WiFi accesses per day*). For each reported wireless AP, we have the information of the AP's geographic location (latitude and longitude), its ownership, the Internet service provider (ISP), and the medium access control (MAC) address. With the location information, we are able to analyze the deployments of wireless APs in four major cities in China. Moreover, the ownership information allows us to classify the APs into different categories (i.e., business, residential, and public). Each reported user-AP connection includes the information of the network's BSSID, the user IP (which was hashed so as not to reveal the real user IP), the server IP, the connect time, and the connect duration. By identifying and merging the records that belong to the same user, we are able to analyze the connection activities of individual users.

How Are the WiFi Access Points Deployed?

AP Distribution and Coverage

We first inspect the locations of deployed wireless APs in our dataset, and investigate the current deployments of WiFi networks. Figure 1 visualizes the deployments of wireless APs in the four major cities, Beijing, Shanghai, Guangzhou, and Shenzhen, which are representative of today's metropolises in China (with urban populations of 21, 24, 20, and 12 million, respectively). As shown in Fig. 1, well developed urban regions

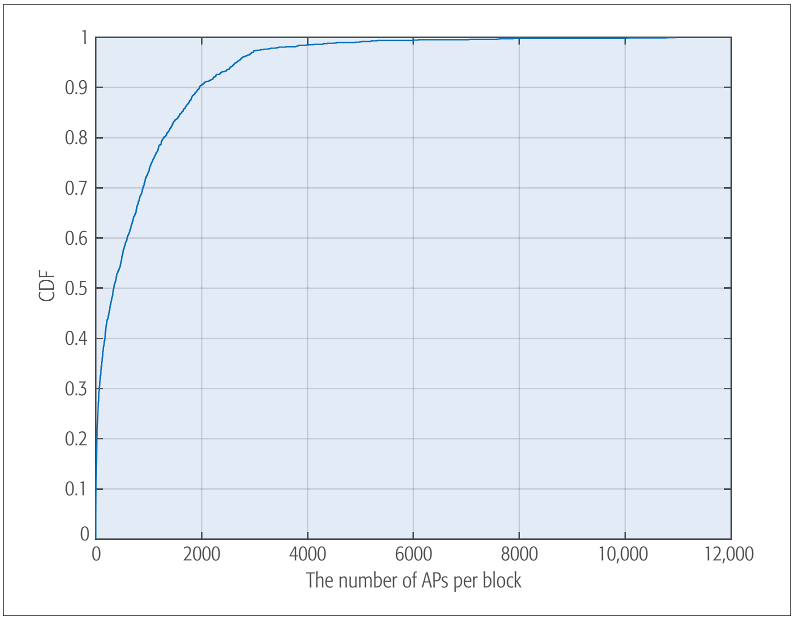


Figure 2. CDF of number of APs per block.

have much denser deployments of wireless APs than less developed suburban regions. It is not surprising that the wireless APs are deployed toward the economically well developed and highly populated regions. The deployment is also confined by natural topography: there is nearly no wireless AP deployed in water, mountains, and forests. In short, the deployments of WiFi networks in today's metropolises are driven by and naturally reflect the intensity of human social and economic activities.

To better understand the WiFi networks' distributions in different areas, we divide the latitude and longitude space into 0.0009° × 0.0009° latitude/longitude grids (approximately 100 m × 100 m). The reason that we adopt this setting is two-fold: first, dividing geographical area into blocks allows us to index locations and better study the relationship between the APs' deployments and locations; second, the APs that are located within this range are likely to have overlapping coverage, and thus the users in the same block probably have opportunities to connect to different WiFi networks, and even to establish device-to-device communications with other users. This will help in explaining and utilizing the potential social relationships between mobile users; for example, early studies discovered that people calling while connected to the same cell tower (colocation) are good proxies for face-to-face meetings [8] .

Applying the above settings, our analysis of the dataset shows that the distribution of the APs is highly skewed: a small number of blocks have deployed very large numbers of APs, while others have far smaller populations of APs. Taking Shenzhen as an example, Fig. 2 shows that about 30 percent of the blocks have over 1000 APs, and around 10 percent of the blocks have more than 2000 APs. Despite the fact that there may be high-rise buildings in developed urban areas, this result suggests that today's metropolises have good coverage of WiFi signals. In fact, many urban areas already have more than enough available APs, and further deployment should be strategically planned and optimized by examining the collective WiFi coverage and possibly removing redundant APs.

In contrast, the statistic of our dataset shows that only 3.2 percent of the blocks in Shenzhen (14.6 percent in Beijing, 9.8 percent in Shanghai, 9.5 percent in Guangzhou) have less than or equal to 1 wireless AP, which we consider insufficient WiFi coverage as few existing commercial APs are powerful enough to cover a 100 m × 100 m block area. After closely examining the blocks that have limited WiFi coverage on the map, we find that these blocks are often in regions that have limited population of residents and rare business activities (e.g., water, mountains, forests, and suburban areas). Our

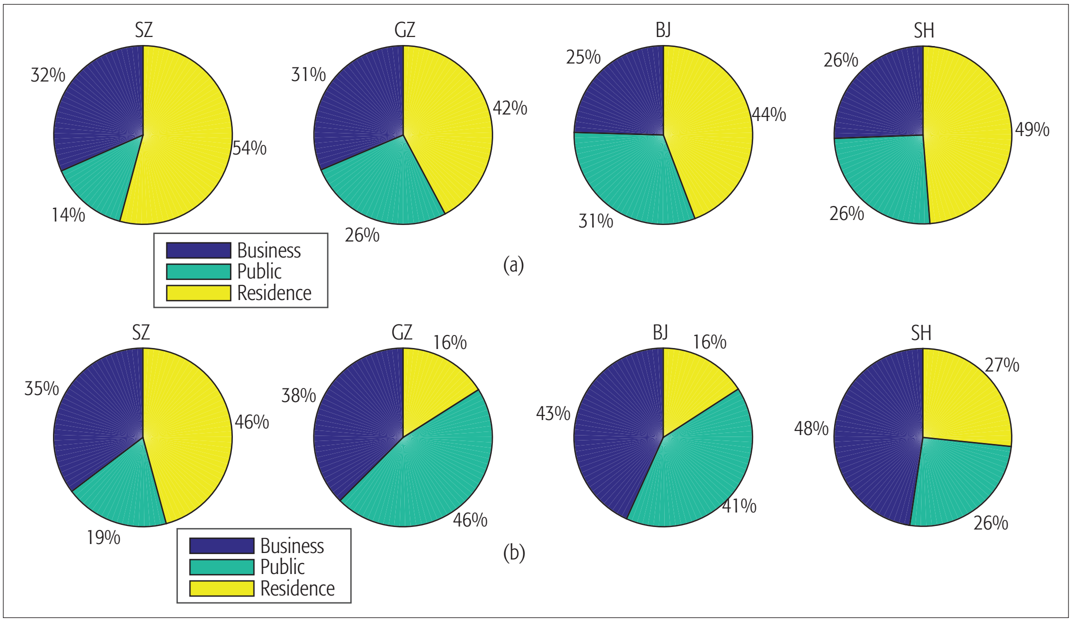


Figure 3. AP/block constitution in four metropolises (SZ stands for Shenzhen, GZ stands for Guangzhou, BJ stands for Beijing, SH stands for Shanghai): a) classification of APs in four metropolises; b) classification of blocks in four metropolises.

measurement shows that today's metropolises already have dense deployments of wireless APs, which can collectively provide near-ubiquitous Internet access. The deployments and applications of wireless mesh networks [9] have been extensively discussed in the literature, which, however, can hardly be realized on a large scale by a sole operator. As more and more wireless APs will be deployed on demand, we can foresee that in the near future wireless mesh networks may be constructed and maintained at the metropolis level with the help of mobile wireless APs that can be carried by vehicles or whose role can be fulfilled by increasingly capable personal mobile devices.

AP Categories

We next analyze the types of wireless APs. Specifically, we classify the APs into three categories according to the collected ownership information: business (e.g., shopping centers, restaurants, hotels, entertainment venues, domestic services, auto services, banks, and other companies), public (e.g., government offices, schools/universities, travel sites, public gyms, cultural venues, and other infrastructures), and residential. We further classify the blocks into these three categories based on the dominant type of wireless APs in this block. Figure 3 shows the constitutions of the APs and location blocks in the four metropolises. In all the cities, residential APs are the largest component, and public APs are the smallest one (except for Beijing, the capital of China). Existing studies on leveraging WiFi networks to offload traffic usually focus on public or business wireless APs. Our results suggest that there remain great opportunities in fully utilizing residential WiFi networks, which can provide substantial bandwidth resources especially during off-peak hours. However, private WiFi network owners may be reluctant to share their paid services from ISPs with strangers, and thus such incentives as monetary reward should be offered, or the social relationship among providers and users should be explored.

According to Fig. 3b, public blocks take up a large portion, especially in Beijing (41 percent) and Guangzhou (46 percent). It makes sense as today's metropolises are well developed in terms of the infrastructure construction. Note that universities play a huge role in modern cities' education systems, and their campus WiFi networks contribute significantly to the public WiFi network coverage. Moreover, we observe that business blocks also account for a considerable percentage in all four cities (35 percent in Shenzhen, 38 percent in Guangzhou, 43 percent in Beijing, 48 percent in Shanghai), implying that today's business operators have considered providing WiFi networks as one of their top priorities to serve customers. Another interesting observation is that residential APs take up a much larger share than residence blocks, while business APs have a much smaller share than business blocks. This can be explained by the fact that business APs usually concentrate on certain locations, and residence APs are more evenly distributed, which again indicates the great potential of exploring residential APs toward ubiquitous coverage.

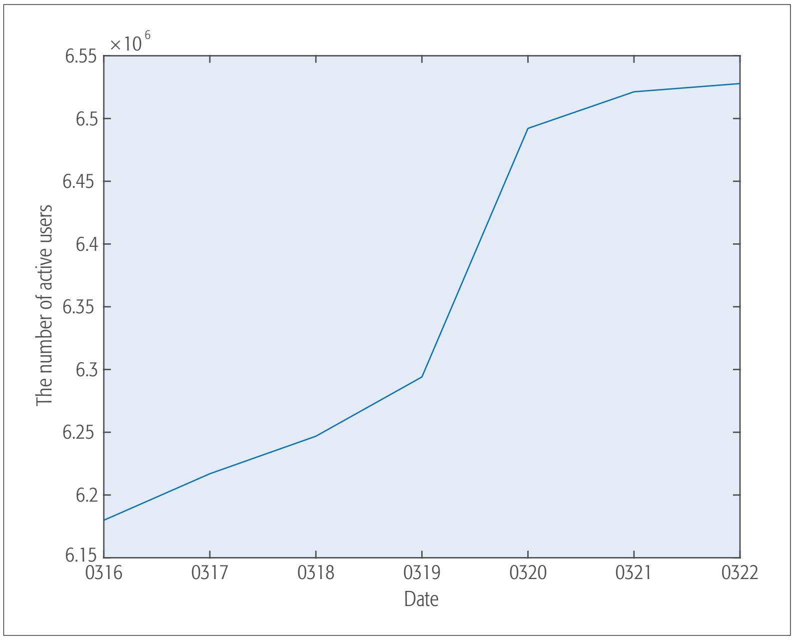


Figure 4. Number of users in each day.

How Does the Individual User Connect to WiFi?

So far we have focused on the information from the wireless APs' perspective. We now examine the data of user-AP connections to study how the users utilize the WiFi networks. Figure 4 shows the number of active WiFi users during the week from March 16 to March 22, 2015. As we can see, there are more users accessing WiFi on weekends (2015, 03, 21–2015, 03, 22) than on weekdays (2015, 03, 16–2015, 03, 20). Our later results also show that, for different scales of WiFi users, the access patterns are quite similar. We next check when each user connects to a WiFi network, and plot the number of user connections in the one-day time span in Fig. 5a. The results are based on the time when each user-AP connection is constructed. It is clear that there are three daily peaks: around 9 a.m., 1 p.m., and between 6 p.m. and 7 p.m. Intuitively, most users arrive at workplaces at 9 a.m., return from lunch breaks at 1 p.m., and get back home before or after dinner between 6 p.m. and 7 p.m. Such mobility patterns (e.g., the home-work/study-home daily routines) for the majority of WiFi users during the weekdays certainly help with developing smart city applications (e.g., smart office/home applications). Knowing the users' mobility routines (even partially) can also inspire the design of content delivery schemes using device-to-device communication with relays of wireless APs.

From the dataset, we also observe that more than 90 percent of users connect to no more than 2 wireless APs in a day. Together with the previous observation, this suggests that for most users, their WiFi accesses are not highly dynamic and may even be predictable. Revealing WiFi access patterns for mobile users can significantly enhance media content delivery. For example, with today's powerful smart routers, the providers can in-network cache the contents close to the APs from which the consumers would demand

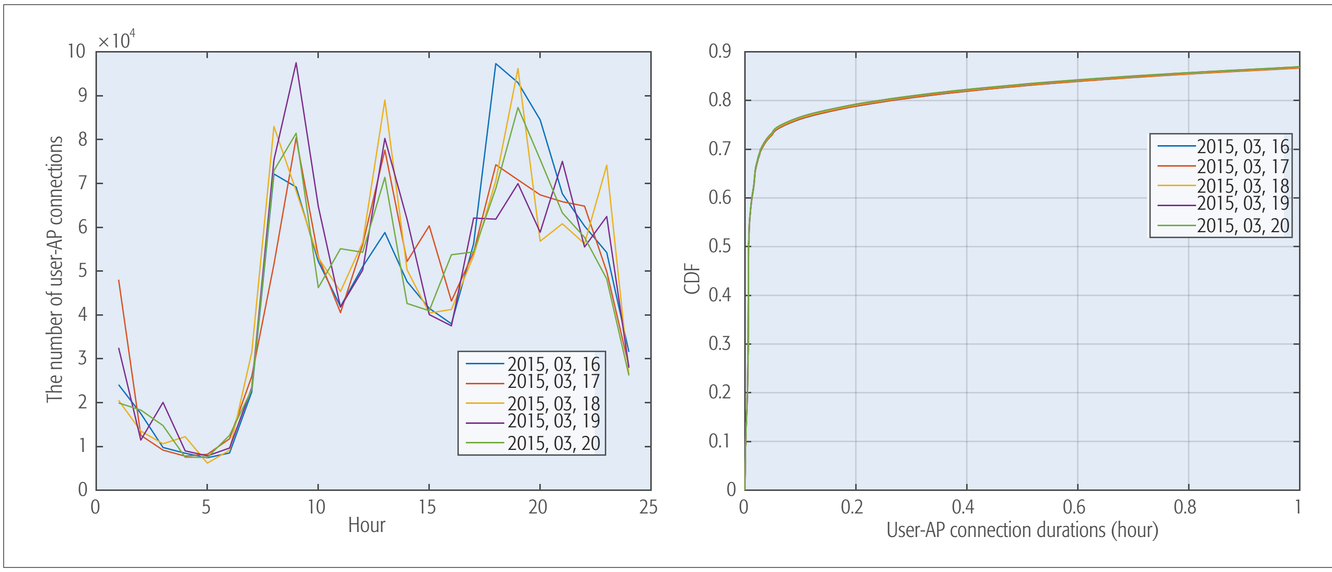


Figure 5. User-AP connection time and duration: a) number of user-AP connections in 24 hours; b) CDF of user-AP connection durations.

the media contents. On the other hand, our observation implies that the wide WiFi coverage in today's metropolises does not directly impact the users' WiFi access behaviors. If ubiquitous WiFi access becomes available in the foreseeable future, mobile users would access WiFi much more frequently and dynamically, which brings both challenges and opportunities in designing smart city applications.

We further investigate how long each time a user connects to a WiFi network during the collection period. Figure 5b shows that the distribution of the connection durations are highly consistent across weekdays. As can be seen, there are mainly two types of user-AP connections: *temporary connections* of short duration and *stable connections* of long duration (together accounting for over 82 percent of the connections), among which about 70 percent last for less than 5 minutes, while over 12 percent are longer than 1 hour. The high percentage of temporary connections implies that smart city applications are demanded to respond to user requests in a timely fashion. On the other hand, the existence of stable connections offer opportunities for long-term use of the corresponding user devices. In particular, once the user-AP connection is sustained for 10 minutes, it probably would last for much longer, in which case one can infer the user is static rather than nomadic, and turn the connected device into a crowd-sensing component in the Internet of Things to collectively contribute to complex tasks, for example, urban noise monitoring and structural health monitoring.

Revealing Implicit Social Relations

A social network is a social structure defined by actors, relationships, and other interactions between actors, which widely exists in both the human world (e.g., family ties between relatives) and the cyber world (e.g., friendships on Facebook). Upon identifying such social structures, social network analysis focuses on patterns of relationships between actors and examines the availability of resources and the exchange of resources between these actors [10]. The resources exchanged can be of many types, for example, information in a communication context. By studying the social network properties, we can understand what kind of information is exchanged, between whom, and to what extent. Today's dense WiFi network deployments in metropolises and the regular accesses, as we have seen, imply that there are certain underlying social structures among the massive number of WiFi users; if identified, they will certainly help us better utilize these WiFi networks. Unlike human or cyber social networks, mobile WiFi users do not have explicit meaningful relations. Our objective in this section is thus to investigate whether WiFi users have implicit social relations of which most of the users are not aware.

Constructing the Social Network

As mentioned, most WiFi users connect to no more than two wireless APs per day, which implies that, in practice, only a very small portion of users connect to the same wireless APs. Defining social relations between users based on sharing common WiFi connections to the same APs will result in a very sparse network that can hardly be interpretable or have practical meanings. Therefore, we assume that there is a social edge between two users if both of them access the WiFi networks in the same location block within a one-day time span. In other words, the social relations between WiFi users are defined on their geographical WiFi access patterns. Given the 100 m × 100 m block size, the WiFi users in the same block can have chances for direct communication or can easily be relayed through certain APs. Accordingly, we have analyzed the 10 million user-AP connection records in Shenzhen, and built the corresponding network of WiFi users with 67,264 nodes and 4,274,997 edges.

The constructed network consists of 245 connected components. We extract the subgraph of the largest connected component (referred to as the *WiFi user network* for the remainder of this article), which contains about 95 percent of the nodes and 99 percent of the edges in the complete graph, and present our analyses based on it hereafter. We summarize the basic statistics of the extracted subgraph in part 1 of Table 1. Note that we do observe some isolated user groups, each of which has a very limited population (from 2 to 100 users). Those isolated WiFi users may live or work in certain suburban areas, which are out of our focus here.

Small-World Network

The small-world network phenomenon is probably the most interesting characteristic of social networks. Watts and Strogatz adopted this concept to describe networks that are neither completely random nor completely regular, but possess characteristics of both [11]. They introduce a measure of one of these characteristics, the cliquishness of a typical neighborhood, as the clustering coefficient of the graph. They define a small-world graph as one in which the clustering coefficient is still large, as in regular graphs, but the measure of the average distance between nodes (the characteristic path length) is small, as in random graphs.

We compute the two small-world metrics for the WiFi user network. The result shows that the WiFi user network has definite small-world characteristics. As shown in Table 1, the average clustering coefficient is extremely high (0.91), which is very close to 1, the clustering coefficient of a regular graph. On the other hand, although the WiFi user network has a diameter of 22, the average shortest path length (the characteristic path length) is only 5.9024, and the 90th percentile diameter is 8, which nearly follows the famous *six degrees of separation* rule. The observation of the small-world network phenomenon confirms that the proposed WiFi user network is indeed a social network, which has never been identified before and is even beyond the awareness of its own members, the WiFi users.

Community Detection

We further apply the k-clique clustering algorithm to the WiFi user network. A k-clique community is the union of all cliques of size k that can be reached through adjacent (sharing $k - 1$ nodes) k-cliques. We vary the value of k from 10 to 150, and for each k present the number of clusters and the average cluster size (the number of nodes in the cluster) in part 2 of Table 1. It is clear that, whatever the value of k is, a considerable number of communities can be detected in the WiFi user network. In particular, for k = 10, 20, 50, 100, and 150, there are 107.2, 103.8, 92.1, 61.1, and 32.0 percent of the nodes clustered into different communities, respectively. It should be noted that there are cases where over 100 percent of nodes are clustered. This implies that overlapping communities are detected when k is small.

The result of k-clique clustering indicates that the WiFi user network is well structured and highly connected, which strongly suggests the existence of social communities. Researchers have discussed the construction [12] and the incentive mechanism design [13] of wireless community networks for years. However, no large-scale wireless community network with long-term impact has been built or observed. For the first time, we observe and provide the evidence of such social communities based on the geographical access patterns of WiFi users. This provides great opportunities for various aspects of network optimization. For instance, social media contents can propagate not only in online social networks but also in geographical WiFi user networks through cascaded physical communications relayed by users and APs [14]; the communities can help optimize the crowdsourcer recruitment for mobile crowdsourced sensing [15].

(1) Miscellaneous statistics	
Number of nodes	64,142
Number of edges	4,232,061
Average degree	65.98
Average clustering coefficient	0.91
Diameter (longest shortest path)	22
90-percentile diameter	8
Average shortest path length	5.9024

(2) k-clique clustering		
k	Number of clusters	Average cluster size
10	893	77
20	774	86
50	532	111
100	258	152
150	103	199

Table 1. Statistics and clustering results.

Based on our observations, we believe that WLANs are shifting from complementary access technology to primary access technology in today's metropolises and providing rich information for the underlying social relations among WiFi users.

Conclusion

In this article, we have conducted a large-scale measurement study on wireless networks in modern metropolises. Based on an extensive dataset, we have first investigated the coverage of today's WiFi networks, and by classifying wireless APs, we uncovered their rich geographical features. We further studied the access patterns of the WiFi users during weekdays and analyzed the polarization of user-AP connection durations. Finally, and most importantly, we identified the existence of geographical social networks and community structures among WiFi users. To the best of our knowledge, this is the first time such autonomous wireless communities have been observed in large-scale networks. Based on our observations, we believe that WLANs are shifting from complementary access technology to primary access technology in today's metropolises and providing rich information for the underlying social relations among WiFi users.

References

[1] K. Lee *et al.*, "Mobile Data Offloading: How Much Can Wi-Fi Deliver?" *IEEE/ACM Trans. Networking*, vol. 21, no. 2, 2013, pp. 536–50.

[2] Cisco, "Cisco Service Provider Wi-Fi: A Platform for Business Innovation and Revenue Generation," tech. tep. doc. ID 1458684054151755, 2012.

[3] A. R. Ekti *et al.*, "On the Traffic Offloading in Wi-Fi Supported Heterogeneous Wireless Networks," *J. Signal Processing Systems*, Springer, vol. 83, no. 2, 2016, pp. 225–40.

[4] M. Afanasyev *et al.*, "Usage Patterns in an Urban WiFi Network," *IEEE/ACM Trans. Networking*, vol. 18, no. 5, 2010, pp. 1359–72.

[5] X. Chen *et al.*, "Network Performance of Smart Mobile Handhelds in a University Campus WiFi Network," *Proc. 2012 ACM Internet Measurement Conf.*, 2012, pp. 315–28.

[6] J.-P. Laulajainen *et al.*, "Study of Youtube Demand Patterns in Mixed Public and Campus WiFi Network," *Proc. 2014 IEEE Int'l. Wireless Commun. Mobile Computing Conf.*, 2014, pp. 635–41.

[7] S. Biswas *et al.*, "Large-Scale Measurements of Wireless Network Behavior," *Proc. ACM SIGCOMM. ACM*, 2015, pp. 153–65.

[8] L. Wu *et al.*, "Mining Face-to-Face Interaction Networks using Sociometric Badges: Predicting Productivity in an IT Configuration Task," available at SSRN 1130251, 2008.

[9] S. Vural, D. Wei, and K. Moessner, "Survey of Experimental Evaluation Studies for Wireless Mesh Network Deployments in Urban Areas Towards Ubiquitous Internet," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 1, 2013, pp. 223–39.

[10] C. Haythornthwaite, "Social Network Analysis: An Approach and Technique for the Study of Information Exchange," *Library & Info. Science Research*, vol. 18, no. 4, 1996, pp. 323–42.

[11] D. J. Watts and S. H. Strogatz, "Collective Dynamics of Small-World Networks," *Nature*, vol. 393, no. 6684, 1998, pp. 440–42.

[12] P. A. Frangoudis, G. C. Polyzos, and V. P. Kemerlis, "Wireless Community Networks: An Alternative Approach for Nomadic Broadband Network Access," *IEEE Commun. Mag.*, vol. 49, no. 5, May 2011, pp. 206–13.

[13] M. H. Manshaei *et al.*, "On Wireless Social Community Networks,"*Proc. 27th IEEE INFOCOM*, 2008.

[14] Z. Wang *et al.*, "Propagation-Based Social-Aware Replication for Social Video Contents," *Proc. 20th ACM Int'l. Conf. Multimedia*, 2012, pp. 29–38.

[15] R. K. Ganti, F. Ye, and H. Lei, "Mobile Crowdsensing: Current State and Future Challenges," *IEEE Commun. Mag.*, vol. 49, no. 11, Nov. 2011, pp. 32–9.

Biographies

Lei Zhang [S'12] received his B.E. degree from the Advanced Class of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2011, and his M.S. degree from Simon Fraser University, Burnaby, British Columbia, Canada, in 2013. He is currently a Ph.D. student at the School of Computing Science, Simon Fraser University. He is a recipient of the C.D. Nelson Memorial Graduate Scholarship (2013). His research interests include mobile cloud computing, social media, and multimedia systems and networks.

Liting Zhao received her B.E. degree in information security in 2015 from Beijing University of Posts and Telecommunications, China. She is currently a Master's candidate in computer science at Tsinghua University. Her research interests focus on urban data mining and user behavior analysis.

Zhi Wang [S'10, M'14] received his B.E. and Ph.D. degrees in computer science in 2008 and 2014, both from Tsinghua University, Beijing, China. He is currently an assistant professor at Tsinghua University. His research areas include online social networks, mobile cloud computing, and large-scale multimedia systems. He is a recipient of the China Computer Federation Outstanding Doctoral Dissertation Award (2014), ACM Multimedia Best Paper Award (2012), and MMM Best Student Paper Award (2015).

Jiangchuan Liu [S'01, M'03, SM'08, F'17] is a professor in the School of Computing Science, Simon Fraser University, and an NSERC E.W.R. Steacie Memorial Fellow. He was an EMC-Endowed Visiting Chair Professor of Tsinghua University (2013–2016). He is a co-recipient of the inaugural Test of Time Paper Award of IEEE INFOCOM (2015), ACM SIGMM TOMCCAP Nicolas D. Georganas Best Paper Award (2013), and ACM Multimedia Best Paper Award (2012).

Vehicular Social Networks: Enabling Smart Mobility

Zhaolong Ning, Feng Xia, Noor Ullah, Xiangjie Kong, and Xiping Hu

ABSTRACT

Vehicular transportation is an essential part of modern cities. However, the ever increasing number of road accidents, traffic congestion, and other such issues become obstacles for the realization of smart cities. As the integration of the Internet of Vehicles and social networks, vehicular social networks (VSNs) are promising to solve the above-mentioned problems by enabling smart mobility in modern cities, which are likely to pave the way for sustainable development by promoting transportation efficiency. In this article, the definition of and a brief introduction to VSNs are presented first. Existing supporting communication technologies are then summarized. Furthermore, we introduce an application scenario on trajectory data-analysis-based traffic anomaly detection for VSNs. Finally, several research challenges and open issues are highlighted and discussed.

INTRODUCTION

The objective of smart cities is to improve the quality of a citizen's life. In this regard, the transportation sector is of great significance due to the rapidly increasing number of vehicles in big cities, which makes traffic management for smart cities rather challenging. As a vehicular user-centric network, the vehicular social network (VSN) deeply integrates social networks and the Internet of Vehicles (IoVs). Recently, vehicular communication networks are at a turning point, with application targets transforming from road traffic safety and transportation efficiency to VSNs, which can provide comprehensive social services for citizens. Various leading IT companies are entering this area. For instance, Apple launched the vehicle system Carplay in March 2014, by which users can participate in social activities easily and safely. Google joined in the development of VSNs by releasing Andriod Auto in June 2014.

Vehicular communication technology has evolved explosively in the past two decades. The novel design of technologies and protocols has been the focus of both industry and academia. For instance, a 75 MHz spectrum was licensed at 5.9 GHz band as dedicated short range communication (DSRC) by the U.S. Federal Communications Commission (FCC) to provide wireless communication among vehicles. IEEE then standardized the whole communication stack as wireless access for the vehicular environment (WAVE) under IEEE 802.11p, which provides support for interconnection among vehicles, and between vehicles and roadways. For the sake of studying access to social networks and social information sharing among car users, a joint research plan for the next generation of IoVs was launched in May 2014 by four top universities in the United States (Carnegie Mellon University, the University of Wisconsin, Duke University, and Boston University), which sparked high attention on VSNs in academia.

Enabling smart mobility and fulfilling high-efficiency content transmission are challenging due to the intrinsic features of VSNs. Although VSN is a brand-new communication paradigm with interest from both academia and industry, the convergence of social networks with IoVs is still in its infancy. With the objective of making VSNs practical and widely utilized, their specific research challenges and feasible solutions deserve to be studied, which motivates our work. The rest of this article is organized as follows. The concept of VSNs is introduced in the next section. Supporting communication technologies in VSNs are then summarized. Following that, a case study for VSNs is developed. Then we discuss the challenges and possible solutions in VSNs, and the final section concludes our work.

WHAT ARE VEHICULAR SOCIAL NETWORKS?

The study of IoV has sprung up, where vehicles perform as sensor hubs to capture information by in-vehicle or smartphone sensors, then publish it for consumers. The integration of intelligent sensors and communication technologies opens up an entirely new frontier for IoVs in smart cities since vehicles have changed dramatically.

FROM IOVS TO VSNS

For the sake of information dissemination and connectivity improvement, opportunistic routing has been extended into IoVs, whose applications have natural contact with social networks. As human factors are involved, this emerging paradigm is often named socially aware networking, which makes use of social relationships among device users for the construction of mobile social networks. Since social features and behaviors of

The authors provide the definition and a brief introduction of VSNs. Existing supporting communication technologies are then summarized. They introduce an application scenario on trajectory data-analysis-based traffic anomaly detection for VSNs. Finally, several research challenges and open issues are highlighted and discussed.

Zhaolong Ning, Feng Xia (corresponding author), and Xiangjie Kong are with Dalian University of Technology, and Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province; Noor Ullah is with Dalian University of Technology; Xiping Hu is with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences.

Digital Object Identifier: 10.1109/MCOM.2017.1600263

0163-6804/17/$25.00

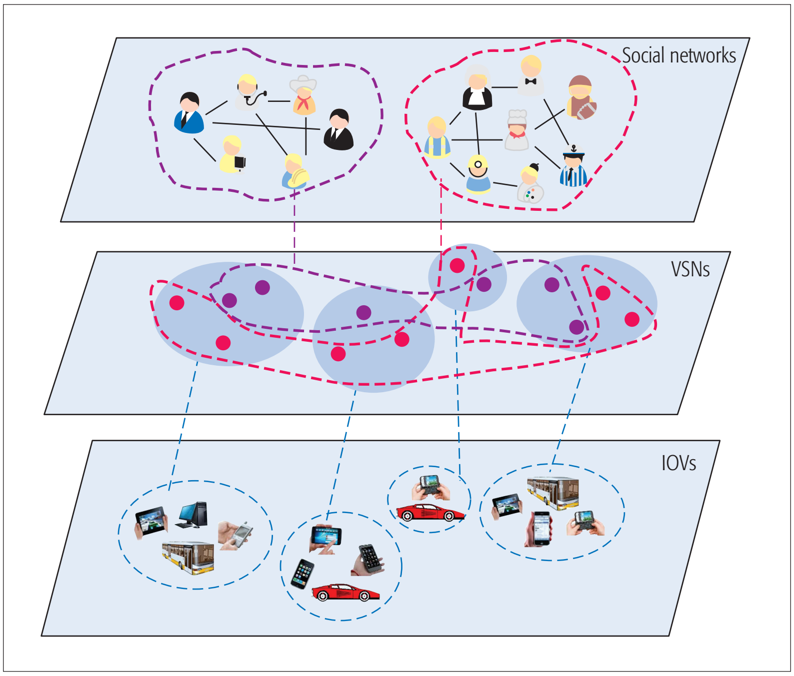


Figure 1. Schematic diagram of VSNs.

individuals are more stable than their mobility patterns, it is promising to combine socially aware networking with IoVs [1]. *We define VSNs as deep integrations of social networks and IoVs, which not only incorporate the social networks describing relationships among vehicle users, but also embrace the vehicular networks for communications among vehicle users with social relationships*. Specifically, VSNs contain not only conventional vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication frameworks, but also human factors that impact vehicular connectivity.

A social framework for vehicular communication along with an application named Roadspeak were proposed in [2]. It is called the VSN, and was initially defined as "a social network of vehicles, enabled by spatial and temporal localities on the road." This opened up a new paradigm enabling drivers and passengers to socialize on the road. VSNs can solve the problems faced by IoVs and delay-tolerant networks (DTNs) by integrating their features with social properties of the users and vehicles.

Architecture of VSNs

A schematic diagram of VSNs is illustrated in Fig. 1. The bottom level embraces the communication networks among vehicles and smart terminals, and can be viewed as a social network including physical communities. The spatiotemporal properties are critical for social relationship construction in VSNs since vehicles connect with each other when they encounter each other on the road. Meanwhile, the virtual social networks are constructed by mobile users in light of their inherent social ties on the upper level. In VSNs, the smart devices can be embedded into onboard units, roadside units (RSUs), and pedestrians. Individuals communicate with each other by exploiting social behaviors, which support various wireless communication technologies and enable communication at close proximities. Interactions among these devices/users can be divided into three types: human-to-human, human-to-machine, and machine-to-machine. The architecture of VSNs is flexible, including: centralized architecture, where a centralized server manages the system enabling V2I communication; decentralized architecture, where V2V communication is carried out opportunistically via DSRC or WAVE; and hybrid architecture, where users can be connected to the Internet via cellular networks like third/fourth generation (3G/4G) through RSUs.

Characteristics of VSNs

VSNs could be broadly leveraged in smart cities since they can obtain individuals' social relationships through network analysis, and extend users' social activities in IoVs by providing data services. The main intrinsic features in VSNs can be summarized from the social and mobile aspects.

Compared with online social networks (OSNs), the construction of VSNs is more dynamic. Social relationships in VSNs are weaker than those in OSNs, because VSNs are no longer active when users depart from the network. Besides, social connections among vehicles may be established even if they do not know each other before. Individuals in VSNs are more likely to have common interests instead of family members or friends.

Several factors influence the mobility of VSNs, including the mobility model, selfishness status, and preferences of human beings. Based on the mobility model in the community, some locations would draw high social attractiveness (e.g., restaurants, malls, and theaters) and become communication hotspots. Furthermore, the mobility model in VSNs is time-varying and has some trajectory characteristics. For example, individuals move toward the office in the morning and return home at night.

Other characteristics of VSNs include large scale, high dynamics, limited bandwidth, data heterogeneity, individual sociality, and so on. The main differences between VSNs and OSNs are summarized in Table 1.

	Network access	Social relationship	Lifetime
VSNs	Given scenarios (specific position, content and social relationship); given time (by encounter, on the fly, sporadic).	Dynamic and weak, people with common interests instead of family members or friends, even anonymous users.	Limited lifetime, no longer active when users depart from the network.
OSNs	Anytime and anywhere.	Stable and relatively strong, mostly based on social relationship and friendship.	Unlimited lifetime, once formulated, "always on."

Table 1. Main differences between VSNs and OSNs.

Applications of VSNs

The traditional applications of VSNs mainly focus on safety-related and entertainment-based applications. The former concentrates on enhancing the safety on roads to lower the probability of road accidents according to the information of a vehicle's position, speed, direction, and so on. The latter focuses on sharing and downloading of multimedia services based on a user's interest.

Recently, studies of data-driven applications are prevalent in VSNs. The periodic moving patterns offer an insightful and concise illustration, and could help to compress trajectory data and forecast the future movement of vehicles [3]. A trajectory, generated by the traveling vehicles in geographical spaces, can be gained by mobile computing techniques. Our daily life emerges as an army of GPS-equipped vehicles, including taxis, buses, and ambulances. These vehicles can report a timestamped location with a certain frequency interval, by which a great number of spatial trajectories can be generated for vehicle behavior analysis, resource allocation, and transportation efficiency improvement in VSNs. Table 2 demonstrates the taxonomy of VSN applications, which can be further divided into social-data-driven vehicular networks, social vehicular ad hoc networks (VANETs), and data-driven social networks [4].

Supporting Communication Technologies

Existing communication methods that might possibly be employed in VSNs can be broadly classified into the following four categories.

Context-Aware Transmission

It was reported by Cisco that 300 million passenger vehicles can generate over 400 million GB of data in wireless communications [5]. A large amount of data calls for effective information dissemination mechanisms, because irrelevant information is unwelcome by drivers. VSNs are promising to filter the information by social relationship recognition, interest comparison, and context-aware transmission. Therefore, filtering information on cloud and delivering relevant information by combining the emerging wireless technologies (e.g., 5G and millimeter-wave) will be essential in future vehicular networks.

Generally, the encountering probability of humans with certain social relationships is large. However, the encountering probability and duration time of vehicles may be low even with the same destination, which makes data forwarding and sharing in VSNs challenging. For the sake of fulfilling the requirements of applications and services in smart cities, the concept of social relationships among vehicles could be employed to promote the effectiveness and efficiency of information dissemination in VSNs. Wan *et al.* [6] investigated an architecture supported by mobile cloud for the vehicular cyber-physical system, and designed a context-aware dynamic parking service for smart cities as a case. Our previous work [7] presented an interest-based forwarding scheme for VSNs, which simulates food foraging behavior of bees to record information passing through different communities.

Vehicular social networks	Social data-driven vehicular networks	Traffic information
		E-advertisements
		News
		Smart calendar
		Entertainment
	Social VANETs	Cooperative driving
		Cruise control
		Navigation
		Location services
		Platooning
		P2P networking
		Content delivery
		Car-share
		Waze
		Uber
	Data-driven social networks	Theft control
		Healthcare
		Diagnostic
		Route planning
		Safety warnings
		Emergency alert
		Vehicle tracking

Table 2. Taxonomy of VSN applications.

Reliable Transmission

The moving speed of vehicles is much faster than human beings, which causes topology changes frequently in VSNs. Data transmission should accommodate highly dynamic topologies. Thus, it is imperative to guarantee the real-time transmission of urgent messages, and the correctness of regular information in both dense and sparse environments. For the sake of promoting stability and guaranteeing transmission reliability in VSNs, some mechanisms have been studied for the stimulation of node selfishness, which can be generally divided into methods based on reputation, tit-for-tat, and virtual currency. A representative game incentive mechanism was presented in [8], and it can be applied in VSNs by ameliorating the traditional game theory. A distributed incentive model was designed for VSNs based on the job market signaling model [9], and the selfish nodes are encouraged to perform cooperatively by compensating a certain amount of virtual currency for packet forwarding.

Trustworthy Transmission

The benefit of sharing common interest among vehicle drivers can be brought into VSNs. However, privacy issues deserve to be considered, which require trustworthy transmission. A trust-

Although various kinds of applications can be enabled by crowdsensing, employing them in real systems still has a long way to go in VSNs. Some issues deserve to be investigated further [15], such as sensing with human participation, data redundancy and inconsistency, and crowdsourced data mining.

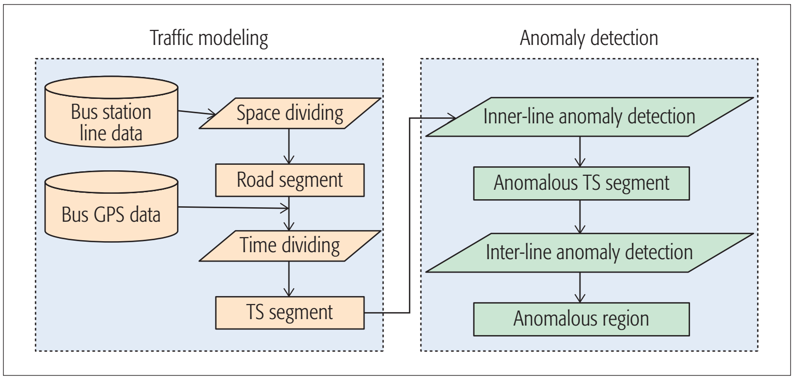


Figure 2. Framework of the trajectory data analysis-based model.

worthy information sharing scheme for VSNs was presented in [10] by evaluating direct trust from past interacting information and deducing indirect trust from social recommendations among vehicles. As an emerging application in VSNs, a dynamic trustworthy parking scheme was studied in [11], by which trusted groups of vehicles can be constructed to seek out parking spaces in the corresponding community area. A framework to construct and maintain VSNs was constructed in [12], where some trust principles were proposed to form a social group for admission and control the interactions among members.

Mobile Crowdsensing

The objective of mobile crowdsensing is to involve participants from the public to contribute sensing data from their mobile terminals in a collaborative way. Crowdsensing has drawn wide attention, due to the ever increasing use of mobile equipment. Not only can an ideal and ubiquitous platform to encourage mobile users to participate by crowdsensing be supplied in VSNs, but also clients in mobile crowdsensing can be provided by excellent assistance through employing social contexts. The effectiveness of mobile crowdsensing in VSNs has already been demonstrated, which can be leveraged for safety improvement and traffic management [13]. Since traffic prediction and congestion alleviation are important for smart cities, a mobile crowdsensing-based scheme was presented for dynamic route selection [14]. A message delivery scheme for the combination of geographical and topological information was proposed in [13]. It is adaptive for crowdsensing in vehicular circumstances, and the V2I connection for cloud services is the main consideration.

Although various kinds of applications can be enabled by crowdsensing, employing them in real systems still has a long way to go in VSNs. Some issues deserve to be investigated further [15], such as sensing with human participation, data redundancy and inconsistency, and crowdsourced data mining.

A Case Study: Trajectory Data-Analysis-Based Traffic Anomaly Detection

Some events (e.g., accidents, celebrations, sports, and disasters) cause traffic anomalies. The objective of traffic anomaly detection is to find the traffic patterns that are not expected, by which traffic problems can be discovered accurately. In this section, we propose a trajectory data-analysis-based model to detect traffic anomalies by crowdsensing. *To the best of our knowledge, this is the first work to utilize bus trajectory data to infer urban traffic conditions*. Since a bus route cannot be changed even if the bus driver encounters traffic conditions on the road, passengers can share the traffic anomaly situations, and provide information for arriving time prediction and other applications for traffic management.

Due to space limitations, we mainly focus on the utilization of results by crowdsensing. The task allocation is distributed to passengers on the bus, who can report the bus trajectory information. Task execution can be fulfilled by the GPS and sensors (mobile phones) equipped with the bus and passengers, respectively.

The framework of our method in Fig. 2 contains two parts: traffic modeling and anomaly detection. We leverage bus station line data to divide the urban road networks into road segments spatially, and extract temporal and spatial (TS)-segments from bus trajectory data according to the divided road segments. Two features are extracted from each TS-segment: average velocity and average stop time. The former models the real traffic condition on the road, while the latter represents the passenger flow volume of the bus. In the anomaly detection part, we use the inner-line anomaly detection method to find the anomalous TS-segments by calculating the local outlier factor (LOF) of TS-segments, and discover anomalous regions with an inter-line-based method.

Traffic Modeling

We first extract the TS segments from bus trajectory data to describe the traffic conditions around the city. Then we partition the urban road networks into road segments based on the bus line data. After that, we get the TS-segments to describe the real traffic situation around the city. Details are illustrated as follows.

Network Partition of the Urban Road: Bus station line data are leveraged to partition the urban road networks into road segments, which contain the information of two adjacent stations. A road segment can be identified uniquely by the bus line ID, line direction, and station number.

TS-Segment Extraction: With the results of road network partition, bus trajectory can be divided spatially. First, we search the stopping items, whose location coordinates are in the range of the line's station area. Then items with adjacent stations in one line can be obtained according to the line ID and station number, which are the results of partition from the spatial aspect.

TS-Segment Feature: Each TS-segment has lots of trajectories. The trajectory speed is the ratio between the Manhattan distance within the adjacent two stations of a TS-segment and the time the bus travels between stations.

TS-Segment Matrix Construction: We formulate a matrix for each bus line, and its item presents a TS-segment as a tuple.

Traffic Anomaly Detection

We first mine the anomalous TS-segments, which affect bus traveling by calculating the LOF of each TS-segment. Then we partition the city into small regions shown as grids on the map. With the results of inner-line detection, we map the anomalous

TS-segments in different lines to the small regions for better understanding traffic condition of the city.

Inner-Line Anomaly Detection: The average velocity and stop time of a TS-segment are leveraged to find the anomalous one by calculating the LOF in each line. LOF indicates a point's degree of outlier and illustrates the level of a point's anomaly numerically. After calculating the LOF of a TS-segment, we can find the collection of anomalous TS-segments, which largely affect the efficiency of a bus traveling in one line.

Inter-Line Anomaly Detection: After the detection of inner-line anomaly, the outlier collection of TS-segments can be obtained. Since hundreds of bus lines with connections coexist in the large city, the TS-segments need to be considered comprehensively. Furthermore, the traffic situation tends to be associated with the number of cars apparently.

We use two real datasets in Hangzhou, China: one is the bus station line dataset, and the other is the bus trajectory dataset. The bus trajectory data were collected in October 2014 and March 2015, respectively. Figures 3 and 4 depict the inferred traffic conditions (on average) in these two months, on weekdays and weekends, respectively. The background of the figures is the road network of Hangzhou. The detected anomalous regions are represented by the small grids with colors, which represent the LOFs of the regions.

It can be observed from Figs. 3a, 3b, 4a, and 4b that two main parts in Hangzhou have serious traffic problems. Deeper colors indicate worse traffic conditions. The first one covers districts like West Lake and Xiacheng, where downtown areas, Zhejiang University, and the governments of Zhejiang Province and Hangzhou City are located. The other one is around the Xiaoshan commercial city, where many office buildings, hotels, and banks are located. Figures 3c and 4c show the changes in traffic conditions from October 2014 to March 2015 on weekdays and weekends, respectively. The green and red colors represent better and worse traffic conditions, respectively. The changes might be caused by some social events or administrative traffic control (e.g., National Day and Spring Festival in China).

Figure 5 demonstrates the inferred traffic conditions of specific regions on weekdays and weekends in March 2015. Figures 5a and 5b correspond to an education region near Zhonghe Street, where more than ten schools are located. Figures 5c and 5d correspond to the scenic area around Xili Lake. These results are quite consistent with the reality, thus demonstrating the effectiveness of our method.

Challenges

A lot of challenges still exist for enabling smart mobility with VSNs. We highlight and discuss some of them below.

Message Dissemination

Delivering relevant messages intelligently to drivers in VSNs is challenging. Since no prior knowledge exists on user interest, the data will be stored on vehicles (or a server), or broadcast to interested vehicles. How to manage local data to reasonably utilize storage space and effectively disseminate messages should be investigated. Generally, a larger number of data copies accompanies a higher

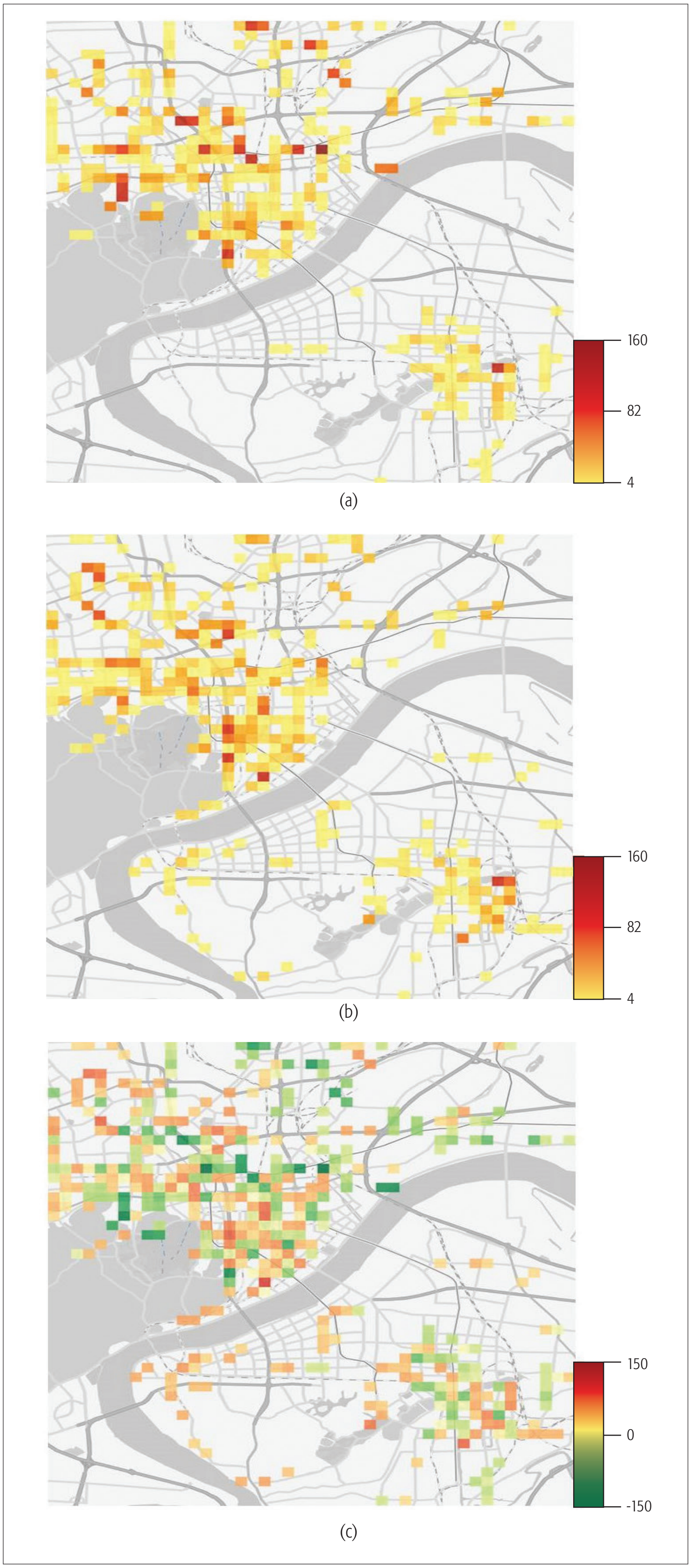


Figure 3. Inferred traffic conditions on weekdays: a) October 2014; b) March 2015; c) changes between the two months.

Figure 4. Inferred traffic conditions on weekends: a) October 2014; b) March 2015; c) changes between the two months.

success probability of transmission. A VSN-oriented multi-copy data duplication method deserves to be studied, because too many copies increase resource overhead. One promising solution is to integrate vehicular networks with content distribution networks, where the information types of vehicles are identified and classified by information relevance estimation for the decision of whether to inform the driver or share the information further.

How to address human factors to forward data under limited bandwidth is still an open issue, because user aspiration, node diversity, and selfishness variability have not yet been fully studied. Furthermore, nodes in VSNs might exhibit social selfishness, where they only cooperate when having social relationships of some sort. This phenomenon causes a lot of problems in reliable data dissemination like increased delay and loss of data packets. One common strategy is to stimulate cooperation among selfish nodes, that is, to suppress selfish node behaviors. However, node selfishness is an intrinsic feature varying with time. One possible solution is to develop selfishness-tolerant data transmission schemes by viewing individual selfishness as a fundamental social attribute in VSNs (just like mobility).

Big Trajectory Data Analysis

The ever increasing trajectory data require novel technologies to discover data from various sources. Selecting ways to mine big trajectory data and other data sources brings new challenges for researchers. Machine learning and data mining models have been leveraged to handle one single dataset. However, the technology to learn mutually reinforced knowledge from multiple datasets still needs to be studied. Knowledge fusion does not imply merely putting together an array of characteristics extracted from multiple datasets, but calls for deep understanding of each dataset and effective utilization of datasets in different parts of a big data computing framework.

The huge volume of trajectory data makes it possible to analyze the mobility patterns of traveling vehicles and their social relationships in VSNs. However, trajectories of vehicles are not perfectly accurate due to sensor noise and other reasons, for example, false positioning signals received in some urban areas [4]. Therefore, it is necessary to filter such noise points from traffic trajectories before starting a mining task. After that, map matching is needed to convert an array of raw latitude/longitude coordinates to an array of road segments.

Trust, Security, and Privacy

In order to construct VSNs for practical applications, the following four aspects need to be fulfilled: social group formation, trust management and evaluation, decentralized architecture, and data integrity. It is noted that all efforts in network optimization are nothing without trust. For trust transmission in VSNs, two main challenges can be summarized as follows. Constant change of topology caused by the movement of vehicles makes the contact time limited, and a third party is required for trust maintenance; VSNs utilize users' information like location, identity, mobility patterns, and social connections to provide services. Potential security threats for users should also be considered, since their personal information can be exposed to malicious attacks and criminal activities. Similarly, false alarms can be raised in an emergency situation, which may mislead the users or even cause mishaps on the roads.

With the objective of protecting users from privacy information leaks due to disclosing users' trajectories, the trajectory would not be manifested so certainly. On one hand, it is encouraged to obscure the user's location; on the other hand, the user's satisfaction of service should be guaranteed. Correlational studies mainly focus on two aspects: location-based services to report the traffic conditions within 1 km around the user, and publication of the historical trajectories. Furthermore, misleading information in utility services may cause waste of time and energy. So far, little related work has been done to eradicate these issues in VSNs. Therefore, social trust-based strategies can be devised in future for reliable and secure communications in VSNs.

Conclusion

In this article, we have emphasized the importance of high-efficiency and reliable transmissions in VSNs for smart cities. Particularly, we study a case of traffic anomaly detection for VSNs by trajectory data analysis. Although VSNs can be regarded as the integration of social networks and IoVs to improve the quality of life for citizens, the avenues of VSN studies are not flat, and many open issues are still ahead. We believe that VSNs will draw extensive attention and research efforts in the near future as the integration of information technology and social network services becomes more compacted.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61572106 and Grant 61502075, the China Post-Doctoral Science Foundation Project under Grant 2015M580224, and Dalian Science and Technology Planning Project 2015A11GX015, 2015R054.

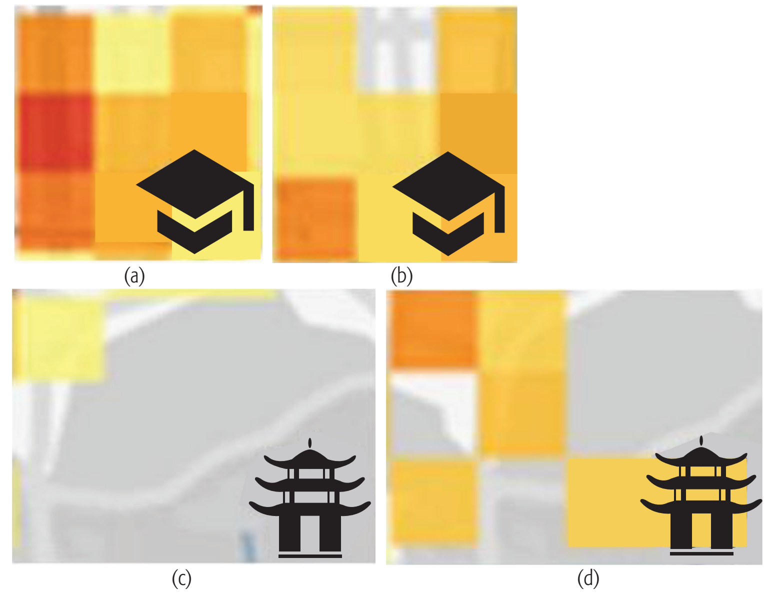


Figure 5. Inferred traffic conditions of specific regions in March 2015: a) an education region on weekdays; b) an education region on weekends; c) a scenic area on weekdays; d) a scenic area on weekends.

References

[1] A. Vegni and V. Loscr, "A Survey on Vehicular Social Networks," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 4, 2015, pp. 2397–2419.
[2] S. Smaldone *et al.*, "RoadSpeak: Enabling Voice Chat on Roadways Using Vehicular Social Networks," *Social Network Sys.*, 2008, pp. 43–48.
[3] R. Hussain *et al.*, "A Hybrid Trust Management Framework for Vehicular Social Networks," *ICCSN*, 2016, pp. 214–25.
[4] Y. Zheng, "Trajectory Data Mining: An Overview," *ACM Trans. Intelligent Sys. Tech.*, vol. 6, no. 3, 2015, pp. 1–41.
[5] Q. Yang *et al.*, "An Architecture of Cloud-Assisted Information Dissemination in Vehicular Networks," *IEEE Access*, vol. 4, 2016, pp. 2764–70.
[6] J. Wan *et al.*, Context-Aware Vehicular Cyber-Physical Systems with Cloud Support: Architecture, Challenges, and Solutions," *IEEE Commun. Mag.*, vol. 52, no. 8, Aug. 2014, pp. 106–13.
[7] F. Xia *et al.*, "BEEINFO: Interest-Based Forwarding Using Artificial Bee Colony for Socially Aware Networking," *IEEE Trans. Vehic. Tech.*, vol. 64, no. 3, 2015, pp. 1188–1200.
[8] J. Zhang *et al.*, Information Dissemination in Vehicular Networks via Evolutionary Game Theory, *IEEE ICC*, 2014, pp. 124–29.
[9] N. Haddadou *et al.*, "A Job Market Signaling Scheme for Incentive and Trust Management in Vehicular Ad Hoc Networks," *IEEE Trans. Vehic. Tech.*, vol. 64, no. 8, 2015, pp. 3657–74.
[10] Q. Yang and H. Wang, "Toward Trustworthy Vehicular Social Networks," *IEEE Commun. Mag.*, vol. 53, no. 8, Aug. 2015, pp. 42–47.
[11] J. Timpner *et al.*, Trustworthy Parking Communities: Helping Your Neighbor to Find a Space, *IEEE Trans. Dependable Secure Computing*, vol. 13, no. 1, 2016, pp. 120–32.
[12] N. Abbani *et al*, "Managing Social Networks in Vehicular Networks Using Trust Rules," *IEEE ISWTA*, 2011, pp. 168–73.
[13] M. Bguena *et al.*, "An Adaptive Anycasting Solution for Crowd Sensing in Vehicular Environments," *IEEE Trans. Industrial Electronics*, vol. 62, no. 12, 2015, pp. 7911–19.
[14] J. Wan *et al.*, "Mobile Crowd Sensing for Traffic Prediction in Internet of Vehicles," *Sensors*, vol. 16, no. 1, 2016, pp. 88–102.
[15] B. Guo *et al.*, "Mobile Crowd Sensing and Computing: The Review of an Emerging Human-Powered Sensing Paradigm," *ACM Computing Surveys*, vol. 48, no. 1, 2015, DOI: 10.1145/2794400.

Biographies

Zhaolong Ning [M'14] received M.S. and Ph.D. degrees from Northeastern University, Shenyang, China. He was a research fellow with Kyushu University, Fukuoka, Japan. He is an assistant professor with the School of Software, Dalian University of Technology, China. He has published over 50 scientific papers in international journals and conferences. His research interests include social computing, network optimization, and big data.

Feng Xia [M'07SM'12] received B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is a full professor with the School of Software, Dalian University of Technology. He has published two books and over 200 scientific papers in international journals and conferences. His research interests include computational social science, big data, and mobile social networks. He is the corresponding author.

Noor Ullah received B.Sc. and M.Sc. degrees in electrical engineering majoring in communication and electronics from the University of Engineering and Technology Peshawar, Pakistan. He is currently a Ph.D. candidate in software engineering at the School of Software, Dalian University of Technology. His research interests include vehicular social networks and intelligent transportation system design.

Xiangjie Kong received a Ph.D. degree from Zhejiang University, Hangzhou, China. Currently, he is an associate professor in the School of Software, Dalian University of Technology. He has published over 30 scientific papers in international journals and conferences. His research interests include big data, social computing, and mobile computing.

Xiping Hu received a Ph.D. degree from the University of British Columbia, Vancouver, Canada. He is currently a professor with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He was the co-founder and CTO of Bravolol Limited in Hong Kong. His research areas are distributed software systems, mobile social networks, service/cloud computing, and crowdsourcing.

Toward Eco-Friendly Smart Mobile Devices for Smart Cities

Abdellaziz Walid, Abdellatif Kobbane, Jalel Ben-Othman, and Mohammed El Koutbi

The authors study the energy efficiency of mobile users for better battery life management of their smart mobile devices through minimal energy consumption during their uplink transmissions. The proposed solution, based on a fully decentralized algorithm for sharing time access executed by SBSs and multi-homing capabilities of macrocellular users, improves the energy efficiency of mobile users for better battery life, and keeps their QoS throughput requirements.

Abstract

It is expected that the incessant use of SBSs by indoor mobile users to achieve high data rate and high energy efficiency for their smart mobile devices will drive ultra dense deployment of SCNs within MCNs, specifically in future smart cities. However, SCN densification has a hard effect on the deterioration of the energy efficiency and QoS throughput of mobile users equipped with smart mobile devices like smartphones, laptops, PDAs, and tablets. This degradation is due to the severe cross-tier interferences occurring in ultra dense co-channel deployments of SCNs within MCNs. The present article intends to study the energy efficiency of mobile users for better battery life management of their smart mobile devices through minimal energy consumption during their uplink transmissions. The proposed solution, based on a fully decentralized algorithm for sharing time access executed by SBSs and multi-homing capabilities of macrocellular users, improves the energy efficiency of mobile users for better battery life, and keeps their QoS throughput requirements. Therefore, it leads to the eco-friendly smart mobile devices required in future smart cities. The results validate our proposed solution and show the improvement achieved compared to the conventional access control mechanisms. Furthermore, our scheme requires no coordination among SBSs and macro base stations, which reduces the huge signaling overhead when using cooperative schemes.

Introduction

Future smart cities are designed to meet citizens' ever increasing requirements efficiently and in an enviromently sound manner [1]. Currently, smart mobile devices such as smartphones, using high data rate and high energy consumption applications like streaming contents and network gaming, are growing significantly. These demanding applications lead to battery drain for smart mobile devices at a faster rate. Consequently, future smart mobile devices will require a high data rate with low energy consumption for satisfying their quality of service (QoS) requirements, improving their battery life and being environmentally friendly. Nonetheless, the 4G cellular networks are still incapable of satisfying these requirements for smart mobile devices in smart cities.

According to the standards of LTE and the requirements of future smart cities, future 5G cellular networks are expected to be ultra dense small cell networks (SCNs), composed of a mixture of dense indoor SCNs such as pico and femtocells installed practically everywhere within macrocell networks (MCNs). SCNs enhance spectrum reuse and coverage while providing high data rate services and seamless connectivity. Meanwhile, SCNs are also considered as a green solution for network operators and smart mobile devices in future smart cities.

The balance of this article is organized as follows. We discuss the SCNs densification in smart cities in the next section. Then we present the advantages and drawbacks of co-channel deployments and our solution. Following that, we present the numerical results. Finally, we make concluding remarks.

Toward SCN Densification in Future Smart Cities

The advantages of SCNs, which appear in ensuring higher data rates with low-power transmission to the users of SCNs as well as offloading from the MCNs, will lead to the spread of SCNs within MCNs, especially in future smart cities. As a consequence, a vast number of neighboring SCNs with overlapping zones will show up (Fig. 1). The extent of the overlapping zones depends on the closeness of neighboring SCNs to each other.

Co-Channel SCN Deployments

Unlike orthogonal channel deployments of SCNs, co-channel SCN deployments share the same frequency band between MCNs and SCNs, thus resulting in high spectral efficiency over the use of self-optimization techniques. However, this happens at the cost of degrading the energy efficiency and QoS requirements of the cellular users due to the cross-tier interferences. Moreover, overlapping zones of neighboring SCNs have a crucial impact on the deterioration of the energy efficiency of cellular users due to the gravity of cross-tier interferences, especifically in ultra dense co-channel SCN deployments. Hence, interference management in heterogeneous networks with SCNs has received more attention from the industrial and academic research communities. Self-optimization techniques such as power control, frequency, and time allocation used by small cell base stations (SBSs) are good methods that need to be implemented based on sensing methods for avoiding cross-tier interferences. On the

Digital Object Identifier: 10.1109/MCOM.2017.1600271

Abdellaziz Walid, Abdellatif Kobbane, and Mohammed El Koutbi are with Mohammed V University of Rabat; Jalel Ben-Othman is with the University of Paris 13.

Figure 1. Ultra dense SCNs deployments in future smart cities.

The energy consumed by wireless networks and mobile devices was ignored in the traditional deployments of wireless cellular networks. This neglect leads to a significant waste of the electricity, poor battery autonomy of mobile devices, and increased CO_2 emissions. Hence, energy efficiency arises as a strongly property of future wireless networks and mobile devices designed for smart cities.

other side, the access control policy to the SCNs also has a major role in defining the impact of interferences on the overall system performance. Two access control methods for SCNs have been proposed:

- Closed access: Only the owner of the SCN and its list of authorized users have the right to use the SCN.
- Open access: Everyone have the right to use the SCN.

Energy Efficiency of Mobile Devices

The energy consumed by wireless networks and mobile devices was ignored in the traditional deployments of wireless cellular networks. This neglect leads to a significant waste of electricity, poor battery autonomy of mobile devices, and increased CO_2 emissions. Hence, energy efficiency [2, 3] arises as a strongly property of future wireless networks and mobile devices designed for smart cities. In addition, current applications of mobile devices increase communication power consumption due to their high data rate requirements, which consequently raises power cost [4]. Therefore, energy efficiency is taking more attention; also, spectrum efficiency and green wireless networks are widely studied nowadays.

Accordingly, the energy aspect needs to be considered in heterogeneous networks with SCNs. In [5], cognitive radio and power control optimization are studied for SCNs. The problem is modeled as a Stackelberg game to obtain the Stackelberg equilibrium solution. Similarly, power control optimization for the uplink communications of two-tier heterogeneous networks is investigated in [6], using fictitious games. Joint energy, spectrum utilization, and outage capacity are addressed in [7]. In [8], the problem of power optimization is formulated via game theory as a non-cooperative game, in which each macro base station (MBS) maximizes its gain based on circuit and transmission power, while each SBS tends to maximizes its utility based on signal-to-interferemce-plus-noise ratio (SINR). Energy efficiency of uplink transmissions of mobile devices is investigated with QoS requirements and multi-homing capabilities [9].

In brief, a plethora of studies on energy-efficient uplink communication for heterogeneous networks (HetNets) are investigated considering only closed or open access control mechanisms. On the other hand, all scenarios of their simulations are simple, composed of a few SCNs within MCNs and assuming no overlapping zones among SCNs. However, in reality, SBSs are deployed in a random way to satisfy the high data rate requirements of mobile users in indoor environments (house, enterprise, etc.). In this case, in dense urban zones a random ultra dense co-channel deployment of SCNs will show up with numerous overlapping zones where there are many macro indoor users that aggravate the effect of cross-tier interferences in the closed access policy. Furthermore, in the open access policy, macrocellular users located inside overlapping zones of neighboring SCNs will face complications in making their decisions in such a way that each macro indoor user will have the right to select a single SBS among all neighboring SBSs involved in the overlapping zones. As a result, each cellular user located inside an overlapping zone of wireless networks will choose the best one among all available wireless networks. Received signal strength (RSS) [10], available bandwidth [11], and maximum SINR are criteria for choosing the best available wireless network. One disadvantage is that

Step 1 : Each SBS_i sets up the closed access as a default access policy $\phi_i^* = 1$;
Step 2 : Each SBS_i sensing periodically for macro users in the area of its coverage;
Step 3 : If they exist in the area of a SBS_i, it gets some parameters n_i, $n_{s,i}$, Ω_h^i, Ω_m^i;
Step 4 : The SBS_i gets the total throughput of homes users in the uplink $T_{total}^{h,i}$, and the total throughput of macro indoor users in the uplink $T_{total}^{min,i}$ received by SBS_i;
Step 5 : The SBS_i computes the fraction of the shared time access ϕ_i and $(1 - \phi_i)$ that will be allocating to home and macro indoor users respectively;
Step 6 : If the fraction is feasible $(0 < \phi_i < 1)$, it will represent the solution $\phi_i^* = \phi_i$, else the SBS_i will switch to the open access policy.

Algorithm 1. The fully decentralized algorithm.

all resources from different neighboring SCNs are not fully exploited with the single network selection mode. Another drawback is the call blocking that occurs when no resources are available for cellular users. Furthermore, a congested network will occur if many macro indoor users choose the best SBS among all neighboring SBSs at the same time [12]. On the contrary, with multi-homing capabilities, a macrocellular user located in the overlapping zone of neighboring SCNs can maintain multiple simultaneous connections with its neighboring SBSs rather than using a single SBS.

In this article, we consider ultra dense co-channel deployments of indoor SCNs within an MCN, analyzing and evaluating the energy efficiency performance of smart mobile devices in smart cities.

The major contributions of this article can be outlined as follows:

- Our research takes into consideration the study of the energy efficiency of cellular users in ultra dense co-channel deployments of SCNs in future smart cities.
- A solution based on a fully decentralized algorithm and multi-homing capabilities of macrocellular users has been proposed, to improve the energy efficiency and to satisfy the QoS requirements of mobile users.
- Our scheme requires no coordination among SBSs and the MBS, which reduces the excessive overhead over the backhaul links [13] induced using cooperative schemes [14].

Energy Efficiency Analysis

Small Cell Network Densification Model

In our model, we consider a macrocell M with radius R_M centered at an MBS and a variable number N_s of SBSs (called femto BSs) distributed randomly within the area covered by macrocell M. N_m macrocellular users are uniformly distributed within the coverage area of macrocell M, and $n_{s,i}$ cellular users (called home cellular users) are randomly distributed within the area of their SBS_i. The macro users and home users use smart mobile devices with fixed transmit powers of p_m and p_h, respectively. We focus our study on LTE technology, and we consider the uplink transmission that uses single-carrier frequency-division multiple access (SC-FDMA). We consider a co-channel deployment where all SCNs and the MCN use the same band frequency. We take into account the cross-tier interferences between each SCN and MCN, co-tier interferences between neighboring SCNs, and path loss and propagation models. We divide all cellular users into three categories: macro indoor users, macro outdoor users, and home users or small BS owners.

Parameter	Value
Macrocell radius (R_m)	1000 m
Small cell radius (R_s)	75 m
Number of macrocellular users (N_m)	500
Number of home users	1–2
Frequency	2 GHz
Macro UE power p_m	0.2 W
Home UE power p_h	1 mW
Indoor walls loss (L_{iw})	5 dB
Outdoor walls loss (L_{ow})	20 dB
Bandwidth of downlink B_w	20 MHz
N_{sc}	12
N_{symbol}	4
N_{rb}	100
Modulation scheme	64-QAM
Subcarrier spacing (Δf)	15 kHz
Bit error rate (BER)	10^{-6}
White noise power density (N_0)	–174 dBm/Hz
Home user QoS requirements Ω_h^i	1 Mb/s
Macro user QoS requirements Ω_m^i	0.1 Mb/s
Number of homogenous interfaces of macrocellular users	4

Table 1. Simulation parameters.

Energy Efficiency Model

For smart mobile devices, energy efficiency is expressed as a measure of the maximum amount of bits that can be delivered per joule of energy consumed. In order to model the system as a set of distinct mobile devices, the definition in [15] represents the total energy efficiency expressed as the sum of energy efficiency for individual mobile devices.

Our Solution : Shared Time and Multi-Homing Access Policy

In this section, we focus on how to maximize the energy efficiency of smart mobile devices by adopting our solution which is based on a fully decentralized algorithm for sharing time access executed by SBSs and multi-homing capabilities of macrocellular users.

Distributed Shared Time Access Algorithm

We consider the shared time access where an SBS_i allocates ϕ_i fraction of time slots to home users and the remaining $(1 - \phi_i)$ fraction of time slots to macro indoor users in the uplink. Each SBS tries to optimize ϕ_i in order to maximize selfishly the total energy efficiency of users located in its area and to satisfy their QoS requirements.

In the shared time access we have

$\phi_i \in [0,1[$ and $p_{m_{in},k} = p_{h,k}$

where $p_{m_{in},k}$ is the transmit power of a macro indoor user on subcarrier k, and $p_{h,k}$ is the transmit power of a home user on subcarrier k. The total energy efficiency of mobile devices in an SBS_i is given as

$$EE_{SBS_i} = \frac{1}{p_{h,k}}\left(\phi_i T_{total}^{h,i} + (1-\phi_i) T_{total}^{m_{in},i}\right) \quad (1)$$

where $T_{total}^{h,i}$, $T_{total}^{min,i}$ are the total uplink throughput of home users and the total uplink throughput of macro indoor users in $SBS_{i,}$ respectively

Maximize EE_{SBS_i}, *subject to*

$$\frac{\phi_i T_{total}^{h,i}}{n_{s,i}} \ge \Omega_h^i \text{ and } \frac{(1-\phi_i) T_{total}^{m_{in},i}}{n_i} \ge \Omega_m^i$$

Ω_h^i represents the average QoS uplink throughput requirement of a home user h.

Ω_m^i represents the average QoS uplink throughput requirement of a macro user m.

The optimal value ϕ_i^* of the time allocation of a SBS_i is given as

if $T_{total}^{h,i} \ge T_{total}^{min,i}$

$$S_1 = \phi_i^* = 1 - \frac{\Omega_m^i n_i}{T_{total}^{m_{in},i}} \quad (2)$$

This solution S_1 is feasible if

$$\phi_i^* \ge \frac{\Omega_h^i n_{s,i}}{T_{total}^{h,i}}$$

if $T_{total}^{h,i} < T_{total}^{min,i}$

$$S_2 = \phi_i^* = \frac{\Omega_h^i n_{s,i}}{T_{total}^{h,i}} \quad (3)$$

Accordingly, this fully decentralized algorithm (Algorithm 1) presents a low-complexity and low-latency implementation, which motivates its deployment in the SBSs due to their low resources in terms of memory, CPU, and so on.

Exploiting Multi-Homing Capabilities of Macrocellular Users

With multi-homing capabilities of macrocellular users located in overlapping zones of SCNs, macrocellular users can maintain multiple simultaneous uplink associations with different SCNs through their multiple homogeneous radio interfaces. Hence, in multi-homing radio resource allocation, the macrocellular user sends its uplink traffic to all available SBSs enabling shared time access by exploiting its multi-homing capabilities. This has the advantage of supporting applications with high required data uplink rate through aggregating the offered resources to different SCNs. Then the achieved energy efficiency of a macro indoor user EE_{min}^{total} with multi-homing capabilities, located in an overlapping zone of a set of SCNs, is the sum of its energy efficiencies in all SCNs involved in the overlapping zone.

Moreover, the multi-homing capabilities of macrocellular users give our solution the advantage of preventing frequent occurrence of handoff that leads to the deterioration of service quality, especially with single network selection solutions. However, our solution presents a drawback due to the technical limitations in deploying many homogeneous LTE interfaces in mobile devices for the support of multi-homing in HetNets.

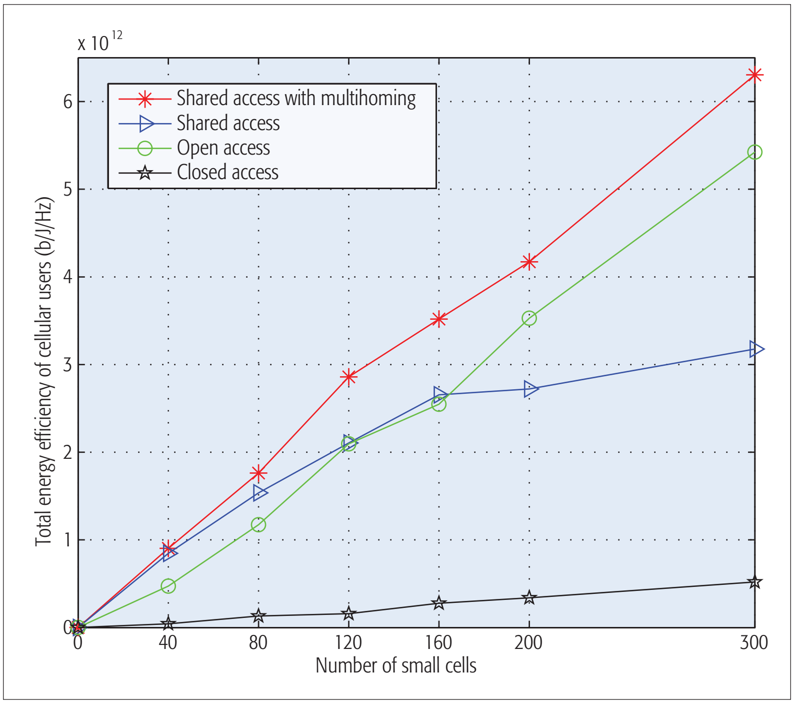


Figure 2. Total energy efficiency of cellular users.

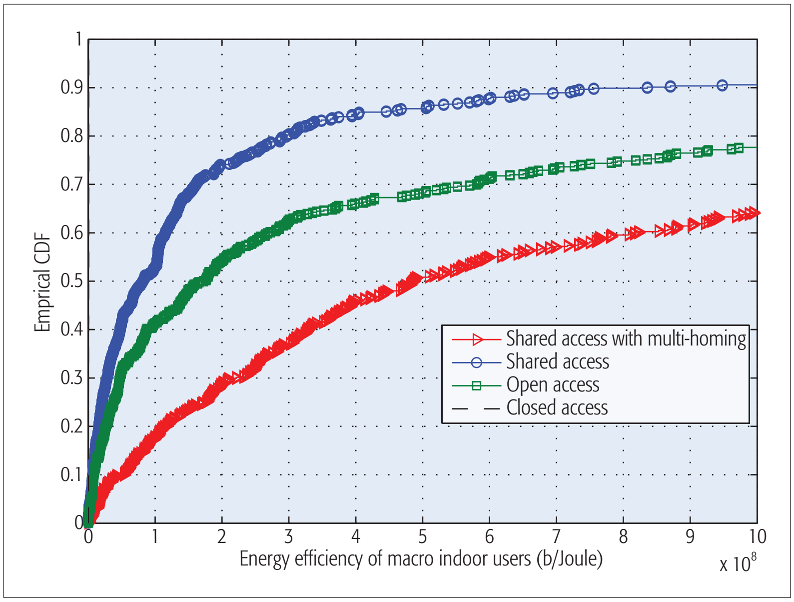


Figure 3. Empirical CDF of energy efficiency of macro indoor users.

Simulation Results

The simulations are event-based and developed according to Third Generation Partnership Project (3GPP) standards. The simulations' scenario is given according to our SCN densification model. The simulation parameters are given in Table 1. The plotted values are an average of 10,000 independent simulations using Matlab.

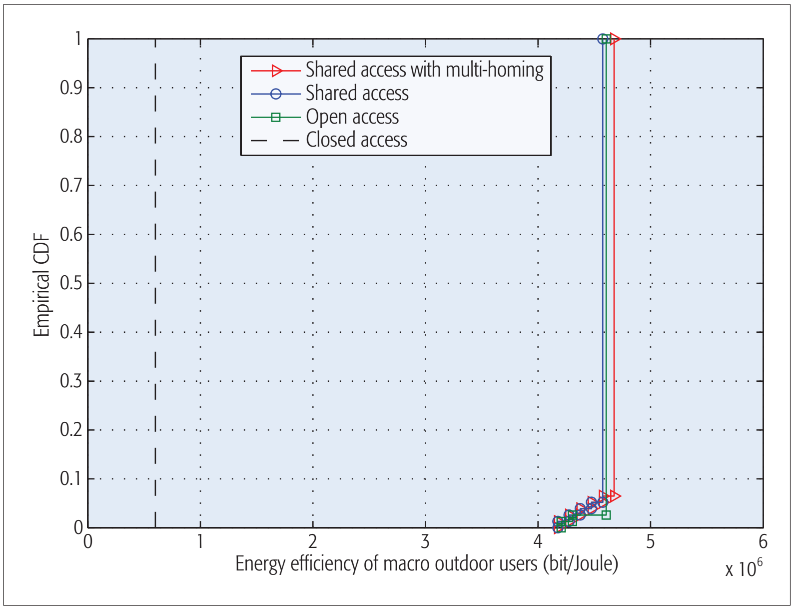


Figure 4. Empirical CDF of energy efficiency of macro outdoor users.

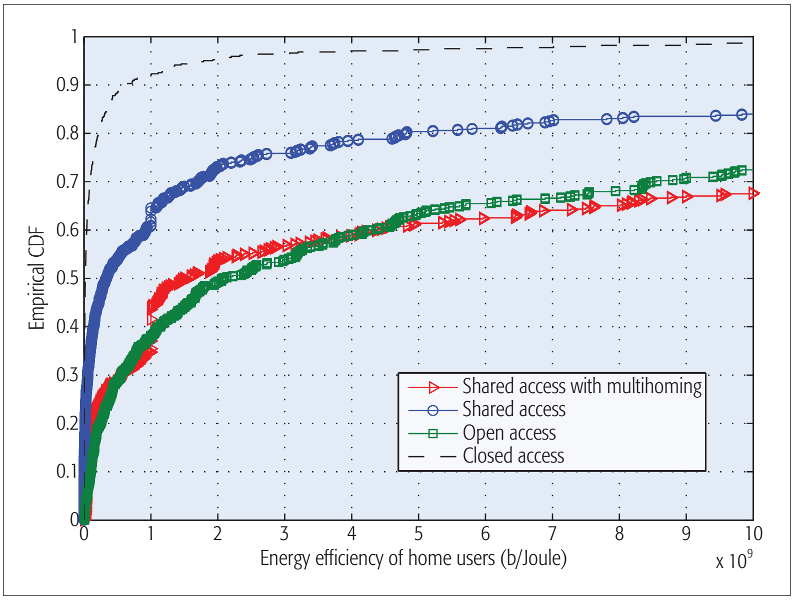


Figure 5. Empirical CDF of energy efficiency of home users.

Impact on System Energy Efficiency

First, we study the impact of our solution on the overall energy efficiency by increasing the number of SCNs within an MCN.

As depicted in Fig. 2, the overall energy efficiency in the uplink reaches its high level when the shared access policy is combined with the multi-homing capabilities of macrocellular users. Shared access with multi-homing mitigates the cross-tier interferences caused by macro indoor users to SBSs in closed mode and offloads from the macrocell. Furthermore, it keeps the QoS parameters of both home and macro indoor users degraded in the open access mode. Moreover, multi-homing improves the energy efficiency with the increasing of SCNs compared to the shared time access with single access mode.

Energy Efficiency Cumulative Distribution Function

Second, we study the impact of our solution on the empirical cumulative distribution of macro indoor users, macro outdoor users, and home users. The simulated network consists of 300 SBSs distributed randomly within the area covered by MCN.

Figure 3 shows the empirical CDF energy efficiency of macro indoor users. Shared time access mode with multi-homing is the most suitable because it mitigates the interferences caused by macro indoor users to SBSs in the closed mode, at the same time avoiding the degradation of QoS parameters of home and macro indoor users in the open access mode. What is more, the multi-homing capabilities of macro indoor users in overlapping zones of SCNs improve the energy efficiency of macro indoor users with the densification of SCNs compared to shared access without multi-homing capabilities of macrocellular users because the greater the number of SCNs, the greater the number of macro indoor users are in overlapping zones.

As depicted in Fig. 4, open access, shared access, and shared access with multi-homing are suitable modes for macro outdoor users because they offload from the macrocell by decreasing the number of macro users using the service of the macrocell, hence increasing the data rate capacity of macro outdoor users. However, this will be at the expense of QoS requirement of home and macro indoor users, which will be degraded in the open access and shared access modes.

As shown in Fig. 5, the preferred access policy for home users is shared access with multi-homing compared to open, closed, and shared access policies. This is because it reduces interferences caused by macro indoor users to home users in the closed mode and the degradation of QoS requirements of home and macro indoor users in the open mode. Moreover, it gives better energy efficiency for home users compared to the shared access mode thanks to the multi-homing capabilities of macrocellular users, which mitigate more interferences in overlapping zones compared to shared access without the multi-homing capabilities of macrocellular users.

Conclusion

In this article, ultra dense co-channel deployment scenarios of indoor SCNs within MCNs located in smart cities are studied. To deal with the challenges from random deployments, cross-tier interferences and huge overhead induced by SBSs and MBSs using coordination and cooperative schemes, a solution based on a fully decentralized algorithm for sharing time access executed by SBSs and multi-homing capabilities of macrocellular users is proposed. Our solution aims to improve the energy efficiency of cellular users for better battery lifetime of their smart mobile devices while meeting their throughput QoS requirements. In fact, numerical results display the effectiveness of our solution compared to closed, open, and shared time access mechanisms in the context of energy efficiency and QoS throughput requirements. It is noteworthy that future studies can combine our proposal with power control methods and sophisticated multi-homing algorithms to achieve far better results for future smart green cities.

References

[1] M. Batty *et al.*, "Smart Cities of the Future," *Euro. Phys. J. Special Topics*, vol. 214, 2012, pp. 481–518.

[2] G. Li *et al*., Energy-Efficient Wireless Communications: Tutorial, Survey, and Open Issues, *IEEE Wireless Commun*., vol. 18, no. 6, Dec. 2011, pp. 28–35.

[3] B. Hamdaoui, T. Alshammari, and M. Guizani, "Exploiting 4G Mobile User Cooperation for Energy Conservation: Challenges and Opportunities," *IEEE Wireless Commun*., vol. 20, no. 5, Oct. 2013, pp. 62–67.

[4] G. Miao *et al*., "Distributed Interference Aware Energy-Efficient Power Optimization," *IEEE Trans. Wireless Commun*., vol. 10, no. 4, 2012, pp. 1323–33.

[5] R. Xie, F.R. Yu, and H. Ji, "Spectrum Sharing and Resource Allocation for Energy-Effcient Heterogeneous Cognitive Radio Networks with Femtocells," *IEEE ICC*, 2012, pp. 1661–65.

[6] T. Su *et al*., "Energy Efficient Power Optimization for Two-Tier Femtocell Networks using Fictitious Game," *IEEE GLOBECOM*, 2012, pp. 3478–83.

[7] Y. Ko and K. Moessner, "Maximum Outage Capacity in Dense Indoor Femtocell Networks with Joint Energy and Spectrum Utilization," *IEEE Trans. Wireless Commun*., vol. 11, no. 12, 2012, pp. 4416–25.

[8] W. Li et al., Energy Efficient Power Allocation with Dual-Utility in Two-Tier Ofdma Femtocell Networks, *IEEE GC Wksps*., 2012, pp. 535–40.

[9] M. Ismail et al., "Uplink Decentralized Joint Bandwidth and Power Allocation for Energy-Efficient Operation in a Heterogeneous Wireless Medium," *IEEE Trans. Commun*., vol. 63, no. 4, 2015, pp. 1483–95.

[10] S. Mohanty and I. F. Akylidiz, "A Cross-Layer (Layer 2 + 3) Handoff Management Protocol for Next Generation Wireless Systems," *IEEE Trans. Mobile Computing*, vol. 5, no. 10, 2006, pp. 1347–60.

[11] W. Shen and Q. Zeng, "Resource Management Schemes for Multiple Traffic in Integrated Heterogeneous Wireless and Mobile Networks," *Proc. ICCCN*, 2008, pp. 105–10.

[12] A. Walid *et al*., "A Decentralized Network Selection Algorithm for Group Vertical Handover in Heterogeneous Networks," *Proc. IEEE WCNC*, Istanbul, Turkey, 2014, pp. 2817–21.

[13] X. Ge *et al*., "5G Wireless Backhaul Networks: Challenges and Research Advances," *IEEE Network*, vol. 28, no. 6, Nov. 2014, pp. 6–11.

[14] Ch. Yang *et al*., "Joint Power Coordination for Spectral-and-Energy Efficiency in Heterogeneous Small Cell Networks: A Bargaining Game-Theoretic Perspective," *IEEE Trans. Wireless Commun*., vol. 15, no. 2, 2016, pp. 1364–76.

[15] Y. Rui *et al*., "Mode Selection and Power Optimization for Energy Efficiency in Uplink Virtual MIMO Systems," *IEEE JSAC*, vol. 31, no. 5, 2013, pp. 926–36.

Biographies

Abdellaziz Walid graduated from ENSIAS (the public school of engineering) in the field of computer science in 2012, and has an M.S. degree in networks, telecommunications, and multimedia from Mohammed V University of Rabat, Morocco (2011). He is currently a Ph.D. candidate at ENSIAS under the supervision of Prof. Abdellatif Kobbane. His research activities focus on heterogeneous networks. He has been a computer teacher at a secondary school of the Ministry of National Education since 2005.

Abdellatif Kobbane [SM] is currently an associate professor at ENSIAS, Mohammed V University of Rabat, since 2009. He received his Ph.D. degree in computer science from Mohammed V-Agdal University, Morocco, and the University of Avignon, France, in September 2008. He is the Founder and General Co-Chair of the International Conference on Wireless Networks and Mobile Communications (WINCOM'15), October 20–23, 2015, Marrakech, Morocco.

Jalel Ben-Othman received his Ph.D. degree from the University of Versailles, France, in 1998. He is a full professor at the University of Paris 13. His research interests are in the areas of wireless ad hoc and sensor networks, broadband wireless networks, the Internet of Things, and security in wireless networks in general and wireless sensor and ad hoc networks in particular. His work appears in highly respected international journals and conferences.

Mohammed El Koutbi is presently working as a full professor at the University Mohammed V at ENSIAS Mohammed V University of Rabat. He received a Ph.D. degree in computer science from Montreal University, Canada, in July 2000. He has been heavily involved in teaching, research, certification activities, lab developments, and staff hiring at the university. His current research is in the area of computer networking (mobile computing and ad hoc networks) and software engineering.

Connecting a City by Wireless Backhaul: 3D Spatial Channel Characterization and Modeling Perspectives

Ruonan Zhang, Xiaohong Jiang, Tarik Taleb, Bin Li, Heng Qin, Zhimeng Zhong, and Xiaomei Zhang

The authors explain the methodology and implementation of the spatial channel measurement, and they reveal the urban wireless backhaul channel characteristics which are useful for the design and deployment of wireless backhaul for a smart city.

ABSTRACT

The backhaul forwards aggregated traffic from massive users, machines, and sensors to the core network, and is the key infrastructure to facilitate a smart city. The beamforming and spatial multiplexing technologies enable the wireless multipoint-to-point backhaul to connect dense small cell eNodeBs (eNBs) and macrocell eNBs. Because the spatial characteristics of the backhaul channels, such as the angular power spectra and spreads, determine the interference and capacity performance of the backhaul links, accurate 3D channel modeling is vital for the evaluation and comparison of candidate proposals. In this article, the concepts and methodology of spatial channel modeling are first introduced, and the state-of-the-art models developed by the standardization bodies in recent years are surveyed. Then we present a field measurement campaign on the 3D backhaul channels in an urban street, including the channel sounder implementation, field measurement, multipath parameter estimation, and propagation modeling. The Rx of the sounder (emulating a donor eNB) was installed on the rooftop of a five-story building, and the Tx was located along a street and at different altitudes to emulate a relay eNB. The channel angular power spectra in the elevation and azimuth domains were measured, and the impact by the relay's distance and altitude on the angular spreads was evaluated. In addition, a Laplace model is proposed for the power spectra in both domains. This article not only explains the methodology and implementation of the spatial channel measurement, but also reveals the urban wireless backhaul channel characteristics which are useful for the design and deployment of wireless backhaul for a smart city.

INTRODUCTION

WIRELESS BACKHAUL AND FD-MIMO

Modern cities experienced a dramatic growth in both populations and demand for information service in the last several decades. In the era of Long Term Evolution (LTE), the heterogeneous network (HetNet) has emerged as a flexible and high-capacity architecture. In a HetNet, a macrocell tier is overlaid with a dense tier of small cells (composed of picocells, femtocells, and relays). The small cell eNodeBs (eNBs) provide connectivity to nearby user equipments (UEs), while the macrocell eNBs provide ubiquitous coverage to the others. The former (acting as a relay) forwards the massive traffic from UEs to the latter (acting as a donor), and it is vice versa for the downlink data. Dense deployment of pico/femtocell eNBs with smaller radius can reuse spectrum geographically and also significantly reduce the transmission power. Nowadays, the HetNet has been regarded as the key infrastructure to provide wireless access to not only dense citizens in urban areas but also networked sensors and actuators, which can make city operations more efficient and environmentally safe.

The deployment and popularity of a HetNet require a high-capacity and flexible backhaul infrastructure to connect the relay and donor eNBs. Small cell (relay) eNBs usually need to be quickly deployable, scalable, and cost-effective [1]. Hence, compared to the wired solution (e.g., fiber), the wireless backhaul is a more suitable and viable alternative for the dense urban deployment. The flexibility also allows temporary or mobile deployment (e.g., for sport or entertainment events) and indoor coverage. The emerging large-scale antenna array systems and full-dimensional multiple-input multiple-output (FD-MIMO) technologies [2] are currently under investigation to realize high-performing and cost-effective wireless backhaul. By using a large number of active antenna elements (AAEs) in a planar array, FD-MIMO can form multiple narrow beams and concentrate radiation power in both the azimuth and elevation dimensions. By extending the spatial separation to the elevation domain as well as the azimuth domain, the 3D digital beamforming is attractive to cater for the network densification in cities. As shown in Fig. 1, the donor eNB communicates with multiple relay eNBs simultaneously in a multipoint-to-point manner by space-division multiplexing (SDM). Meanwhile, the radiated power is more concentrated spatially toward the intended receivers, leading to much higher signal strength and less interference to other users.

The current 4G/LTE-Advanced systems operate at 2.6 GHz. Operating backhaul and access links in the same frequency band (in-band wire-

Digital Object Identifier:
10.1109/MCOM.2017.1600234

Ruonan Zhang, Bin Li, and Heng Qin are with Northwestern Polytechnical University; Xiaohong Jiang is with Future University Hakodate; Tarik Taleb is with Aalto University and Sejong University; Zhimeng Zhong and Xiaomei Zhang are with Huawei Technologies Ltd.

less backhaul) is preferable for operators. First, the additional cost of buying separate frequency licenses for backhaul is eliminated. Second, the current radio units can be reused to serve the backhaul links. Therefore, the 3D beamforming working in the current frequency band is a promising solution to enable wireless backhaul for a smart city.

3D Channel Characterization for Wireless Backhaul

Capacity and reliability of wireless hackhaul are still challenging issues. In SDM by 3D beamforming, the inter-sector interference is the limiting factor on the link throughput and system capacity. The received signal-to-interference ratio (SIR) depends on the design of AAE arrays and the spatial multipath propagation characteristics (e.g., the distribution and spread of the angular power spectrum). The accurate characterization of the spatial backhaul channels provides the fundamental assumptions for the theoretical analysis and performance simulations of the candidate systems. To describe the multipath propagation in the vertical and horizontal dimensions, a 3D channel model defines the spatial parameters of *azimuth/elevation angle of arrival* (AoA/EoA), *root-mean-square* (RMS) *azimuth/elevation spread of arrival* (ASA/ESA), and the counterparts of these parameters for path departure on the transmitter side. In addition, the angular power spectra in the two domains represent the power arrival distributions over the incident angles, called the *azimuth/elevation power spectrum* (APS/EPS). The angular spreads (ASA, ESA, ASD, and ESD) are the second-order moments of the corresponding power spectra, and indicate the spreads of the power arrival/departure. They have a significant impact on the inter-sector interference in directional beamforming and spatial multiplexing.

Contributions of This Article

The main contributions of this article are twofold. First, we introduce the fundamental concept and methodology for the 3D channel modeling and then survey the current major standardization efforts in recent years. The wireless backhaul channels are essentially different from those between UEs and BSs (as discussed in detail later), but are still less explored.

Second, since the channel characterization and modeling are based on precise field measurements and parametrization, a field measurement campaign on the typical 2.6 GHz in-band backhaul channels is reported. In particular, the design of a MIMO channel sounder developed jointly by Huawei and Northwestern Polytechnical University (NPU) is presented, including the hardware implementation and path parameter estimation method. Then the measurement results of the angular power spectra and spreads in both the elevation and azimuth domains are provided. The Laplace distribution model for APS and EPS is proposed, and the impact of relay eNB's distance and altitude on the angular spreads (ASA and ESA) is revealed based on the measurement data.

The system design of the sounder provides a valuable reference for future spatial channel measurements. The proposed channel models can be used for the deployment and evaluation of the SDM-based wireless backhaul in realistic urban environments.

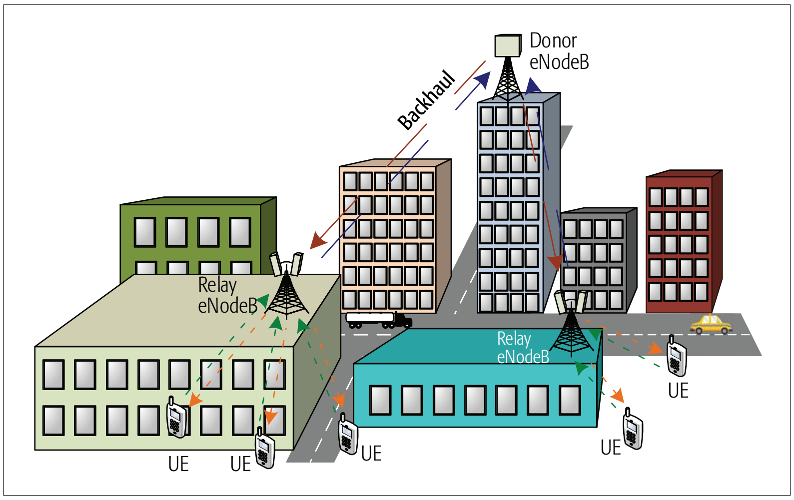


Figure 1. Multipoint-to-point wireless backhaul deployment scenario.

Current 3D Channel Modeling Works

Concept and Methodology of 3D Channel Modeling

The geometry-based stochastic channel models (GSCMs) are ray-based double-directional multipath models for MIMO links. The small-scale (SS) ray parameters (e.g., delay, power, and arrival and departure directions of each propagation path) and the large-scale (LS) channel parameters (e.g., delay spread, angular spread, shadow fading, and cross-polarization ratio) are defined, and their empirical distributions are statistically extracted from field measurements for typical scenarios. GSCMs can generate impulse responses in both the temporal and angular domains, called *channel realizations*, which are then used in wireless system simulations. Channel realizations are the superposition of contributions of rays (plane waves) with three levels of randomness. At first, the LS parameters are drawn stochastically from scenario-dependent distribution functions (by using random number generators and suitable filters). Next, the SS parameters are generated randomly also according to the tabulated distributions and the LS parameters (second moments). Third, by picking (randomly) different initial phases of the scatterers, an unlimited number of different realizations can be generated. The models are usually antenna independent, that is, different antenna configurations and element patterns can be inserted. The GSCMs have become an enabling tool for the link and system-level simulation and performance evaluation of radio interface technologies in realistic propagation conditions.

The traditional GSCMs, such as Third Generation Partnership Project (3GPP) TR25.996-v11 (spatial channel model, SCM) [3] and WINNER I [4], are 2D in the sense that only the azimuth-related parameters (i.e., AoA/AoD and ASA/ASD) are characterized, and the elevation angles of propagation paths are neglected. Inspired by the attractive features and potential benefit of FD-MIMO and 3D beamforming, several standardization projects in recent years, such as WINNER, The system design of the sounder provides a valuable reference for future spatial channel measurements. The proposed channel models can be used for the deployment and evaluation of the SDM based wireless backhaul in realistic urban environments.

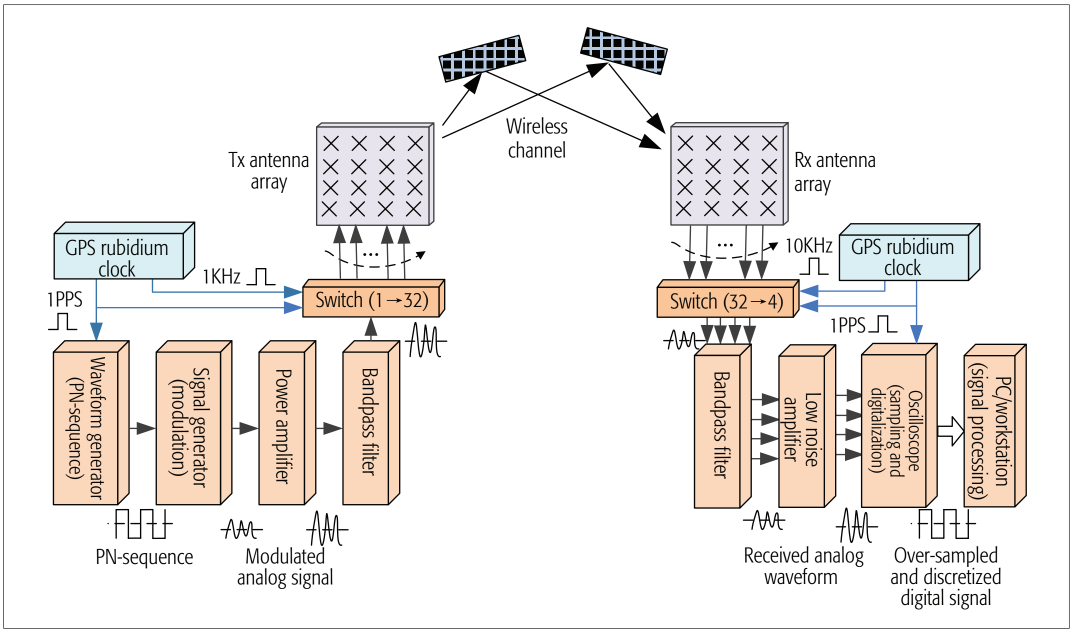


Figure 2. The architecture of the MIMO channel sounder.

3GPP, COST2100, and METIS, have made significant achievements in defining the 3D bidirectional, dual-polarized MIMO channel models. True elevation angles are associated with signal paths. The new SS parameters EoA/EoD and two new LS parameters ESA/ESD are defined in the extended models. Their distributions and correlations with other LS fading parameters such as the path loss and delay spread are also given for specific scenarios.

Current 3D Channel Models

The major efforts in the 3D channel modeling by the standardization bodies are briefly introduced as follows.

The WINNER family is a set of GSCMs, describing various environments including outdoor, indoor, and outdoor-to-indoor. The models are parameterized using results from an extensive set of measurement campaigns. The 3D channel model, WINNER+ [5], has been evolved from the earlier phases of WINNER I and II (SCM wideband extension, SCME). The novel features include the elevation parametrization, scenario-dependent polarization modeling, new scenarios, and so on. The generalization from 2 to 3D uses the same modeling approach in generating the elevation-related parameters. The composite APSs of all clusters are modeled by the wrapped Gaussian distribution. The AoAs of paths are generated by applying the inverse Gaussian function with input parameters of cluster power and ASA given for specific scenarios. The cluster EPSs have the Laplace distribution. Then the elevation angles of paths are determined by the same procedure with the original ASA values replaced by ESAs. Due to the limit of available measurement results, the cross-correlation coefficients of the new LS parameters are set as zero if the measured values are less than 0.4. Furthermore, if the measurement data for a scenario are not available, it is proposed to borrow values for other scenarios.

3GPP has also paved the way to full 3D channel models. 3GPP established a study item in 2012 [6] on 3D channel modeling. The findings are captured in the latest technical report (TR36.873) released in June 2015 [7]. The new scenarios for the UE-specific elevation beamforming, such as the UMa with high outdoor/indoor UE density, are specified. The modeling of the elevation-domain parameters is similar to that in WINNER.

European Cooperation in the Field of Scientific and Technical Research (COST) Actions in the field of wireless communications started from Action 207 (Digital Land Mobile Communications), followed by a series of successful Actions of COST 231, 259, 273, and, finally, 2100 [8]. The channel model in COST2100 characterizes the 3D channel at the individual cluster level. A cluster (scattering object) is not specific to one single link but depicted as an ellipsoid in space. The cluster angular spread (CAS) includes the arrival/departure power spread in both the azimuth and elevation planes. It is proposed that the CAS follows the lognormal distribution with a base of 10.

Another important effort is from the Mobile and Wireless Communications Enablers for the Twenty-twenty Information Society (METIS) project [9]. METIS has the goal to provide solutions for the new challenges in the fifth generation (5G) systems. Based on extensive measurement campaigns in the propagation scenarios relevant to 5G and at various frequencies between 2 and 60 GHz, new channel models are proposed for accurate 3D propagation characterization, massive MIMO with high spatial resolution, and dual-mobility links (e.g., vehicle-to-vehicle communications) [10]. The modeling approaches include the GSCM, a map-based site-specific deterministic model (based on ray-tracing), and a hybrid model for scalability. In the GSCM approach, the generation of arrival and departure directions in the azimuth and elevation domains follows the distribution models in 3GPP TR14-36873 [7] discussed above.

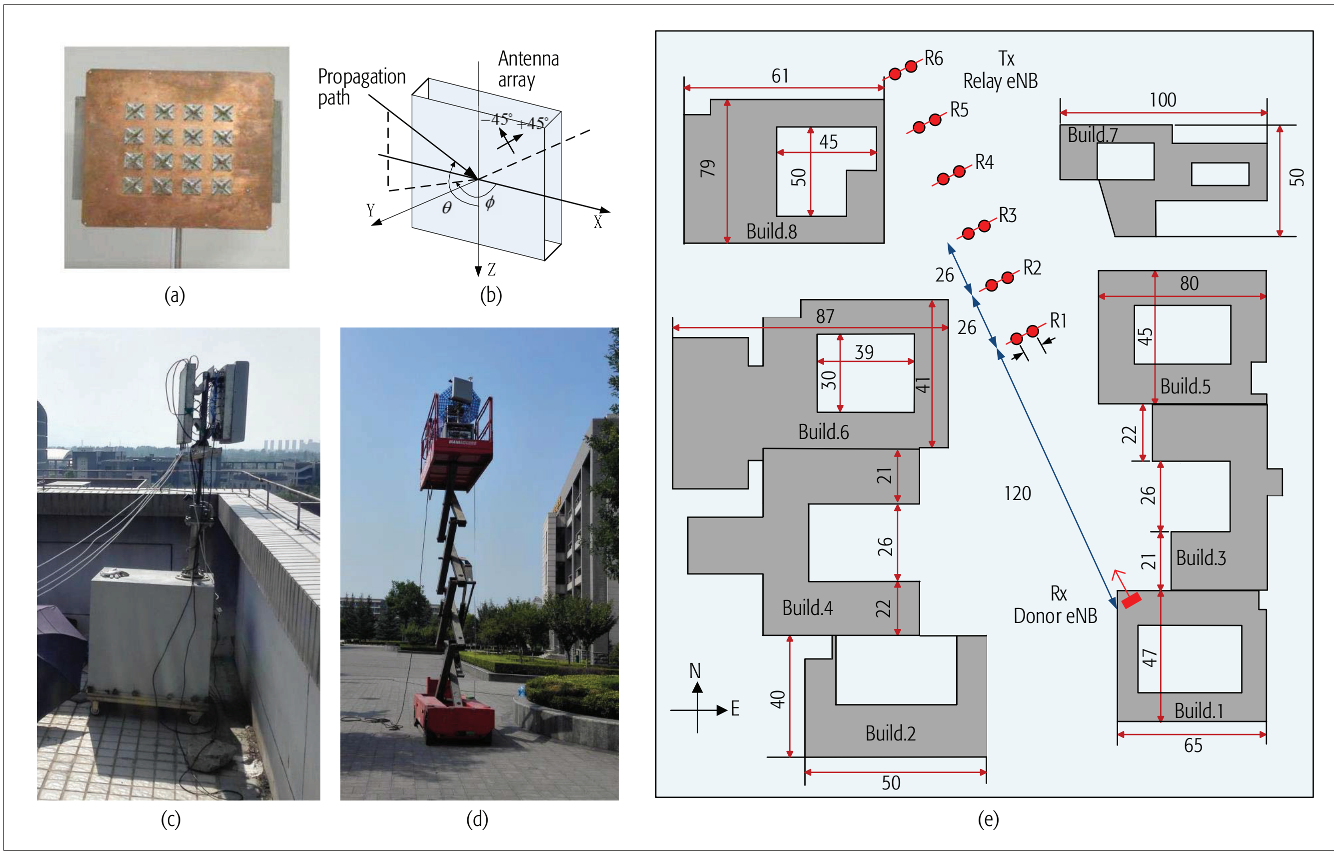


Figure 3. The channel sounder setting and measurement positions: a) antenna array; b) coordinate; c) receiver (donor eNB); d) transmitter (relay eNB); e) measurement positions (distance unit: meter).

Modeling for Wireless Backhaul Channels

The existing 3D channel models developed in recent years mainly focus on the cellular communication links between fixed BSs and mobile UEs [11]. However, the backhaul scenario, which has been less explored, is essentially different, in particular in the following three aspects:

- Relay eNBs are above the height of pedestrians. They are usually mounted on lampposts or exterior walls of buildings. The altitude will have a significant impact on the channel characteristics.
- The relay eNBs employ directional antennas (or antenna arrays) instead of omnidirectional antennas used in UEs. The main lobe of a relay eNB is pointing to its target donor eNB to concentrate radiated power. Since there are no paths emitting from the back of the directional antenna, the angular power spectra on both the relay and donor eNBs may change.
- The eNBs are fixed, and thus there is no Doppler effect. Therefore, the channel modeling focuses on the spatial properties. The high-performing wireless backhaul techniques set new requirements for radio channel and propagation modeling, which has motivated this work.

Measurement System and Scenario

Measurement System

The measurement campaign on the wireless backhaul channels was conducted using a MIMO radio sounder. Its architecture is illustrated in Fig. 2, and photographs of the transceivers are presented in Figs. 3c and 3d. The sounder probes channels in the time domain based on the direct sequence spread spectrum (DSSS) approach. The carrier of 2.6 GHz frequency is modulated by cyclic pseudo-noise (PN)-sequences and transmitted over the air. The length of a PN-sequence is 1023 chips and the rate is 62.5 Mchips/s.

The Tx and Rx are both equipped with a dual-polarized uniformed planar array (UPA). One UPA is composed of 16 pairs of cross-patches organized in a 4 × 4 matrix with spacing of half a wavelength (Fig. 3a). Each pair includes two dipoles in ± 45° polarizations, and hence there is a 4 × 4 dipole array in either polarization. A dipole has 7 dBi peak gain and a half-power beamwidth of −70° to +70° in both the azimuth and elevation dimensions. The incident angle of an impinging wave is defined with respect to the coordinate in Fig. 3b. The spatial and polarization domains are covered by measuring the channel responses between multiple transmitting and receiving antenna pairs sequentially in a time-division multiplexing (TDM) manner. On the Tx side, the signal source is connected to the UPA with a microwave switch with 1 input and 32 output ports. The 32 dipoles are stimulated one by one through the switch, and each dipole transmits 10 frames for 1 ms to probe the channel. Each frame contains five PN-sequences plus header and tail, and lasts for 0.1 ms. On the Rx side, a microwave switch with 32 input and 4 output ports is connected to the UPA, which selects 4 co-polarized dipoles in one column and outputs the received signals. The radio signals are then processed by low noise amplifiers (LNAs) and bandpass filters,

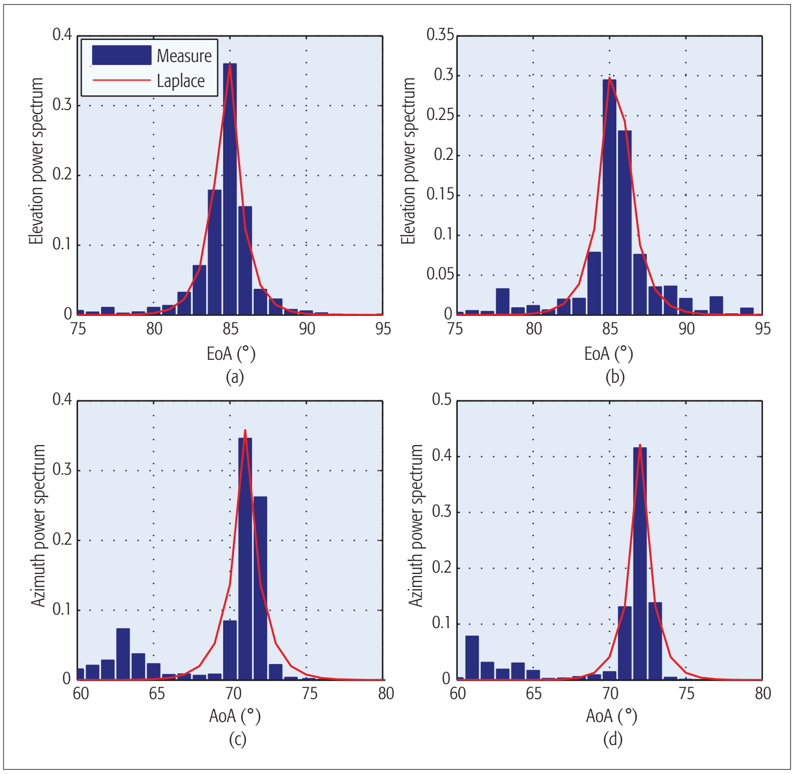


Figure 4. The measured angular spectra (EPS and APS) at different relay-to-donor distances and relay altitudes: a) region = 1, altitude = 4.5 m; b) region = 3, altitude = 7 m; c) region = 1, altitude = 4.5 m; d) region = 2, altitude = 7 m.

and finally input into the 4 ports of a Tex70608 oscilloscope where they are sampled by the rate of 625 MHz and stored for offline processing. The microwave switch connects 4 dipoles in one column for 0.1 ms to receive one complete frame and then switch to the next column. Thus, within the transmitting time of one Tx dipole (1 ms), the 8 columns of the Rx UPA take 8 × 0.1 = 0.8 ms for reception, while all 32 dipoles receive a complete frame of the sounding signal. The antenna elements are switched almost instantaneously so that the channel response remains practically constant within the Rx antenna switching period. The waveforms captured on the 4 × 4 rectangular array in one polarization during this period is a *spatial channel snapshot*, which is then used to estimate the power and 2D angle of arrival of the propagation paths. To ensure signal timing and antenna switching synchronization, two GPS-triggered rubidium clocks are used at the transceivers to provide reference clocks of 1 Hz, 1 kHz, and 10 kHz pulses per second (PPS).

The path parameters are extracted using super-resolution techniques including two tasks. First, paths are resolved from the captured radio waveforms jointly by sliding correlation (between the received signals and the spreading code) in the time domain and spatial smoothing in the space domain. Second, based on the array complex impulse responses received on the UPA, the 2D incident angles of the paths are estimated using the subspace-decomposition algorithm. Readers are referred to [12] for detailed presentation of the parameter estimation algorithm. System calibration is performed on each measurement day to compensate the transfer functions of the radio chains of the Rx in the captured channel impulse responses. The 3D radiation pattern of the Rx UPA (steering vectors) has been measured in an anechoic chamber and is incorporated in the angle of arrival estimation algorithm to remove the antenna effect. Thus, the antenna-independent channel characteristics are obtained.

Measurement Scenario

The measurement campaign was conducted on the campus of NPU. One important deployment scenario of small cells is along urban streets to forward high traffic load generated from pedestrians and passengers on vehicles. We consider the multipoint-to-point scenario where a donor eNB connects multiple relay eNBs by SDM, as shown in Fig. 1. The wireless backhaul channels in such a street canyon environment with line of sight (LOS) were measured in this campaign. Since multiple beams would be formed on the donor eNB antenna array in this scenario, we focused on the power arrival profile at the donor eNB. Furthermore, according to the channel reciprocity, the power departure profiles were also obtained by swapping the directions of the incoming paths.

The Rx was installed on the rooftop of a five-story office building by the roadside as the donor eNB (Fig. 3c). Its antenna panel was 24.5 m high and above the other buildings in the surrounding environment. The Tx was placed in six regions along the street, where the distances to the Rx changed from 120 to 250 m horizontally, as illustrated in the map in Fig. 3e. The regions were evenly spaced by 26 m. In each region, the Tx was placed at two positions which were parallel and spaced by 2 m. Furthermore, to emulate a relay eNB, the Tx was raised up to three altitudes (2, 4.5, and 7 m) by an elevator at each position (Fig. 3d). Thus, the impact of the relay distance, altitude, and scattering environment on the wireless backhaul channel could be studied.

In this work, the signals when both transceiver antennas were +45° polarized (with respect to the Rx coordinate system) were used and the co-polarization measurements were performed. At each relay position and altitude, the 32 dipoles on the Tx UPA transmitted for 25 cycles. Thus, the channel was sounded by 32/2 × 25 = 400 times in one polarization, and 400 independent channel snapshots were captured (as mentioned earlier, one snapshot was the array signal received by the Rx UPA during 0.8 ms when one Tx dipole was transmitting). Furthermore, from 2 measured positions in one region, 400 × 2 = 800 snapshots were obtained for a given relay-donor distance and altitude. From the array signal of each snapshot, the 10 most significant multipath components (MPCs) were resolved from the waveforms and parameterized. Based on the EoA, AoA, and power of these MPCs, the angular power spectra and spreads were obtained.

There was variation among the 800 snapshots at one altitude in a region, caused by the movements of the Tx over a small distance and the people and vehicles in the scattering environment. However, the relay-donor distance, altitude, and surrounding buildings remained stationary. Therefore, the variation among the snapshots is referred to as the *small-scale spatial fading*. The statistics of the snapshots collected from one Tx region and altitude represent the small-scale fading of the channel. On the other hand, the variation of the model parameters (e.g., mean and variance) among different relay regions and altitudes reflects the *large-scale spatial fading*.

Measurement Results and Modeling

Angular Power Spectrum and Spread in the Elevation Domain

For a given relay-donor distance and relay's altitude, the EPS and ESA are calculated from the power and EoAs of the 800 × 10 = 8000 sample MPCs (800 channel snapshots and 10 MPC samples per snapshot). As illustrative examples, the obtained EPSs when the Tx was placed 4.5 m high in region 1 and 7 m high in region 3 are plotted in Figs. 4a and 4b, respectively.

As shown in Fig. 4, the power arrival mostly concentrates in the main clusters in this LOS scenario. Hence the statistical model for the EPS of the main cluster is established. Since the main clusters are symmetric with one peak, we consider the candidate distribution functions of *Laplace*, *normal*, and *Cauchy* probability distribution functions (PDFs). By using the AIC-testing method, they are compared in fitting the empirical power spectra, and the Laplace distribution can model the measurement graphs best. The Laplace PDF is denoted by $Lap(\mu_E, b_E)$, where μ_E and b_E are the location parameter and the scale parameter, respectively. For example, the best-fit Laplace PDFs, obtained with minimum mean square error (MMSE), are superimposed in Figs. 4a and 4b. The location parameter μ_E is approximated by the LOS direction, that is, $\mu_E = \theta_{LOS}$ where θ_{LOS} is the EoA of the LOS path and calculated geometrically according to the geographic locations of the relay and donor eNBs.

The scale parameter b_E describes the spread of the main cluster, and $\sqrt{2}b_E$ (the second moment of the Laplace distribution function) is actually the ESA of the main cluster. The values of b_E for all the measurement regions and altitudes are plotted in Fig. 5. First, it is observed that the impact of the relay altitude, denoted by h_{rel}, is remarkable. For a given relay-donor distance (denoted by d_{rel}), b_E is larger when the relay altitude is increased. This shows that higher relay altitude leads to larger ESA at the donor eNB. Second, as d_{rel} increases, b_E has fluctuations, and the linear regressions on the measured values show two tendencies. When the relay is low (2 m high), the expectation of b_E is almost constant for various d_{rel}. But b_E has a common decreasing tendency at the altitudes of 4.5 and 7 m. This indicates that when the relay is moved away from the donor, the channel ESA almost does not change if the relay is close to the ground, but decreases if the relay is placed higher. Third, it can be seen that when the distance is increased, the values of b_E at the three altitudes converge. Therefore, the impact of a relay's altitude is more significant when it is close to the donor horizontally, and vice versa.

Figure 5 shows that b_E varies linearly with respect to the relay-donor distance, and the slope and intersect depend on the relay altitude. Therefore, b_E can be modeled by a linear function with respect to the independent variable d_{rel} and the parameter h_{rel}, expressed as

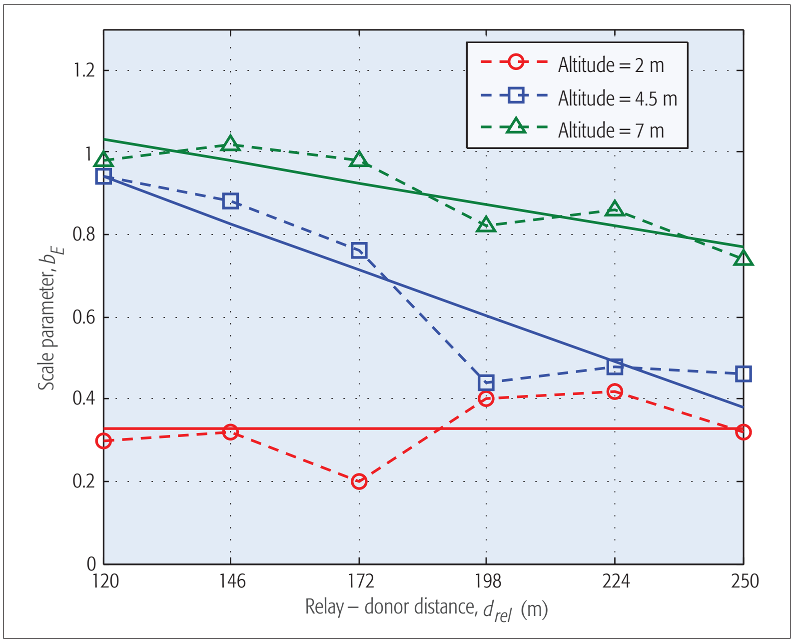


Figure 5. The scale parameter b_E of the best-fit Laplace distribution functions for the measured main cluster EPS.

$$b_E = \alpha\,(h_{rel})d_{rel} + \beta(h_{rel}). \quad (1)$$

The coefficients of the linear function for the three measured relay altitudes, h_{rel} = 2, 4.5, and 7 m, are obtained directly by the linear regression as plotted in Fig. 5. The results are (α, β) = (0, 0.33), (–0.004, 1.457), and (–0.002, 1,274). The coefficients for other h_{rel} from 2 to 7 m can be derived by interpolation. The linear model in Eq. 1 reflects the impact of the relay's altitude and horizontal distance and their interaction on the channel ESA. Based on the linear model, the value of b_E for the continuous relay distance (in the range of $d_{rel} \in$ [120, 250] m) and altitude (in the range of $h_{rel} \in$ [2, 7] m) can be determined. Then the realizations of the main cluster EPS can be generated by the Laplace model of $Lap(\theta_{LOS}, b_E)$.

Angular Power Spectrum and Spread in the Azimuth Domain

Following the same approach for the elevation domain, the measured APSs when the Tx was located at 4.5 m in region 1 and 7 m in region 2 are plotted in Figs. 4c and 4d for demonstration. In addition to the main clusters, due to the scattering and reflection by the roadside buildings, small clusters can be observed. Again, the symmetric shape of the main clusters suggests the candidate functions of *Laplace*, *Normal*, and *Cauchy* PDFs. The AIC-testing indicates that the APSs of the main clusters can be best fit by the Laplace distribution, which is denoted by $Lap(\mu_A, b_A)$. $\sqrt{2}b_A$ is the ASA of the main cluster.

The location parameter μ_A is the AoA of the LOS direction, denoted by ϕ_{AoA} and calculated geometrically. The scale parameter b_A of the best-fit Laplace PDFs is shown in Fig. 6a for the relay distances and altitudes measured in this campaign. As can be observed, as the horizontal distance d_{rel}

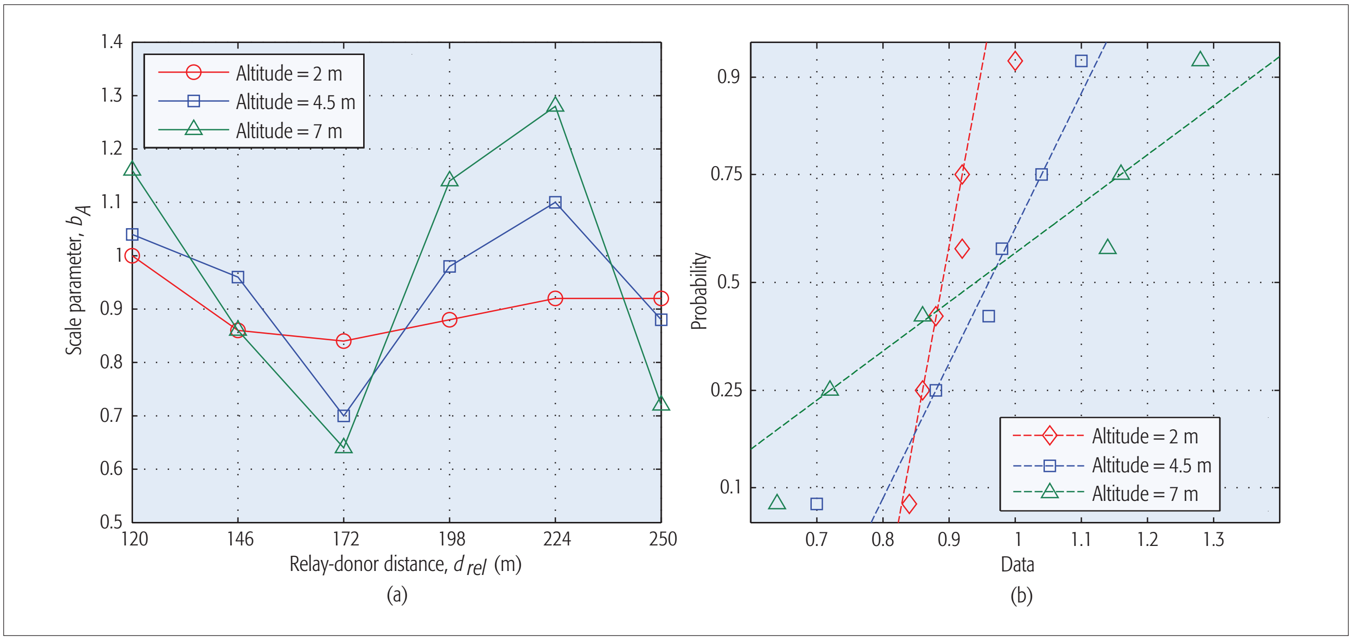


Figure 6. The scale parameter b_A of the best-fit Laplace distribution functions for the measured main cluster APSs: a) measurement results of b_A; b) probability plot of b_A.

increases, b_A presents considerable fluctuations at all the altitudes. From the street map in Fig. 3e, the building layout by the roadsides is irregular. There are gardens of grass and small trees between the buildings, and the distance between the Tx and roadside buildings varies. This indicates that the change of the roadside building layout has a significant impact on the ASA of the backhaul channel. Furthermore, interesting properties of b_A can be found. First, the variation tendencies at the three altitudes are coincident with each other, as plotted in Fig. 6a. Therefore, the impact of surrounding environment is consistent for different relay altitudes. Second, when h_{rel} is larger, the fluctuation of b_A is more significant. This indicates the interaction between h_{rel} and d_{rel}. Higher relay altitude leads to larger variance of b_A when the relay is located at different distances.

To model the distribution of b_A, the probability plots in Fig. 6b testify to the measurement values against the normal distribution using the function *normplot* in MATLAB. As observed, the sample values are close to the lines and hence follow the normal distributions well. In addition, the Kolmogorov-Smirnov test has shown that the measurement results of b_A have the normal distribution at the significance level of 0.05. Therefore, we propose the normal distribution model for b_A of the main cluster APS, that is, $b_A \sim N(\mu_b, \sigma_b^2)$, where μ_b and σ_b are the mean and standard deviation (STD), respectively. According to the fitting results plotted in Fig. 6b, μ_b of the normal PDFs for the three relay altitudes are 0.90°, 0.94°, and 0.97°. Hence, we can approximate μ_b by a constant, that is, $\mu_b \approx 0.94°$. The values of σ_b of the normal PDFs are 0.05°, 0.15°, and 0.26° which increase quite linearly for the relay altitudes. Therefore, we can model σ_b by a linear function with respect to h_{rel}. By using the linear regression, it can be derived as $\sigma_b = 0.042\, h_{rel} - 0.036$. The value of σ_b for the relay altitude between 2 and 7 m can be calculated simply by the linear function.

To generate channel realizations, μ_b and σ_b are first determined according to d_{rel} and h_{rel} as specified above. Then the scale parameter b_A is drawn stochastically by the normal distribution $N(\mu_b, \sigma_b^2)$. Finally, the main cluster APS is obtained as the Laplace distribution of $Lap(\phi_{AoA}, b_A)$.

Conclusion

High-capacity and scalable wireless backhaul will be essential for the HetNet in providing ubiquitous coverage for dense citizens and sensors in smart cities. The successful deployment of wireless backhaul using high-throughput spatial multiplexing requires a solid understanding and accurate propagation models of the radio channels between relay and donor stations. In this article, we have surveyed the development of 3D channel models defined by the standardization bodies in recent years. The principles, methodologies, and channel sounder implementation for 3D channel modeling are introduced. Then a field channel measurement campaign on the 3D co-polarized wireless backhaul channels in a typical urban street environment is presented. We have shown that the APS and EPS of the main MPC cluster can be well modeled by the Laplace distribution. The location and scale parameters have been formulated with respect to the relay altitude and distance. The measurement results provide an initial insight into the 3D multipath propagation of street backhaul channels and how the relay altitude and distance affect the channel characteristics. Furthermore, the Laplace model and stochastically generated channel realizations can be used for the analysis and simulations of the wireless backhaul technologies, such as the SIR evaluation and interference cancellation in beamforming and spatial multiplexing. Third, the proposed power spectrum model can extend current 3D GSCMs reviewed previously.

The street backhaul channel model can be further extended to the packet level. By considering the wireless channel dynamics, the packet-level model describes the link SIR or packet error rate on the basis of packet transactions [11, 13]. It will be a useful tool for fast simulations of upper-layer protocols and applications of the SDM wireless backhaul. Another avenue of interest is the channel characterization for the millimeter-wave (mmWave) capabilities at frequencies such as 28 and 38 GHz. The mmWave communications are expected to provide multi-gigabit rates to further enhance the backhaul capacity.

Acknowledgment

This work was partly supported by NSFC (61571370 and 61601365), Natural Science Basic Research Plan in Shaanxi Province (2015JM6349 and 2016JQ6017), and National Civil Aircraft Major Project of China (MIZ-2015-F-009)).

References

[1] "Macrocell Mobile Backhaul Equipment and Services Market Share and Forecast Report," *Infonetics Research*, tech. rep., 2012.

[2] Y.-H. Nam *et al.*, "Full-Dimension MIMO (FD-MIMO) for Next Generation Cellular Technology," *IEEE Commun. Mag.*, vol. 51, no. 6, June 2013, pp. 172–79.

[3] 3GPP Tech. Spec. Group Radio Access Networks, "Spatial Channel Model for Multiple Input Multiple Output (MIMO) Simulations (Release 11)," tech. rep. TR 25.996 v11.0.0, Sept. 2012.

[4] "D5.4: Final Report on Link and System Level Channel Models," WINNER, tech. rep., Oct. 2005.

[5] "D5.3: WINNER+ Final Channel Models," WINNER, tech. rep. v. 1.0, June 2010.

[6] "Study on 3D-Channel Model for Elevation Beamforming and FD-MIMO Studies for LTE," 3GPP TSG RAN Plenary, Barcelona, Spain, Study Item Description RP-122034, Dec. 2012.

[7] 3GPP Tech. Spec. Group Radio Access Networks, "Study on 3D Channel Model for LTE (Release 12)," tech. rep. TR 36.873 v. 12.2.0, June 2015.

[8] R. Verdone and A. Zanella, Eds., *Pervasive Mobile and Ambient Wireless Communications COST Action 2100*, Springer, 2012.

[9] 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.

[10] "METIS Channel Models," tech. rep. ICT-317669, July 2015.

[11] R. Zhang *et al.*, "Channel Measurement and Packet-Level Modeling for V2I Spatial Multiplexing Uplinks Using Massive MIMO," *IEEE Trans. Vehic. Tech.*, vol. 65, no. 10, Oct. 2016, pp. 7831–43.

[12] R. Zhang *et al.*, "Two-Dimensional DoA Estimation for Multipath Propagation Characterization Using the Array Response of PN-Sequences," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, Jan. 2016, pp. 341–56.

[13] R. Zhang and L. Cai, "A Packet-Level Model for UWB Channel with People Shadowing Process Based on Angular Spectrum Analysis," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, Aug. 2009, pp. 4048–55.

Biographies

Ruonan Zhang [S'09, M'10] received his B.S. and M.Sc. degrees from Xi'an Jiaotong University, China, in 2000 and 2003, respectively, and his Ph.D. degree from the University of Victoria, Canada, in 2010, all in electrical and electronics engineering. He worked as an IC architecture engineer at Motorola Inc. and Freescale Semiconductor Inc. in Tianjin, China, from 2003 to 2006. Since 2010, he has been with the Department of Communications Engineering at Northwestern Polytechnical University, Xi'an, Shaanxi, China, and he is currently a professor. His current research interests include wireless channel measurement and modeling, architecture and protocol design of wireless networks, and satellite communications. He has been a recipient of the New Century Excellent Talent Grant from the Ministry of Education of China, and the best paper award of IEEE NaNA 2016. He has served as a Local Arrangement Co-Chair for IEEE ICCC 2013, and an Associate Editor for the *Journal of Communications and Networks*.

Xiaohong Jiang [SM] is currently a full professor of Future University Hakodate, Japan. Before joining Future University, he was an associate professor, Tohoku University, from February 2005 to March 2010. His research interests include computer communications networks, mainly wireless networks and optical networks, network security, interconnection networks for massive parallel computing systems, routers/switches design for high performance networks, and so on. He has published over 280 technical papers in premium international journals and conferences, which include over 60 papers published in top IEEE journals and top IEEE conferences, like *IEEE/ACM Transactions on Networking*, the *IEEE Journal of Selected Areas on Communications*, *IEEE Transactions on Parallel and Distributed Systems*, *IEEE Transactions on Communications*, and IEEE INFOCOM. He was the winner of the Best Paper Award and Outstanding Paper Award of IEEE HPCC 2014, IEEE WCNC 2012, IEEE WCNC 2008, IEEE ICC 2005 Optical Networking Symposium, and IEEE/IEICE HPSR 2002. He is a member of ACM and IEICE.

Tarik Taleb [M] (talebtarik@ieee.org) is currently a professor at the School of Electrical Engineering, Aalto University. He has worked as senior researcher and 3GPP standards expert at NEC Europe Ltd. Prior to his work at NEC, until March 2009, he worked as an assistant professor at the Graduate School of Information Sciences, Tohoku University, Japan, in a lab fully funded by KDDI. He received his B.E. degree in information engineering with distinction, and his M.Sc. and Ph.D. degrees in information sciences from Tohoku University in 2001, 2003, and 2005, respectively. His research interests lie in the field of architectural enhancements to mobile core networks (particularly 3GPP's), mobile cloud networking, mobile multimedia streaming, and social media networking. He has also been directly engaged in the development and standardization of the Evolved Packet System. He is a member of the IEEE Communications Society Standardization Program Development Board and serves as Steering Committee Chair of the IEEE Conference on Standards for Communications and Networking. He has received many awards for his many contributions to the area of mobile networking.

Bin Li [S'12, M'14] received his Ph.D. degree from the School of Electronics and Information Engineering, Xi'an Jiaotong University, in 2014. From 2012 to 2013, he was a visiting student in the University of Louisville, Kentucky. He is currently an assistant professor at Northwest Polytechnical University. His research interests include network coding, channel measurement, the Internet of Things, and array signal processing.

Heng Qin [S'14] received his B.S. degree from Northwestern Polytechnical University in 2014, where he is currently an M.Sc. student. His research interests include wireless channel modeling, communications signal processing, and massive MIMO.

Zhimeng Zhong received his B.Sc., M.Sc., and Ph.D. degrees from Xi'an Jiaotong University in 2002, 2005, and 2008, respectively, all in electronic engineering. From 2009 to 2013, he was an algorithm engineer with Huawei Technologies Ltd., Xi'an. After working with the Research Institute of Aerospace Radio Technology in 2014, he rejoined Huawei Technologies Ltd., Shanghai. His current research interests include digital wireless communications, wireless channel measurement and modeling, and multiple-input multiple-output systems.

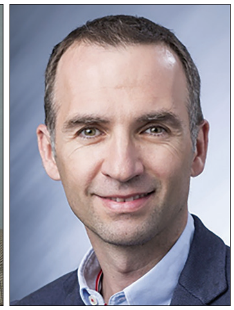
Xiaomei Zhang received her B.S. and M.S. degrees from Xi'an Jiaotong University in 2005 and 2008, respectively. Since 2008, she has been an algorithm engineer of Huawei Technologies Ltd.. Her research interests include wireless channel measurement and modeling and MIMO systems.

Another avenue of interest is the channel characterization for the millimeter-wave (mmWave) capabilities at the frequencies such as 28 and 38 GHz. The mmWave communications are expected to provide multi-gigabit rates to further enhance the backhaul capacity.

5G NETWORK SLICING: PART 1 – CONCEPTS, PRINCIPLES, AND ARCHITECTURES

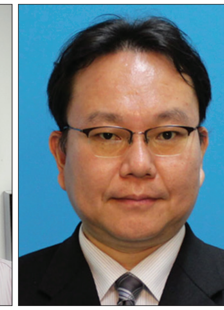

Konstantinos Samdanis


Steven Wright


Albert Banchs


Antonio Capone


Mehmet Ulema


Kazuaki Obana

Network slicing has evolved from a simple fixed network overlay concept to a fundamental feature of the emerging multi-provider fifth generation (5G) systems, enabling new business opportunities by facilitating flexible and agile support for multi-service and multi-tenancy. Network slicing can drastically transform the monolithic "one network fits all" architecture by abstracting, isolating, and separating logical network behaviors from the underlying physical network resources, opening the network to third parties and providing the means of integrating vertical market segments. Network operators can exploit network slicing not only for reducing capital and operational expenditures, but for enabling network programmability and innovation, in order to enrich the offered services from providing simple communication pipelines to a wider range of business solutions.

Network slicing in 5G systems defines logical, self-contained networks that consist of a mixture of shared and dedicated resource instances, such as radio spectrum or network equipment, and virtual network functions. Virtualization allows the decoupling of network functions from proprietary hardware appliances in order to create distinct building blocks that can be flexibly chained to create value-added communication services.

The notion of resources in 5G network slicing includes network, compute, and storage capacity resources, virtualized network functions, and radio resources. Service designers can select the optimal control/user plane split, as well as compose and allocate virtualized network functions at particular locations inside the core or radio access network depending on the service requirements. 5G network slicing enables a particular communication service exploiting the principles of software defined networks (SDN) and network functions virtualization (NFV) to fulfill the business and regulatory requirements. The achieved networking and service flexibility enables a radical change, beyond network sharing, enabling tailored services to third parties and vertical market players.

This Feature Topic includes five outstanding articles that focus on 5G network slicing — elaborating the network slicing concept, principles, and architectures from different perspectives considering distinct network parts, for instance, the radio access, mobile core network, backhaul, and fixed IP, as well as the support of different tenants and services. In addition, these articles provide insight on the architecture aspects, pointing out enhancements of the current mobile networks and on key enabling technologies, highlighting operational features for efficient management and orchestration in order to achieve the desired flexibility, service isolation, and performance customization.

The first article, "Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks," by P. Rost, C. Mannweiler, D. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker, presents the concept of network slicing in 5G networks focusing on the architectural aspects associated with the coexistence of dedicated and shared slices, bringing light to the revenue potential. In addition, it analyzes the different options for providing a flexible radio access network considering the impact on the 5G mobile network design.

The second article, "Slicing for 5G with SDN/NFV: Concepts, Architectures and Challenges," by J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, L. Lorca, and J. Folgueira elaborates the concept of 5G network slicing, analyzing the Open Networking Foundation (ONF) SDN and European Telecommunications Standards Institute (ETSI) NFV technologies in an effort to introduce a combined solution for realizing network slicing.

The following article, "PERMIT: Network Slicing for Personalized 5G Mobile Telecommunications," by T. Taleb, M. Corici, A. Nakao, and H. Flinck, brings light to the 5G network architecture in terms of elasticity and scalability requirements with respect to network slicing for providing deep customization at different granularity levels, while ensuring the desired service delivery on the top of a common infrastructure.

The next article, "Network Slicing in 5G: Survey and Challenges," by X. Foukas, G. Patounas, A. Elmokashfi, and M. Marina, analyzes the concept of network slicing with respect to multi-service support and provides a comprehensive overview of different network slicing proposals, presenting a framework that can evaluate their maturity and identify potential open research challenges.

The last article, "5G-ICN: Delivering ICN Services over 5G using Network Slicing," by R. Ravindran, A. Chakraborti, S. O. Amin, A. Azgin, and G. Wang, elaborates the notion of application-centric network slicing over a programmable compute, storage, and transport infrastructure considering information-centric networking (ICN). The article emphasizes the flexibility offered by NFV/SDN over which 5G-ICN can be realized and introduces the concept of mobility as a service.

We hope that these five articles provide an overview to the readers with a representative taste of the 5G network slicing concepts, principles, and architectures.

Biographies

Konstantinos Samdanis (konstantinos.samdanis@huawei.com) is a principal researcher at Huawei for 5G carrier networks. He is involved in research for 5G SDN/NFV architectures and network slice OS, being also active at BBF in wireless-wired converged networks and the 5GPPP Architecture WG. Previously he worked for NEC Europe, Germany, as a senior researcher and a broadband standardization specialist, involved in numerous EU projects and 3GPP. He received his Ph.D. and M.Sc. degrees from Kings College London.

Steven A. Wright, M.B.A., Ph.D., J.D. (s.wright@ieee.org), has been involved in the communications industry in research, development, program management, product management, university faculty, and board roles, resulting in 50+ patents and a number of diverse publications. He has presented his research at international conferences in Europe, Asia, Australia, and the United States, and founded the IEEE NFV SDN conference.

Albert Banchs (banchs@it.uc3m.es) is a professor with the University Carlos III of Madrid (UC3M) and has a double affiliation as deputy director of the IMDEA Networks institute. Before joining UC3M, he was at ICSI Berkeley, Telefonica I+D, and NEC Europe Ltd., respectively. He has authored more than 100 peer-reviewed publications, as well as several patents and standardization proposals. A current major focus is on the 5G architecture work within the European 5G NORMA project.

Antonio Capone [SM] (antonio.capone@polimi.it) is a professor at Politecnico di Milano, where he is the director of the ANTLab and Vice-Dean of the School of Engineering. He serves on the TPCs of major conferences in networking, is an Editor of *IEEE Transactions on Mobile Computing*, *Computer Networks*, and *Computer Communications*, and was an Editor of *ACM/IEEE Transactions on Networking* from 2010 to 2014.

Mehmet Ulema (mehmet.ulema@manhattan.edu) is a professor at Manhattan College, New York. Previously, he was with AT&T Bell Laboratories and Bellcore. He also serves as Director of Standards Development in ComSoc. He has had leadership roles for many major IEEE conferences. He is on the Editorial Boards of the *IEEE Journal of IoT* and the *Journal of Network and Services Management*. He received his Ph.D. from Polytechnic University, Brooklyn, New York, and his B.S. and M.S. from Istanbul Technical University.

Kazuaki Obana (kazuaki.obana.uz@nttdocomo.com) is an executive research engineer and leads the ICT Network Cloud Research Group in the Research Laboratories, NTT DoCoMo. He is involved in research on the future mobile core network for the 5G era, standardization in ETSI ISG NFV, and open source activity in OPNFV. He holds B.E./M.E. degrees in electrical engineering from Waseda University, Tokyo.

Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks

Peter Rost, Christian Mannweiler, Diomidis S. Michalopoulos, Cinzia Sartori, Vincenzo Sciancalepore, Nishanth Sastry, Oliver Holland, Shreya Tayade, Bin Han, Dario Bega, Danish Aziz, and Hajo Bakker

The authors argue for network slicing as an efficient solution that addresses the diverse requirements of 5G mobile networks, thus providing the necessary flexibility and scalability associated with future network implementations. They elaborate on the challenges that emerge when designing 5G networks based on network slicing.

ABSTRACT

We argue for network slicing as an efficient solution that addresses the diverse requirements of 5G mobile networks, thus providing the necessary flexibility and scalability associated with future network implementations. We elaborate on the challenges that emerge when designing 5G networks based on network slicing. We focus on the architectural aspects associated with the coexistence of dedicated as well as shared slices in the network. In particular, we analyze the realization options of a flexible radio access network with focus on network slicing and their impact on the design of 5G mobile networks. In addition to the technical study, this article provides an investigation of the revenue potential of network slicing, where the applications that originate from this concept and the profit capabilities from the network operator's perspective are put forward.

INTRODUCTION

Future mobile networks will be subject to manifold technical and service requirements with respect to throughput, latency, reliability, availability, as well as operational requirements such as energy efficiency and cost efficiency. These requirements stem from an increasing diversity of services carried by the mobile network as well as novel application areas such as Industry 4.0, vehicular communication, or smart grid. In order to provide cost- and energy-efficient solutions, it is necessary to avoid a largely segmented solution space with deployments of individual mobile network solutions for each use case. Hence, there is the need for a flexible and scalable mobile network. Thus, *flexibility and scalability* go hand in hand and ensure that the mobile network can be appropriately adopted to the network environment of a particular use case (e.g., available bandwidth, transport network, or access point density). Furthermore, the actual quantitative technical requirements may differ significantly; for example, while packet error rates of 10^{-4} are acceptable in a mobile broadband system, industrial use cases require significantly lower packet *and* frame error rates, particularly if latency constraints must be met [1].

DEFINITION OF NETWORK SLICES

In order to cope with the above requirements, the concept of network slicing has been proposed as a means of providing better resource isolation and increased statistical multiplexing [2]. The Next Generation Mobile Network Alliance (NGMN) defines network slicing as a concept for running multiple logical networks as independent business operations on a common physical infrastructure [2]. Each network slice represents an independent virtualized end-to-end network and allows operators to run different deployments based on different architectures in parallel. In the following, the term *network slice* refers to a specific instance of such a logical network (instantiated according to a pre-defined *network slice blueprint*).

A network slice as a logical end-to-end construct is self-contained, having customized functions also including those in the user equipment (UE), and using network function chains for delivering services to a given group of devices. Employing network slicing in 5G networks engenders a number of challenges, in part due to difficulties in virtualizing and apportioning the radio access network (RAN) into different slices, as discussed in the ensuing subsection.

DESIGN CHALLENGES

In the following, we provide a detailed explanation of the potential challenges associated with the implementation of network slicing in future networks.

Granularity constraints in spectrum and radio-level resource sharing: Unlike fixed network slices, which can be scaled up by adding more hardware resources, RAN slicing quickly runs into a *physical* constraint: the limited availability of spectrum. This limitation is deteriorated if dedicated carriers are assigned to individual slices, since this approach does not leverage the network's potential for multiplexing gains.

Radio access technology (RAT) heterogeneity and spatial diversity: It is expected that 5G will incorporate several kinds of RATs and air interfaces, each with different capabilities and needs. General-purpose infrastructure providers will need to carefully plan and apply different technologies to serve diverse tenant needs. However, it may be infeasible to satisfy the needs of each application at any location. For instance, tactile Internet may

Peter Rost, Christian Mannweiler, Diomidis S. Michalopoulos, Cinzia Sartori, Danish Aziz, and Hajo Bakker are with Nokia Bell Labs Germany; Vincenzo Sciancalepore is with NEC Laboratories Europe; Nishanth Sastry and Oliver Holland, are with King's College London; Shreya Tayade and Bin Han are with University of Kaiserslautern; D ario Bega is with Univeristy Carlos III of Madrid and IMDEA Networks Institute.

Digital Object Identifier: 10.1109/MCOM.2017.1600920

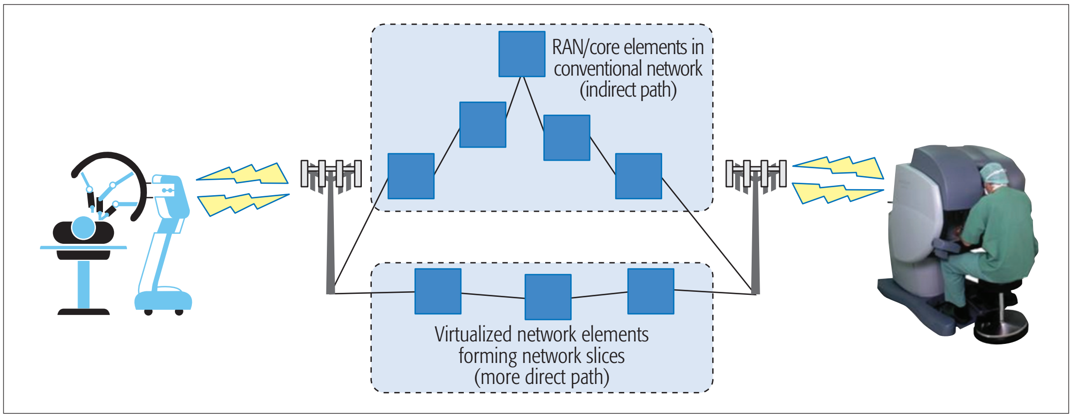


Figure 1. Virtualized edge network slices achieving a more direct path compared to (fixed) network elements in a tactile Internet remote surgical operation example.

Two exemplary applications for network slicing are "smart factory" industrial communications and the "Tactile Internet." In both cases, wireless communication conveys force (or "kinaesthetic") information to a client, and in the Tactile Internet case especially touch sensations such as texture might be conveyed.

require careful positioning of resources to minimize latency. In another example, an industrial control network might *have* to use a certain computational resource in a given location for security reasons.

Managing information exposure and sharing constraints: Different flavors of network slices can be defined based on the extent of network elements that are shared, for example, whether only the physical layer (PHY) is shared, whether the medium access control (MAC) layer is shared, or even whether the complete RAN is shared. The more information that can be provided by the infrastructure about the shared parts to the network slice, the more efficiently the slice can be operated. However, exposing information also creates new potential security vulnerabilities between infrastructure providers and their clients (also known as "tenants" [2]), as well as between tenants themselves. Security requirements of specific tenant applications, such as traffic associated with emergency services or machine control (e.g., remote surgery or vehicular control), could put constraints on how the slices are partitioned, or even prevent network slices coexisting and thus share the same hardware at all.

Transparency of network slicing: A major question is whether a slice can be extended all the way to the UE; that is, whether the definition of the slice will be transparent to the UE, or whether the UE will be aware of the network slice. A slicing-aware UE may open up new possibilities (e.g., simplification of multi-slice connectivity). However, it also creates new challenges for network slices; for example, UE mobility may need to be handled by the slice provider as part of the slice setup and maintenance.

Network slice requests brokerage: Network slicing in 5G networks enables a new ecosystem in which different tenants issue requests to an infrastructure provider for acquiring network slices. Since spectrum is a scarce resource for which overprovisioning is not possible, applying an "always accept" strategy for all incoming requests is not feasible. This calls for novel algorithms and solutions to allocate network resources among different tenants, allowing an infrastructure provider to accept or reject network slice requests with the objective of maximizing the overall utility.

Network Slicing Applications and Profitability

This section highlights the major applications where the slicing concept is expected to play a key role in future networks, along with a profitability assessment as seen through the lens of the operator.

Slicing Applications: Smart Factory and the Tactile Internet: Two exemplary applications for network slicing are smart factory industrial communications and the tactile Internet. In both cases, wireless communication conveys force (or kinesthetic) information to a client, and in the tactile Internet case especially, touch sensations such as texture might be conveyed. The purpose of these applications is to achieve the touching or manipulation of remote real or virtual objects by a human or machine. If kinesthetic information is conveyed to a machine client, the latency requirement might correspond to the challenging 1 ms in 5G. For human clients, this is relaxed to around 5 ms, or more than 100 ms for tactile information alone conveyed to humans. Both applications also require extremely high reliability and security requirements, noting the mission-critical characteristics associated with them.

Network slicing can address the latency, reliability, and security requirements of these applications. Referring to the remote surgery example shown in Fig. 1, virtualization allows the instantiation of network elements at appropriate locations for the communication to proceed as close to a direct path as possible, reducing propagation delay and hence latency. The instantiation of virtualized elements collectively forming network slices allows multiple instances of such applications to viably share available computational and other resources end to end, making virtualization viable from a management point of view. Slicing also assists reliability through the reservation of hardware and other resources as distinct slices, even in some cases potentially down to spectrum resources. Security can benefit through slicing (e.g., tenant isolation and "sandboxing" capabilities). Furthermore, slices may only be operated locally within a factory in order to ensure data privacy while its operation is coordinated with slices operated by public mobile network operators (MNOs) offering Internet services or specific network functionality such as mobility management.

Slicing as a Means to Increase Network Revenue: Besides the flexibilities provided by net-

We underline the key elements that enable the coexistence of dedicated and shared slices within a common network architecture, and elaborate on the implementation of the notion of network slicing in the RAN and in CN, putting particular emphasis on the concept of software defined mobile network control (SDM-C).

work slicing, it is also important to demonstrate the economic profit of applying network slicing from the MNO's perspective. The cost in terms of capital expenditures (CAPEX) and operation expenditures (OPEX) of a network is often much higher in comparison to the revenue expected by the operators. One reason for low revenue is underutilization of the network. According to the KPI requirements, different use cases may have highly specified resource demands. Nevertheless, in the current framework, the operator can only provide the network with an unspecified resource bundle for general utilization. Hence, most of the resources are often reserved for use cases with only slight demands on them, and are thus wasted. With network slicing, the MNOs are able to efficiently analyze the operational cost and revenue generated from the respective slice. According to the analysis, they can allocate different network resource bundles to different slices, which makes the resource management much more structured, flexible, and efficient. As a result, the very same network can be utilized to seamlessly provide more and better services (i.e., generate more revenue without any increase in CAPEX).

Moreover, concepts such as cooperative slicing and inter-operator network sharing can be efficiently implemented by optimizing the network cost model for increasing the overall revenue, and simultaneously providing network scalability. For example, the sliced network of operator A is serving several services and still has a few resources unutilized. Hence, the network can implement another slice that requires less resources but more coverage area, and might belong to operator B. The moderated approach of implementing slicing is beneficial for both the operators for providing more services without increasing CAPEX while simultaneously generating revenue from the unutilized resources. Hence, the network provider needs a new algorithm (e.g., based on a threshold rule) that allows it to decide whether to accept or reject an incoming network slice request while maximizing its revenue.

Related Work

A simplified network slice concept has been exhaustively studied in the literature, wherein a dedicated portion of RAN elements are fully reserved to particular services such as an "isolated slice." However, with the advent of advanced network virtualization techniques, the notion of network slicing in 5G has evolved to more flexible sharing, aiming to attain a significant multiplexing gain while still guaranteeing isolation and separation. The network virtualization substrate (NVS) was introduced in [2], allowing the infrastructure provider to control the resource allocation toward each virtual instance of an eNB before each virtual operator customizes scheduling within the allocated resources. In [5], relevant technologies for network slicing are discussed with particular focus on synchronous functions, such as multi-dimensional resource management, dynamic traffic steering, and resource abstraction. A particular architecture for network slicing has been introduced and discussed in the context of the 5G NORMA project [6]. Another network slicing solution considering a gateway-based approach is illustrated in [7], wherein a controller provides application-oriented resource abstraction of the underlying RAN. A capacity broker for slice resources was introduced first by the Third Generation Partnership Project (3GPP) and extensively evaluated in [8] by enabling on-demand slice resource allocation. The infrastructure provider instantiates a network slice by allocating specific resources to a mobile virtual network operator (MVNO), service providers, and vertical segments for a specified time duration. A study that explores the different options of network sharing based on a centralized broker is provided in [9] considering mobility means for redirecting users to other networks, spectrum transfer policies, and the application of resource virtualization. Finally, [10] discusses a dynamic slicing scheme that flexibly schedules radio resources based on the requested service level agreement (SLA), while maximizing the user rate and applying fairness criteria.

Our Contribution

This work elaborates on the fundamental pillars for efficient utilization of the concept of network slicing in mobile networks, based on the mobile network architecture framework investigated in the research project 5G NORMA [5]. Particular focus is put on the basic architectural principles for accommodating network slicing in the 5G ecosystem as well as on RAN and core network (CN) aspects. In this regard, we underline the key elements that enable the coexistence of dedicated and shared slices within a common network architecture, and elaborate on the implementation of the notion of network slicing in the RAN and in CN, putting particular emphasis on the concept of software defined mobile network control (SDM-C).

Mobile Network Slicing Architecture

Dedicated and Shared Sub-Slices

Network slices operate on top of a partially shared infrastructure, which is composed of generic hardware resources such as network functions virtualization infrastructure (NFVI) resources, as well as dedicated hardware such as network elements in the RAN. Network functions running on NFVI resources (referred to as virtual network functions, VNFs) are typically instantiated in a customized manner for each network slice. However, this approach cannot be applied to network functions (NFs) relying on dedicated hardware. Therefore, a key issue for network slicing is the identification and design of common NFs, which are either physical or virtual and have to be shared by multiple end-to-end slices.

Examples of common NFs include distributed and monolithic eNBs, and the radio scheduler in the RAN domain. In the CN domain, candidates for shared VNF instances include the home subscribe server (HSS) and mobility management. Generally, three solution groups are discussed with varying levels of common functionality in 3GPP standards [11]: Group A is characterized by a common RAN and completely dedicated CN slices, that is, independent subscription, session, and mobility management for each network slice handling the UE. Group B also assumes a common RAN, where identity, subscription, and mobility management are common across all net-

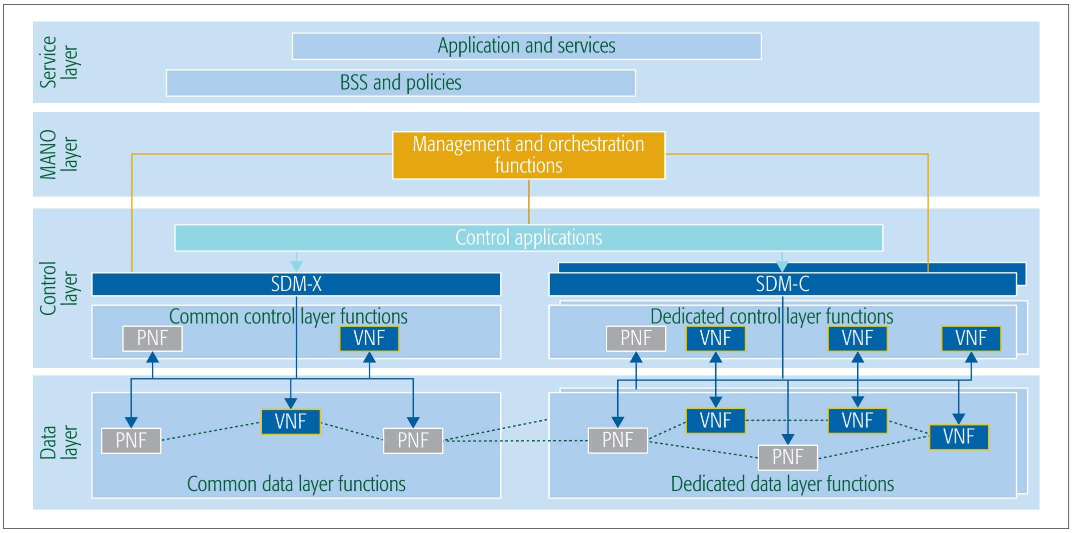


Figure 2. Combining dedicated and shared sub-slices to form e2e mobile network instances.

work slices, while other functions such as session management reside in individual network slices. Group C assumes a completely shared RAN and a common CN control plane, while CN user planes belong to dedicated slices.

In line with the above grouping considered by 3GPP [11], the framework of the 5G NORMA project [5] introduces dedicated network functions, which together form a dedicated sub-slice and are controlled by the software-defined mobile network controller (SDM-C). As illustrated in Fig. 2, shared network functions are aggregated in common sub-slices that are controlled by the SDM coordinator (SDM-X), reflecting the fact that these functions have to coordinate and, if necessary, prioritize the quality of service (QoS) requirements of multiple slices.

End-to-End Network Slicing: Common and Dedicated Network Functions

When sharing NFs and resources between distinct network slices, a central entity in charge of managing and controlling the process is needed (i.e., the SDM-X). This entity ensures attaining high resource efficiency while guaranteeing individual SLAs. Based on the SDM-C paradigm, this entity resides on the common control layer; it also includes NFs, either virtual or physical, on which the network slices rely. While a fixed splitting of common NFs (and resources) simplifies the network management and operation, it may lead to inefficient network utilization. Conversely, dynamic adjustments of common resources might bring multiplexing gains at the expense of less determinism. Hence, the main objective of the SMD-X is to properly administer the trade-off between flexible and static resource assignments by taking into account sharing policies set by the service provider.

Let us consider the system spectrum as a shared resource pool (divided into several resource blocks, RBs) fully managed by the SDM-X. The flexibility introduced by the SDM-X enables dynamic and short-term scheduling decisions based on slice requirements. Specifically, the SDM-X facilitates a "masked" view of the shared resource pool toward the network slices. The resource mask is defined as a group of physical RBs dynamically assigned to each network slice. The advantage of this solution relies on the SDM-X channel monitoring phase and on the subsequent dynamic adjustment of slice resource masks needed to cope with the fast channel dynamics. In a multi-tenancy context [5], a dedicated resource scheduler per tenant may be directly connected to the SDM-X interface, acting as an SDN application. The scheduler uses the slice resource mask and applies its own scheduling policies, while preserving slice isolation constraints. The SDM-X plays a key role in assigning priority to network slices: Different objective functions can be dynamically implemented in order to achieve fairness, maximize spectral efficiency, and mitigate interference.

Implementation of RAN and CN Slicing

Realization of Network Slicing in CN and RAN

Figure 3 illustrates how RAN slicing can be realized such that existing and well proven principles of radio access are utilized. In this regard, the network slice selection function (NS-SF), which is part of the SDM-X concept (Fig. 2), is responsible for selecting the appropriate slice per user. In addition, it configures the RAN-CN interface such that the control and user plane traffic is routed to the accordingly configured functional elements in the CN slice. The user plane anchor (UP-Anchor) is responsible for distributing the traffic according to the configured slice policy, and for encryption with slice-specific security keys.

Radio resource management and control in the base station, and correspondingly in the UE if several slices are configured, is responsible for configuring the RAN protocol stack and QoS according to the slice requirements. For example, for a slice with high throughput requirements, radio bearers are configured to support multi-connectivity (MC), for example, similar to the split bearer approach as in LTE dual connectivity or the equivalent in 5G. For slices with low latency and high robustness requirements, lower frame error rates as well as multi-point diversity techniques may be utilized.

In the example illustrated in Fig. 3, the radio flow in network slice A, which could correspond to a radio bearer in LTE, is configured with two

When sharing NFs and resources between distinct network slices, a central entity in charge of managing and controlling the process is needed (i.e., the SDM-X). This entity ensures attaining high resource efficiency while guaranteeing individual SLAs. Based on the SDM-C paradigm, this entity resides on the common control layer; it also includes NFs, either virtual or physical, on which the network slices rely.

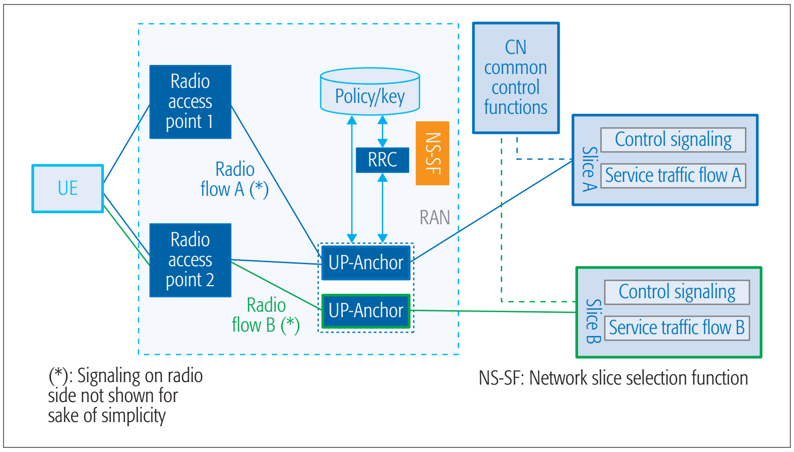


Figure 3. Multi-connectivity anchor — the interface between network slicing and RAN.

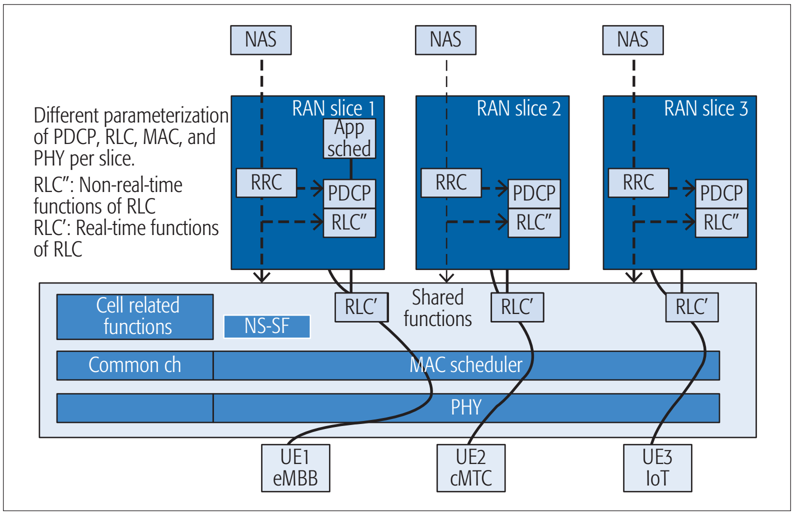


Figure 4. Example of common spectrum shared by multiple slices.

radio connections, while network slice B is configured with only one connection according to the provided policy configuration. In summary, network slicing can be realized by appropriate mapping control and configuration of radio network functions without changing fundamental paradigms of the RAN.

Multiplexing Network Slices in RAN

The RAN is a typical example of a shared network function controlled by a single authority, where spectrum is shared among mobile virtual network and service operators. Figure 4 illustrates an example of a common spectrum shared by three network slices, each with its own RAN and CN part. The layer 2 control plane is split into cell related functions that are common to all slices, and session or user-specific radio resource control (RRC). Depending on the underlying service, RRC can configure and tailor the user plane protocol stack. For example, for a slice supporting low-delay services, IP and related header compression (HC) may not be used, and RLC can be configured in transparent mode. In contrast, for services requiring high quality of experience (QoE) and excellent QoS, IP as well as acknowledged RLC must be initiated. In addition, there would be the possibility to chain proprietary and operator-specific functions within a network slice. In this regard, the intra-slice application scheduler (which prioritizes sessions within the related slice) is chained in RAN slice 1, while the inter-slice radio scheduler (which schedules different slices) resides in the common RAN part (Fig. 4) and makes use of multi-service scheduling capabilities. Multi-service scheduling is part of a flexible RAN and provides the capabilities to differentiate traffic classes and to assign resources according to QoS requirements. Hence, service flows from different slices can be individually treated; for example, flexible numerologies can be used to fulfil QoS constraints and even semi-persistently reserved resources for deterministic traffic requirements.

It is worth mentioning that although the higher RAN layers can be configured to operate in a slice-specific mode relatively easily, this is not the case for the lower-layer radio interface. In contrast to current 3GPP LTE, where radio slices are represented by new variants of 3GPP such as narrowband Internet of Things (NB-IOT), 5G requires the inherent coexistence of diverse services. Hence, in contrast to 4G LTE where adding a new radio slice requires modification to the legacy LTE radio, the new radio proposed for 5G is designed to be forward compatible [11], among other ways by utilizing new radio framing and protocols. This means that future addition of new services and thus radio slices will not require changes in the 5G radio framework.

In a similar context, it is worth pointing out that the new radio framing involves the so-called tiling concept [12]. That is, time and frequency resources of the new 5G radio are tiled so that they it can be allocated for the needs of certain slices with given requirements. An illustration of the tiling concept is provided in Fig. 5.

Exemplary Architecture with Shared RAN Slices

An exemplary architecture with shared RAN slices is presented in Fig. 6a, which shows how the different aspects may integrate, based on three options [13]:

1. Option 1 shows two network slices where each slice carries two different services. Each slice may be operated by a different MNO. Furthermore, for each slice an individual RAN protocol stack is implemented down to the upper part of the physical layer. Only the lower part of the physical layer is shared across slices. The multiplexed access to the transponder part of the physical layer (PHY-TP) is coordinated by the SDM-X, which makes use of flexible and efficient radio resource management for supporting different numerologies within the same spectrum. One could think of option 1 as implementing all user-specific functions such as forward error correction encoding, layer mapping, and precoding in an individual fashion, while TP-specific functionality such as transmission of synchronization and cell-specific reference signals are shared.

2. Option 2 again depicts two network slices from two operators. Compared to the previous example, each slice uses an individual implementation of service-specific functionality such as Packet Data Control Protocol (PDCP), radio link control, RLC, and slice-specific radio resource control (RRC). In addition, the tenant may implement customized QoS scheduling to perform

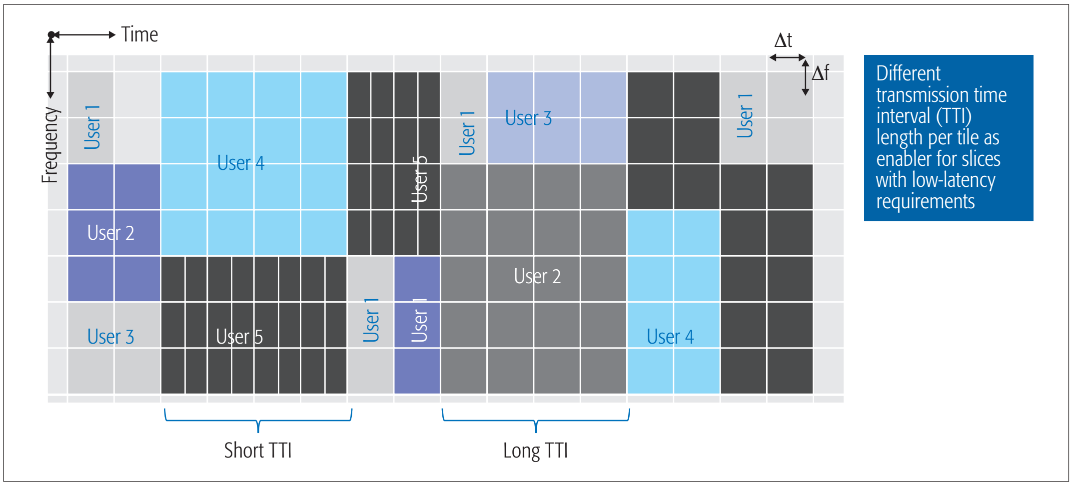


Figure 5. The radio tiling concept proposed for 5G radio with network slicing.

Those multiple PHY interfaces are leveraged to deliver enhanced performance capabilities, which are translated into aggregated throughput or increased reliability. A major challenge is to enforce different QoS requirements, differentiation, and prioritization within a RAN exploiting MC and Multi-RAT through a single scheduler.

pre-scheduling. The access to the MAC layer is then controlled by the SDM-X where resource fairness across tenants and QoS guarantees corresponding to individual SLAs must be met. Furthermore, resource isolation must be provided to alleviate side-effects.

3. Option 3 illustrates the case of two operators using the same RAN as shared resource, that is, the SDM-X is the interface between CN and RAN. In this example, no customization of radio resource management beyond SDM-X parameters and configuration would be possible.

Flexible RAN Technologies as Enablers for Shared RAN

In addition to the flexible architecture considerations mentioned above, further flexible RAN technologies enable a shared RAN for network slicing and accommodating highly diverse services. In the following, some of them are briefly explained and how they facilitate network slicing.

Multi-connectivity (MC): The term RAN MC refers to the versatile scenario where a UE connects to the network via multiple cells. For the sake of the current explanation, it suffices to consider that a multi-connectivity approach takes place whenever the connection of the UE to the RAN involves multiple PHY interfaces. Those multiple PHY interfaces are leveraged to deliver enhanced performance capabilities, which are translated into aggregated throughput or increased reliability. A major challenge is to enforce different QoS requirements, differentiation, and prioritization within a RAN exploiting MC and multi-RAT through a single scheduler.

Next, we consider two MC options, the *common PDCP* and *common MAC approach*, which are shown in Fig. 6b, as part of the exemplary architecture option 2 discussed earlier. The *common PDCP approach* dictates that the PDCP layer of the protocol stack is shared between the individual connections of the RAN multi-connectivity (henceforth called "radio leg"), and all layers below PDCP are separate logical entities. This approach resembles that of dual connectivity in 3GPP LTE, and offers the advantage of flexibility in terms of the physical location of the protocol stack layers. The main advantage of the common PDCP approach is the flexibility it offers in terms of the physical location of the protocol stack layers. In particular, since the interface between PDCP and RLC is not a time-critical interface, the common PDCP layer is not necessarily co-located with RLC; hence, mobility-related signaling can be hidden from the CN. In the *common MAC approach*, the multi-connectivity anchor point is the MAC layer, similarly as carrier aggregation in 3GPP LTE. Owing to the time-critical interface between MAC and PHY, the common MAC approach requires that either the multi-connectivity legs originate from the same site, or they are interconnected via a high-capacity transport link. Nevertheless, the common MAC approach offers the advantage of fast information exchange between the different multi-connectivity legs. This facilitates coordinated scheduling, interference mitigation, and other schemes related to MAC scheduling [14].

Multi-RAT and millimeter wave (mmWave) technology: It is envisioned that mmWave technology will play a key role in the fulfillment of 5G network requirements. MC will be an essential requirement to support mmWave deployments, which are anticipated to cover both mobile broadband and machine-type applications. The design characteristics of such deployments will depend on factors that span a wide area of architecture requirements, such as transport capabilities, low-band integration, propagation impairments, and (edge or core) cloud implementations. Consequently, a flexible architecture incorporating mmWave support is required to meet different slice requirements.

User-centric signaling: User-centric signaling and mobility management for services including short, sporadic, and delay-tolerant data packets is proposed based on a user-centric connection area (UCA) [15]. The UCA consists of a set of radio nodes selected by the flexible 5G-RAN. One radio node acts as an anchor node within the UCA, which shares the user context with all other nodes within the UCA. The CN connections (bearers) are terminated at the anchor node. With the help of a shared context, mobility is managed by the RAN instead of the CN as long as the UE moves within the UCA. This implies that mobility is hidden from the CN, which reduces mobility and connection related signaling. Based on the con-

Based on this analysis, a strong potential of network slicing was revealed for addressing the diverse requirements of future 5G systems. Nevertheless, network slicing remains still at an early stage in terms of its development, hence one should anticipate a long way before it becomes a mature technology and thus be adopted by network standards.

text sharing, the UE is able to send uplink packets and receive downlink packets by any node within a UCA. The user-specific aspect provides flexibility and reconfigurability in the realization of a UCA; that is, each UCA can be configured according to specific requirements taking into account QoS parameters.

Mobile edge computing and edge cloud processing: Advanced 5G services are envisioned to be offered at the network edge so as to reside much closer to the user in order to enhance delay and perceived performance, for example., adopting the European Telecommunications Standards Institute (ETSI) mobile edge computing (MEC) paradigm.[1] Therefore, a flexible service chaining should also be improved to establish dynamic services considering edge network locations and might be combined with VNFs to ensure joint optimization of services and networking operations. Edge server locations can also be exploited for storage, computation, and dynamic service creation within a given network slice by verticals and over-the-top providers, introducing another multi-tenancy dimension.

Conclusions and Further Challenges

An overview of the basic implementation features of network slicing was presented, along with its potential to provide revenue to the network operator. The analysis included the basic principles behind the mapping of dedicated and shared slices, as well as implementation-specific aspects when the concept of network slicing is employed over RAN and CN. Special focus was put on the connection of network slicing with RAN concepts. Based on the above analysis, strong potential of network slicing was revealed for addressing the diverse requirements of future 5G systems. Nevertheless, network slicing remains at an early stage in terms of its development; hence, one should anticipate a long way before it becomes a mature technology and thus be adopted by network standards.

Acknowledgment

This work has been performed in the framework of the H2020-ICT-2014-2 project 5G NORMA. The authors would like to acknowledge the contributions of their colleagues. This information reflects the consortium's view, but the consortium is not liable for any use that may be made of any of the information contained therein.

References

[1] A. Frotzscher *et al.*, "Requirements and Current Solutions of Wireless Communication in Industrial Automation," *IEEE ICC*, Sydney, Australia, June 2014, pp. 67–72.
[2] NGMN Alliance, "Description of Network Slicing Concept," https://www.ngmn.org/uploads/media/160113_Network_Slicing_v1_0.pdf, accessed Jan. 2017.
[3] G. Fettweis, "The Tactile Internet: Applications and Challenges," *IEEE Vehic. Tech. Mag.*, vol. 9, no. 1, Mar. 2014, pp. 64–704.
[4] R. Kokku *et al.*, "NVS: A Virtualization Substrate for WiMAX Networks," *Int'l. Conf. Mobile Computing and Networking*, Chicago, IL, Sept. 2010, pp. 1333–46.
[5] EU H2020 METIS-II, "Draft Synchronous Control Functions and Resource Abstraction Considerations," Deliv. D5.1, https://metis-ii.5g-ppp.eu/wp-content/uploads/deliverables/METIS-II_D5.1_V1.0.pdf, May 2016, accessed Jan. 2017.
[6] EU H2020 5G NORMA, "5G NORMA Network Architecture — Intermediate Report," https://5gnorma.5g-ppp.eu/wp-content/uploads/2016/11/5g_norma_d3-2.pdf, Deliv. D3.2, Jan 2017, accessed Jan. 2017.
[7] J. He, W. Song, "App RAN: Application-Oriented Radio Access Network Sharing in Mobile Networks," *IEEE ICC*, London, U.K., May 2015, pp. 3788–94.
[8] K. Samdanis, X. Perez-Costa, and V. Sciancalepore, "From Network Sharing to Multi-Tenancy: The 5G Network Slice Broker," *IEEE Commun. Mag. Standards Supp.*, vol. 54, no. 7, July 2016, pp. 32–39.
[9] J. Panchal, R. Yates, and M. Buddhikot, "Mobile Network Resource Sharing Options: Performance Comparisons," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, Sept. 2013, pp. 4470–82.
[10] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network Slicing Management & Prioritization in 5G Mobile System," *Euro. Wireless Conf.*, Oulu, Finland, May 2016, pp. 197–202.
[11] 3GPP, "TR23.799, V0.7.0, Technical Specification Group Services and System Aspects; Study on Architecture for Next Generation System," Aug. 2016.
[12] K. I. Pedersen *et al.*, "A Flexible 5G Frame Structure Design for Frequency-Division Duplex Cases," *IEEE Commun. Mag.*, vol. 54, no. 3, Mar. 2016, pp. 53–59.
[13] EU H2020 5G NORMA, "RAN Architecture Components — Intermediate Report," Deliv. D4.1, https://5gnorma.5g-ppp.eu/wp-content/uploads/2016/12/5g_norma_d4-1.pdf, Nov. 2016 accessed Jan. 2017.
[14] A. Ravanshid *et al.*, "Multi-Connectivity Functional Architectures in 5G," *2016 IEEE ICC Wksp.*, Kuala Lumpur, Malaysia, 2016, pp. 187–92.
[15] D. Aziz *et al.*, "Signaling Minimization Framework for Short Data Packet Transmission in 5G," *IEEE VTC-Fall*, Montreal, Canada, Sept. 2016.

Biographies

Peter Rost [SM] (peter.m.rost@nokia-bell-labs.com) received his Ph.D. degree from Technische Universität Dresden in 2009 and his M.Sc. degree from the University of Stuttgart in 2005. Since May 2015, he has been a member of the Radio Systems research group at Nokia Germany, contributing to the European projects 5G-NORMA and METIS-II, and business unit projects on 5G architecture. He serves as a member of IEEE ComSoc GITC, VDE ITG Expert Committee on Information and System Theory, and as Executive Editor of *IEEE Transactions on Wireless Communications*.

Christian Mannweiler (christian.mannweiler@nokia-bell-labs.com) received his M.Sc. (Dipl.-Ing.) and Ph.D. (Dr.-Ing.) degrees from the University of Kaiserslautern, Germany, in 2008 and 2014, respectively. Since 2015, he has been a member of the Cognitive Network Management research group at Nokia Bell Labs Germany, where he has been working in the area of network management automation and SON for 5G systems. He has contributed to several nationally and EU funded projects covering architecture design of mobile communication systems, among them 5G NORMA.

Diomidis S. Michalopoulos [SM] (diomidis.michalopoulos@nokia-bell-labs.com) received his Ph.D. from Aristotle University, Greece, in 2009. From 2009 to 2015 he was with the University of British Columbia, Canada, as a Killam and Banting postdoctoral fellow, and FAU, Germany, as a researcher and teaching instructor. In 2015 he joined Nokia Bell Labs, Germany, contributing toward radio and architecture aspects of 5G. He has received several awards for academic achievement including the Marconi Young Scholar Award, and served as an Editor of *IEEE Communications Letters*.

Cinzia Sartori (cinzia.sartori@nokia-bell-labs.com) is a principal expert in the field of end-to-end mobile network architecture with focus on 5G at NOKIA, Munich. Until mid-2013 she led the Self-Organizing Network (SON) Research and Standardization project in Nokia Siemens Networks. She is co-author of the book *LTE Self-Organizing Network* (January 2012). Earlier she worked in network telecom, O&M, RRM, and SS7 in Nokia Siemens Networks, Siemens, and GTE. She graduated as an engineer in Pavia -Italy.

Vincenzo Sciancalepore [S'11, M'15] (vincenzo.sciancalepore@neclab.eu) received his M.Sc. degree in telecommunications engineering and telematics engineering in 2011 and 2012, respectively, and a double Ph.D. degree in 2015. From 2011 to 2015, he was a research assistant with IMDEA Networks, where he was involved in intercell coordinated scheduling for LTE-Advanced networks and device-to-device communication. He is currently a research scientist with NEC Europe Ltd., Heidelberg, where he is involved in network virtualization and network slicing challenges.

Nishanth Sastry (nishanth.sastry@kcl.ac.uk) is a senior lecturer at King's College London. He holds a Ph.D. from the University

[1] http://www.etsi.org/technologies-clusters/technologies/mobile-edge-computing

of Cambridge, United Kingdom, a Master's degree from the University of Texas at Austin, and a Bachelor's degree from Bangalore University, India, all in computer science. He has over six years of experience in the industry (Cisco Systems, India, IBM Software Group) and Industrial Research Labs (IBM TJ Watson Research Center).

Oliver Holland [M] (oliver.holland@kcl.ac.uk) is a senior researcher at King's College London, working on spectrum sharing and 5G technologies among others. He has an established record in various standards activities, including leadership of various standards, and has undertaken project, academic, event, and other leaderships, within the IEEE and elsewhere. He has achieved various accolades for his work, most recently creating and leading a team/submission/idea that won the 0.5m EUR European Union Collaborative Spectrum Sharing Prize.

Shreya Tayade (tayade@eit.uni-kl.de) received her B.E. from Government College of Engineering Aurangabad, India, in 2010 and her M.Sc. degree in communication engineering from RWTH Aachen University, Germany, in 2015. Since 2015, she has been a researcher at the Technical University of Kaiserslautern Germany and is working on 5G-based EU and industrial projects.

Bin Han (binhan@eit.uni-kl.de) received his B.E. (2009) and M.Sc. (2012) degrees from Shanghai Jiao Tong University and the Technische Universität Darmstadt, respectively. In 2016 he was granted a Ph.D. (Dr.-Ing.) degree in electrical engineering and information technology from the Karlsruhe Institute of Technology. He is currently a postdoctoral researcher at the University of Kaiserslautern, researching in the broad area of communication systems and signal processing, with a special focus on 5G.

Dario Bega (dario.bega@imdea.org) received his B.Sc. and M.Sc. degrees in telecommunication engineering in 2010 and 2013, respectively, from the University of Pisa, Italy. Since 2015, he has been a research assistant with Imdea Networks in Madrid and a Ph.D. student in telematic engineering with Universidad Carlos III de Madrid. His research work focuses on wireless networks and the multi-tenancy approach for 5G.

Danish Aziz (danish.aziz@nokia-bell-labs.com) is a research engineer at Nokia Bell Labs Germany. He received his M.Sc. and Ph.D. (Dr.-Ing.) degrees in communications engineering from the University of Stuttgart, Germany. He started his career in 2006 with the development of 4G-LTE systems. He is currently involved in the development of flexible radio access network architecture for 5G. He has worked in several EU-funded projects (e.g., METIS-I). He is actively contributing to the ongoing 5G-NORMA project.

Hajo Bakker (hajo.bakker@nokia-bell-labs.com) is team leader within End-To-End Mobile Networks Research at Nokia Bell Labs, Stuttgart, Germany. He received his Dipl.-Ing. degree in electrical engineering telecommunication from the Technical University of Brunswick, Germany, and joined Alcatel research in 1986. He participated in the 3GPP RAN 2 working group for the standardization of LTE. He is the author/co-author of more than 20 scientific publications, and holds more than 80 granted patents and patent applications.

Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges

Jose Ordonez-Lucena, Pablo Ameigeiras, Diego Lopez, Juan J. Ramos-Munoz, Javier Lorca, and Jesús Folgueira

The authors present the network slicing concept, with a particular focus on its application to 5G systems. They start by summarizing the key aspects that enable the realization of so-called network slices. Then they give a brief overview on the SDN architecture proposed by the ONF and show that it provides tools to support slicing. They argue that although such architecture paves the way for network slicing implementation, it lacks some essential capabilities that can be supplied by NFV.

ABSTRACT

The fifth generation of mobile communications is anticipated to open up innovation opportunities for new industries such as vertical markets. However, these verticals originate myriad use cases with diverging requirements that future 5G networks have to efficiently support. Network slicing may be a natural solution to simultaneously accommodate, over a common network infrastructure, the wide range of services that vertical-specific use cases will demand. In this article, we present the network slicing concept, with a particular focus on its application to 5G systems. We start by summarizing the key aspects that enable the realization of so-called network slices. Then we give a brief overview on the SDN architecture proposed by the ONF and show that it provides tools to support slicing. We argue that although such architecture paves the way for network slicing implementation, it lacks some essential capabilities that can be supplied by NFV. Hence, we analyze a proposal from ETSI to incorporate the capabilities of SDN into the NFV architecture. Additionally, we present an example scenario that combines SDN and NFV technologies to address the realization of network slices. Finally, we summarize the open research issues with the purpose of motivating new advances in this field.

INTRODUCTION

Fifth generation (5G) systems are nowadays being investigated to satisfy the consumer, service, and business demands of 2020 and beyond. One of the key drivers of 5G systems is the need to support a variety of vertical industries such as manufacturing, automotive, healthcare, energy, and media and entertainment [1]. Such verticals originate very different use cases, which impose a much wider range of requirements than existing services do nowadays. Today's networks, with their "one-size-fits-all" architectural approach, are unable to address the diverging performance requirements that verticals impose in terms of latency, scalability, availability, and reliability. To efficiently accommodate vertical-specific use cases along with increased demands for existing services over the same network infrastructure, it is accepted that 5G systems will require architectural enhancements with respect to current deployments.

Network softwarization, an emerging trend that seeks to transform networks using software-based solutions, can be a potential enabler for accomplishing this. Through technologies like software-defined networking (SDN) and network functions virtualization (NFV), network softwarization can provide the programmability, flexibility, and modularity that is required to create multiple logical (virtual) networks, each tailored for a given use case, on top of a common network. These logical networks are referred to as network slices. The concept of separated virtual networks deployed over a single network is indeed not new (e.g., virtual private networks, VPNs); however, there are specificities that make network slices a novel concept. We define network slices as end-to-end (E2E) logical networks running on a common underlying (physical or virtual) network, mutually isolated, with independent control and management, which can be created on demand. Such self-contained networks must be flexible enough to simultaneously accommodate diverse business-driven use cases from multiple players on a common network infrastructure (Fig. 1).

In this article, we provide a comprehensive study of the architectural frameworks of both SDN and NFV as key enablers to achieve the realization of network slices. Although these two approaches are not yet commonplace in current networking practice, especially in public wide area networks (WANs), their integration offers promising possibilities to adequately meet the slicing requirements. Indeed, many 5G research and demonstration projects (e.g., 5GNORMA, 5GEx, 5GinFIRE, and 5G!Pagoda) are addressing the realization of 5G slicing through the combination of SDN and NFV. Thus, we present a deployment example that illustrates how NFV functional blocks, SDN controllers, and their interactions can fully realize the network slicing concept. Furthermore, we identify the main challenges arising from implementing network slicing for 5G systems.

The remainder of this article is organized as follows. We commence by providing a background on key concepts for network slicing. Then we describe the SDN architecture from the Open Networking Foundation (ONF) and the NFV architecture from the European Telecommunications Standards Institute (ETSI), respectively. Following that, we show a network slicing use case with NFV and SDN integration. Finally, we provide the main challenges and future research directions.

Digital Object Identifier: 10.1109/MCOM.2017.1600935

Jose Ordonez-Lucena, Pablo Ameigeiras, and Juan J. Ramos-Munoz are with the University of Granada; Diego Lopez, Javier Lorca, and Jesús Folgueira are with Telefónica I+D – Global CTO

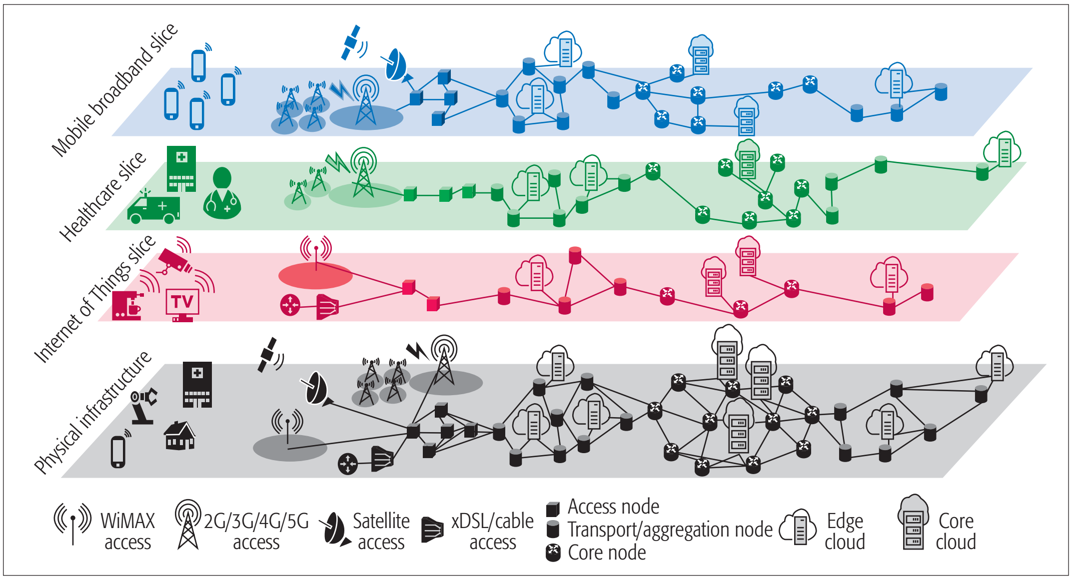


Figure 1. 5G network slices running on a common underlying multi-vendor and multi-access network. Each slice is independently managed and addresses a particular use case.

We define network slices as end-to-end (E2E) logical networks running on a common underlying (physical or virtual) network, mutually isolated, with independent control and management, which can be created on demand. Such self-contained networks must be flexible enough to simultaneously accommodate diverse business-driven use cases from multiple players on a common network infrastructure.

Background on Key Concepts for Network Slicing

In this section, we provide a background on key aspects that are necessary to realize the network slicing concept.

Resources

In its general sense, a resource is a manageable unit, defined by a set of attributes or capabilities that can be used to deliver a service. A network slice is composed of a collection of resources that, appropriately combined, meet the service requirements of the use case that such a slice supports. In network slicing, we consider two types of resources.

Network Functions (NFs): Functional blocks that provide specific network capabilities to support and realize the particular service(s) each use case demands. Generally implemented as software instances running on infrastructure resources, NFs can be physical (a combination of vendor-specific hardware and software, defining a traditional purpose-built physical appliance) and/or virtualized (network function software is decoupled from the hardware it runs on).

Infrastructure Resources: Heterogeneous hardware and necessary software for hosting and connecting NFs. They include computing hardware, storage capacity, networking resources (e.g., links and switching/routing devices enabling network connectivity), and physical assets for radio access. Suitable for use in network slicing, the aforementioned resources and their attributes have to be abstracted and logically partitioned leveraging virtualization mechanisms, defining virtual resources that can be used in the same way as physical ones.

Virtualization

Virtualization is a key process for network slicing as it enables effective resource sharing among slices. Virtualization is the abstraction of resources using appropriate techniques. Resource abstraction is the representation of a resource in terms of attributes that match predefined selection criteria while hiding or ignoring aspects which are irrelevant to such criteria, in an attempt to simplify the use and management of that resource in some useful way. The resources to be virtualized can be physical or already virtualized, supporting a recursive pattern with different abstraction layers.

Just as server virtualization [2] makes virtual machines (VMs) independent of the underlying physical hardware, network virtualization [3] enables the creation of multiple isolated virtual networks that are completely decoupled from the underlying physical network and can safely run on top of it.

The introduction of virtualization to the networking field enables new business models, with novel actors and distinct business roles. We consider a framework with three kinds of actors:

- Infrastructure provider (InP): owns and manages a given physical network and its constituent resources. Such resources, in the form of WANs and/or data centers (DCs), are virtualized and then offered through programming interfaces to a single or multiple tenants.
- Tenant: leases virtual resources from one or more InPs in the form of a virtual network, where the tenant can realize, manage, and provide network services to its users. A network service is a composition of NFs, and it is defined in terms of the individual NFs and the mechanism used to connect them.
- End user: consumes (part of) the services supplied by the tenant, without providing them to other business actors.

As discussed above, virtualization is naturally recursive, and the first two actors can happen in a vertical multi-layered pattern, where a tenant at one layer acts as the InP at the layer immediately above. The recursion mentioned here implies that a tenant can provide network services to an end user, but also to another tenant (Fig. 2). In such a case, this second tenant would provide more advanced network services to its own users.

Orchestration

Orchestration is also a key process for network slicing. In its general sense, orchestration can be defined as the art of both bringing together and

In its general sense, orchestration can be defined as the art of both bringing together and coordinating disparate things into a coherent whole. In a slicing environment, where the players involved are so diverse, an orchestrator is needed to coordinate seemingly disparate network processes for creating, managing and delivering services.

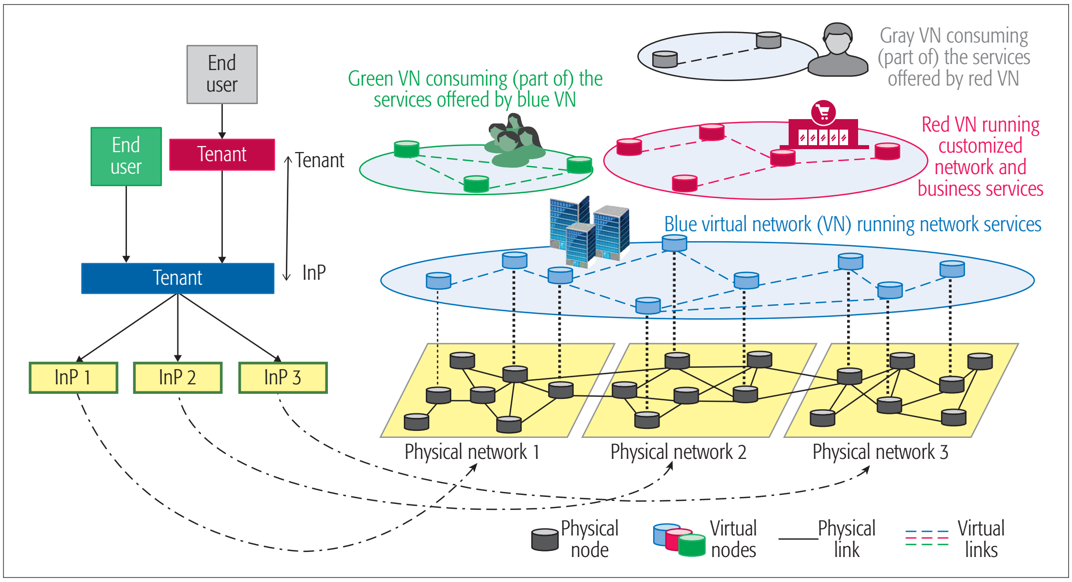


Figure 2. InPs and tenants as virtualization actors. This scenario shows the recursion principle, where these actors happen in a vertical multi-layered pattern.

coordinating disparate things into a coherent whole. In a slicing environment, where the players involved are so diverse, an orchestrator is needed to coordinate seemingly disparate network processes for creating, managing, and delivering services.

A unified vision and scope of orchestration has not been agreed upon. According to the ONF [4], orchestration is defined as the continuing process of selecting resources to fulfill client service demands in an optimal manner. The idea of optimal refers to the optimization policy that governs orchestrator behavior, which is expected to meet all the specific policies and service level agreements (SLAs) associated with clients (e.g., tenants or end users) that request services. The term continuing means that available resources, service demands, and optimization criteria may change in time. Interestingly, orchestration is also referred to in [4] as the defining characteristic of an SDN controller. Note that client is a term used in the SDN context.

The ONF states that the orchestrator functions include client-specific service demand validation, resource configuration, and event notification. For a more detailed description of these functions, see [4, Sec. 6.2].

However, in network slicing, orchestration cannot be performed by a single centralized entity, not only because of the complexity and broad scope of orchestration tasks, but also because it is necessary to preserve management independence and support the possibility of recursion. In our view, a framework in which each virtualization actor (Fig. 2) has an entity performing orchestration functions seems more suitable to satisfy the above requirements. The entities should exchange information and delegate functionalities between them to ensure that the services delivered at a certain abstraction layer satisfy the required performance levels with optimal resource utilization.

Isolation

Strong isolation is a major requirement that must be satisfied to operate parallel slices on a common shared underlying substrate. The isolation must be understood in terms of:

Performance: Each slice is defined to meet particular service requirements, usually expressed in the form of key performance indicators (KPIs). Performance isolation is an E2E issue, and has to ensure that service-specific performance requirements are always met on each slice, regardless of the congestion and performance levels of other slices.

Security and privacy: Attacks or faults occurring in one slice must not have an impact on other slices. Moreover, each slice must have independent security functions that prevent unauthorized entities to have read or write access to slice-specific configuration/management/accounting information, and be able to record any of these attempts, whether authorized or not.

Management: Each slice must be independently managed as a separate network.

To achieve isolation, a set of appropriate, consistent policies and mechanisms have to be defined at each virtualization level, following the recursion principle introduced earlier. The policies (what is to be done) contain lists of rules that describe how different manageable entities must be properly isolated, without delving into how this can be achieved. The mechanisms (how it is to be done) are the processes that are implemented to enforce the defined policies. From our point of view, to fully realize the required isolation level, the interplay of both virtualization and orchestration is needed.

ONF Network Slicing Architecture

The SDN architecture provided by the ONF comprises an intermediate control plane that dynamically configures and abstracts the underlying forwarding plane resources so as to deliver tailored services to clients located in the application plane (see the SDN basic model in [4]). This is well aligned with the requirements of 5G network slicing, which needs to satisfy a wide range of service demands in an agile and cost-effective manner. Thus, the SDN architecture is an appropriate tool for supporting the key principles of slicing. The purpose of this section is to describe the SDN architecture and how it can be applied to enable slicing in 5G systems.

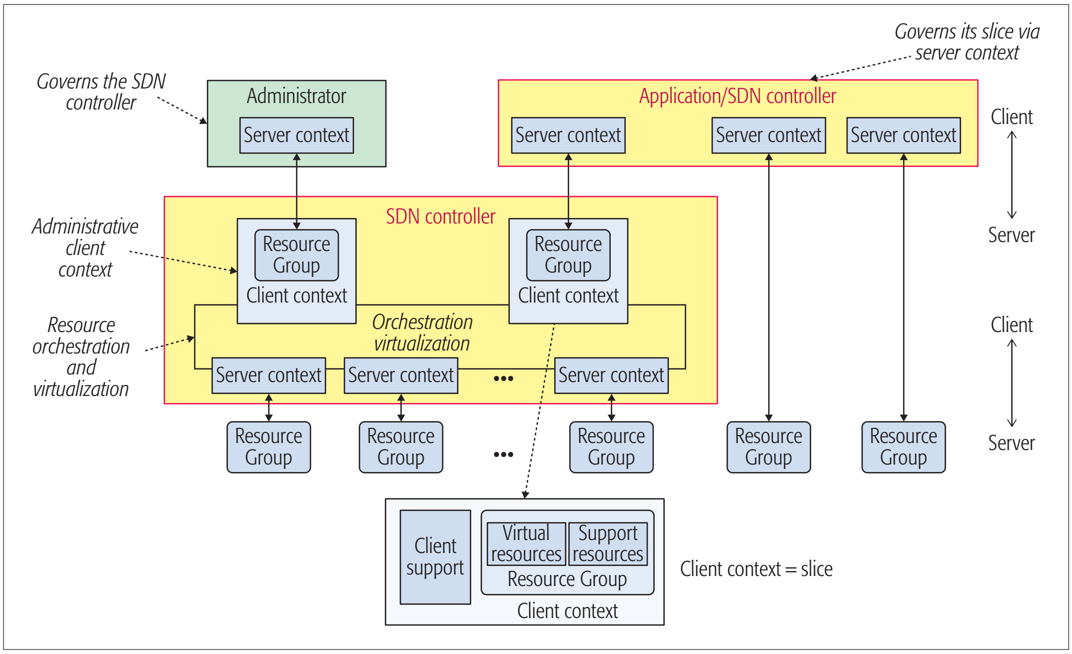


Figure 3. ONF SDN network slicing architecture [5].

Through orchestration, the SDN controller optimally dispatches the selected resources to such separate Resource Groups. The interplay of both controller functions enables the fulfillment of the diverging service demands from all clients while preserving the isolation among them.

According to [4], the major SDN architectural components are resources and controllers. For SDN, a resource is anything that can be utilized to provide services in response to client requests. This includes infrastructure resources and NFs, but also network services, in application of the recursion principle described earlier. A controller is a logically centralized entity instantiated in the control plane which operates SDN resources at runtime to deliver services in an optimal way. Therefore, it mediates between clients and resources, acting simultaneously as server and client via client and server contexts, respectively. Both contexts are conceptual components of an SDN controller enabling the server-client relationships (Fig. 3):

Client context: Represents all the information the controller needs to support and communicate with a given client. It comprises a Resource Group and a Client support function. The Resource Group contains an abstract, customized view of all the resources that the controller, through one of its northbound interfaces, offers to the client, in order to deliver on its service demands and facilitate its interaction with the controller. Client support contains all that is necessary to support client operations, including policies on what the client is allowed to see and do [4], and service-related information to map actions between the client and the controller.

Server context: Represents all the information the controller needs to interact with a set of underlying resources, assembled in a Resource Group, through one of its southbound interfaces.

The process of transforming the set of Resource Groups accessed through server contexts to those defined in separate client contexts is not straightforward, and it requires the SDN controller to perform virtualization and orchestration functions.

When performing the virtualization function, the SDN controller carries out the abstraction and the aggregation/partitioning of the underlying resources. Thanks to virtualization, each client context provides a specific Resource Group that can be used by the client associated with that context to realize its service(s). Through orchestration, the SDN controller optimally dispatches the selected resources to such separate Resource Groups. The interplay of both controller functions enables the fulfillment of the diverging service demands from all clients while preserving the isolation among them.

The SDN architecture also includes an administrator. Its tasks consist of instantiating and configuring the entire controller, including the creation of both server and client contexts and the installation of their associated policies.

According to the ONF vision, the SDN architecture naturally supports slicing [5], as the client context provides the complete abstract set of resources (as a Resource Group) and the supporting control logic that constitute a slice, including the complete collection of related client service attributes.

Another key functional aspect that makes SDN architecture ideal to embrace 5G slicing is recursion. Because of the different abstraction layers that the recursion principle enables, the SDN control plane can involve multiple hierarchically arranged controllers that extend the client-server relationships at several levels (Fig. 4). According to these premises, it is evident that SDN can support a recursive composition of slices [5]. This implies that the resources (i.e., Resource Group) a given controller delivers to one of its clients in the form of a dedicated slice (i.e., client context) can, in turn, be virtualized and orchestrated by such a client in the case of being an SDN controller. This way, the new controller can utilize the resource(s) it accesses via its server context(s) to define, scale, and deliver new resources (and hence new slices) to its own clients, which might also be SDN controllers.

NFV Reference Architectural Framework

Although the SDN architecture described above gives a comprehensive view of the control plane functionalities enabling slicing, it lacks capabilities

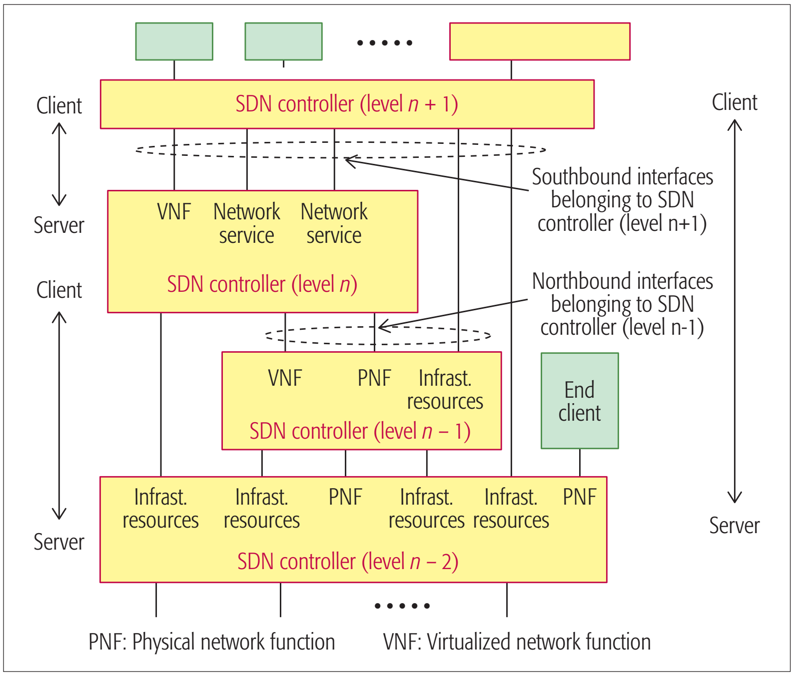


Figure 4. Complex client-server relationships enabled by the recursion in the SDN control plane, adapted from [7].

that are vital to efficiently manage the life cycle of network slices and its constituent resources. In this respect, the NFV architecture [6] is ideal to play this role, as it manages the infrastructure resources and orchestrates the allocation of such resources needed to realize VNFs and network services.

To benefit from the management and orchestration functionalities of NFV, appropriate cooperation between SDN and NFV is required. However, embracing SDN and NFV architectures into a common reference framework is not an easy task [7, 8]. In this section, we briefly describe the tentative framework that ETSI presents in [8] to integrate SDN within the reference NFV architecture. This framework incorporates two SDN controllers, one logically placed at the tenant and another at the InP level. We commence providing a brief overview of the NFV architectural framework, and later describe the integration of the two SDN controllers (Fig. 5).

The NFV architecture comprises the following entities:

Network Functions Virtualization Infrastructure (NFVI): A collection of resources used to host and connect the VNFs. While the broad scope of SDN makes resource a generic concept, the current resource definition in the NFV framework comprises only the infrastructure resources.

VNFs: Software-based implementations of NFs that run over the NFVI.

Management and Orchestration (MANO): Performs all the virtualization-specific management, coordination, and automation tasks in the NFV architecture. The MANO framework [9] comprises three functional blocks:

- Virtualized infrastructure manager (VIM): responsible for controlling and managing the NFVI resources.
- VNF manager (VNFM): performs configuration and life cycle management of the VNF(s) on its domain.
- Orchestrator: According to ETSI, it has two set of functions performed by the Resource Orchestrator (RO) and Network Service Orchestrator (NSO), respectively. The RO orchestrates the NFVI resources across (potentially different) VIMs. The NSO performs the life cycle management of network services using the capabilities provided by the RO and the (potentially different) VNFMs.

Network Management System (NMS): Framework performing the general network management tasks. Although its functions are orthogonal to those defined in MANO, NMS is expected to interact with MANO entities by means of a clear separation of roles [9]. NMS comprises:

- Element management (EM): anchor point responsible for the fault, configuration, accounting, performance, and security (FCAPS) of a VNF.
- Operation/business support system (OSS/BSS): a collection of systems and management applications that network service providers use to provision and operate their network services. In terms of the roles we considered earlier, tenants would run these applications.

The ETSI proposal includes two SDN controllers in the architecture. Each controller centralizes the control plane functionalities and provides an abstract view of all the connectivity-related components it manages. These controllers are:

Infrastructure SDN controller (IC): Sets up and manages the underlying networking resources to provide the required connectivity for communicating the VNFs (and its components [10]). Managed by the VIM, this controller may change infrastructure behavior on demand according to VIM specifications adapted from tenant requests.

Tenant SDN controller (TC): Instantiated in the tenant domain [11] as one of the VNFs or as part of the NMS, this second controller dynamically manages the pertinent VNFs used to realize the tenant's network service(s). These VNFs are the underlying forwarding plane resources of the TC. The operation and management tasks that the TC carries out are triggered by the applications running on top of it (e.g., the OSS).

Both controllers manage and control their underlying resources via programmable southbound interfaces, implementing protocols like OpenFlow, NETCONF, and I2RS. However, each controller provides a different level of abstraction. While the IC provides an underlay to support the deployment and connectivity of VNFs, the TC provides an overlay comprising tenant VNFs that, properly composed, define the network service(s) such a tenant independently manages on its slice(s). These different resource views each controller offers through its interfaces have repercussions on the way they operate. On one side, the IC is not aware of the number of slices that utilize the VNFs it connects, nor the tenant(s) which operate(s) such slices. On the other side, for the TC the network is abstracted in terms of VNFs, without notions of how those VNFs are physically deployed. Despite their different abstraction levels, both controllers have to coordinate and synchronize their actions [8]. Note that the service and tenant concept mentioned here can be extended to higher abstraction layers by simply applying the recursion principle, as shown in Fig. 2.

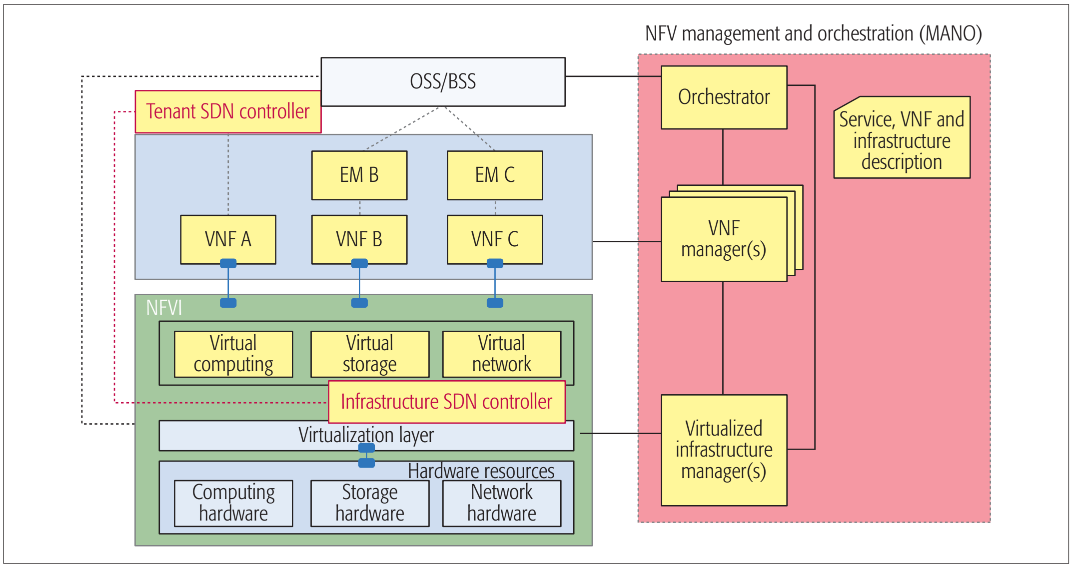


Figure 5. Integrating SDN controllers into the reference NFV architectural framework at the two levels required to achieve slicing [8].

Both VIMs and WIMs act as SDN applications, delegating the tasks related to the management of networking resources to their underlying ICs. Although in this example the ICs are deployed on the NFVI, it would be possible to integrate them into their corresponding VIMs

Network Slicing Use Case with SDN-NFV Integration

In this section, we describe an SDN-enabled NFV deployment example that illustrates the network slicing concept, with several slices running on a common NFVI (Fig. 6). This deployment includes two tenants, each managing a particular set of slices. In the example, we only consider a single level of recursion, and thus the tenants directly serve the end users. Each slice consists of VNFs that are appropriately composed to support and build up the network service(s) the slice (and thus the tenant) delivers to its users. Note that the deployment includes two distinct phases: first, a slice creation phase, in which an end user requests a slice from a network slice catalog, and then the tenant instantiates the slice; and next, a runtime phase, where the different functional blocks within each slice have already been created and are now operative. For simplicity, in Fig. 6 we only depict the runtime phase.

The example considers that the tenants access NFVI resources from three InPs. InP1 provides compute and networking resources, both deployed on two NFVI points of presence (NFVI-PoPs) [12] in the form of DCs. InP2 and InP3 provide SDN-based WAN transport networks, used to communicate such NFVI-PoPs. The VMs and their underlying hardware, instantiated in the NFVI-PoPs and in charge of hosting VNFs (and their components), are directly managed by the VIMs. The networking resources, supporting VM (and hence VNF) connectivity at the infrastructure level, are programmatically managed by the ICs following the VIM and the WAN infrastructure manager (WIM) premises. Both VIMs and WIMs act as SDN applications, delegating the tasks related to the management of networking resources to their underlying ICs. Although in this example the ICs are deployed on the NFVI, it would be possible to integrate them into their corresponding VIMs, as [8] suggests.

On top of the InPs, the tenants independently manage a set of network slices. Each slice comprises an OSS, a TC, and an NSO. The OSS, an SDN application from the TC's perspective, instructs the controller to manage slice's constituent VNFs and logically compose them to efficiently realize the network service(s) the slice offers. The life cycle of such network service(s) is managed by the NSO, which interacts with the TC via the OSS. The TC, deployed as a VNF, relies on the capabilities provided by virtual switches/routers (in the form of VNFs as well) to enable the VNF composition, forwarding pertinent instructions to such virtual switches/routers via its southbound interfaces. Through its northbound interfaces, the TC provides a means to securely expose selected network service capabilities to end users. Such interfaces allow end users to retrieve context information (real-time performance and fault information, user policies, etc.), operate, manage, and make use of the slice's network service(s), always within the limits set by the tenant.

The fact that each slice is provided with its own NSO, OSS, and TC instances enables the required management isolation.

Each tenant must efficiently orchestrate its assigned resources to simultaneously satisfy the diverging requirements of the slices that are under its management. The RO is the functional block that performs such task on behalf of the tenant, providing each slice with the required resources via interfaces with each slice's NSO. The RO must perform the resource sharing among slices while fulfilling their required performance, following an adequate, effective resource management framework that must comply with both tenant and slice-specific policies. Such a framework is required so that the RO enables performance isolation among slices.

All the NFVI resources available for use by a tenant (i.e., those that RO orchestrates) are supplied by the different InPs. Each InP rents part of the virtual resources according to a business lease agreement that both InP and tenant have previously signed. To access, reserve, and request such resources, the tenant's RO interacts with the VIM(s)/WIM(s) by means of interfaces those func-

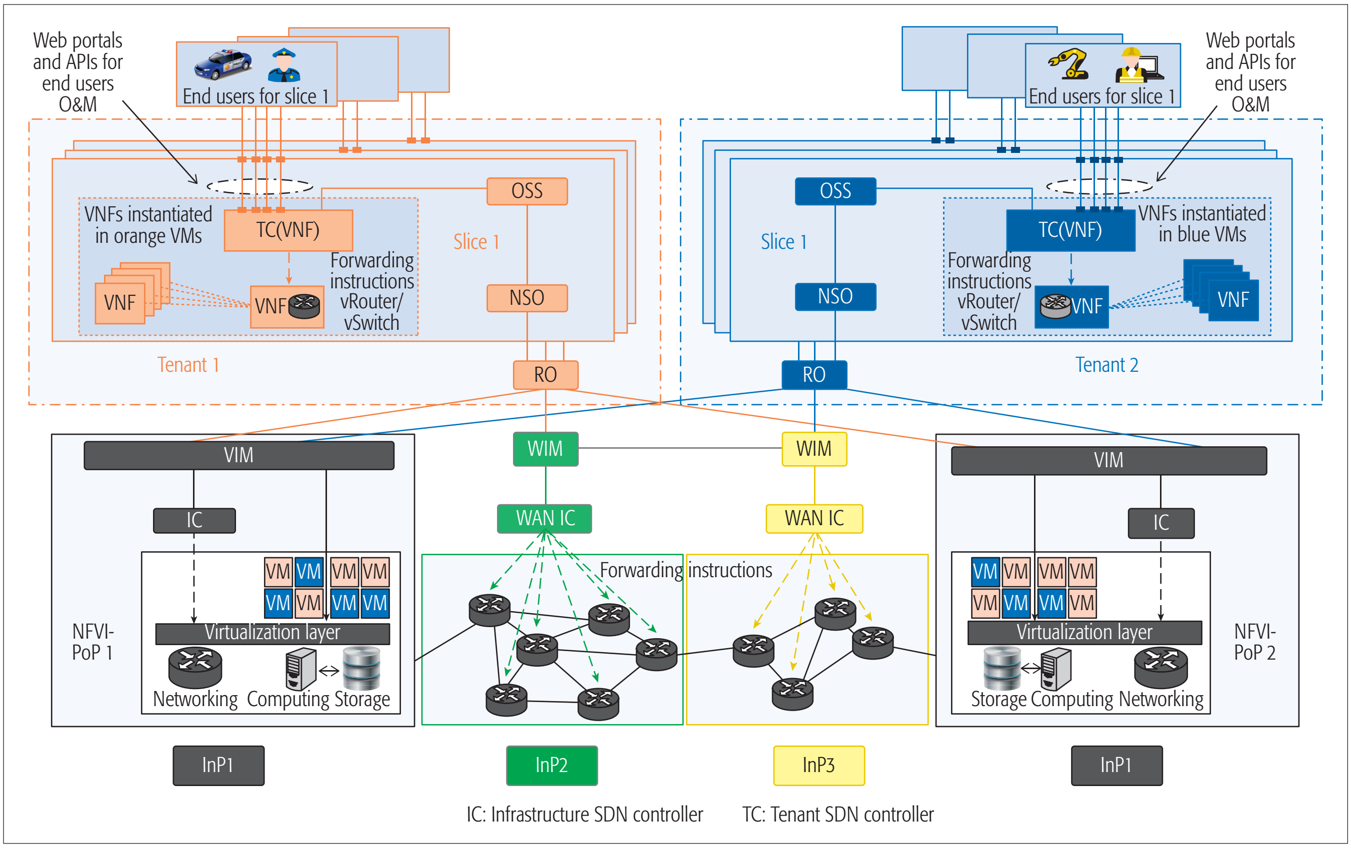


Figure 6. Network slicing deployment in a common framework, integrating both SDN and NFV.

tional blocks expose and that tenant's RO consumes. Indeed, we assume that VIMs and WIMs support multi-tenancy. We also assume that WIMs can communicate with each other according to predefined business agreements. In this respect, the interaction between a WIM and an RO might be achieved indirectly through another WIM.

As Fig. 6 suggests, the resource management must be performed at two levels: at the infrastructure level, where a slice-agnostic VIM/WIM provides the subscribed tenants with (virtualized) infrastructure resources, and at the tenant level, where the RO delivers its assigned resources to the corresponding slices. Both the VIM(s)/WIM(s) and the RO have to collect accurate resource usage information (each at its domain) and in turn to forecast resource availability in relatively short timescales to satisfy tenant and slice demands, respectively.

Please note that, with the exception of hardware resources, the functional blocks (e.g., VIM, RO, NSO, SDN controllers) are modeled as independent software components. The need for separate access, configuration, and management suggests this modeling, wherein the software relationships are enabled with the help of the application programming interfaces (APIs) that each component provides.

To preserve security and privacy isolation among slices, it is required to apply the compartmentalization principle at each virtualization level. In addition, each functional block and manageable resource (e.g. VNF) within a given slice must have its own security mechanisms, ensuring operation within expected parameters and preventing access to unauthorized entities. This is intended to guarantee that faults or attacks occurring in one slice are confined to that slice, preventing their propagation across slice boundaries.

Additionally, although recursion has not been addressed in this example, it is readily applicable to this scenario by simply assuming that some of the slice's users are tenants that in turn can deploy and operate their own slices.

Challenges and Research Directions

In this section, we identify the main challenges and future research arising from implementing slicing in 5G systems.

Performance Issues in a Shared Infrastructure

When network slices are deployed over a common underlying substrate, the fulfillment of the performance isolation requirement is not an easy task. If a tenant's RO only assigns dedicated resources to network slices, their required performance levels are always met at the cost of preventing slices sharing resources. This leads to overprovisioning, an undesired situation bearing in mind that the tenant has a finite set of assigned resources. One way to resolve this issue is to permit resource sharing (see, e.g., [13]), although this means slices are not yet completely decoupled in terms of performance. Thus, it is required to design adequate resource management mechanisms that enable resource sharing among slices when necessary without violating their required performance levels. To accomplish the sharing issue, the RO could use policies and strategies similar to those used in VIMs (e.g., the OpenStack Congress module or Enhanced Platform Awareness attributes).

Management and Orchestration Issues

Given the dynamism and scalability that slicing brings, management and orchestration in multi-tenant scenarios are not straightforward. To flexibly assign resources on the fly to slices, the optimization policy that governs the RO must deal with situations where resource demands vary considerably in relatively short timescales. To accomplish this:

- An appropriate cooperation between slice-specific management functional blocks and RO is required.
- Policies need to be captured in such a way that they can be automatically validated. This automation enables both the RO and slice-specific functional blocks to be authorized to perform the corresponding management and configuration actions in a timely manner.
- It is necessary to design computationally efficient resource allocation algorithms and conflict resolution mechanisms at each abstraction layer.

Security and Privacy

The open interfaces that support the programmability of the network bring new potential attacks to softwarized networks. This calls for a consistent multi-level security framework composed of policies and mechanisms for software integrity, remote attestation, dynamic threat detection and mitigation, user authentication, and accounting management. The security and privacy concerns arising from 5G slicing (see [14]) are today a major barrier to adopting multi-tenancy approaches.

New Business Models

The innovative partnerships between several players, each providing services at different positions of the value chain, and the integration of new tenants such as verticals, over-the-top service providers, and high-value enterprises, empowers promising business models. Given this business-oriented approach, new transition strategies must be broadly analyzed, allowing for a gradual evolution to future 5G networks and ensuring compatibility with past infrastructure investments. To accomplish this, a deep review of the telecom regulatory framework has to be made. Innovative ways of pricing, new grounds for cost sharing and standardized solutions, which provide the required support for interoperability in multi-vendor and multi-technology environments, must be studied as well.

Given this business-oriented approach, new transition strategies must be broadly analyzed, allowing for a gradual evolution to future 5G networks and ensuring compatibility with past infrastructure investments. To accomplish this, a deep review of the telecom regulatory framework has to be made.

Acknowledgments

This work is partially supported by the Spanish Ministry of Economy and Competitiveness, the European Regional Development Fund (Project TIN2013-46223-P), the Ministry of Education, Culture, and Sport of the Government of Spain, under grant Beca de Colaboración (2015-2016), and the University of Granada, under grant Beca de Iniciación a la Investigación (2016-2017).

References

[1] 5G-PPP, ERTICO, EFFRA, EUTC, NEM, CONTINUA and Networld2020 ETP, "5G Empowering Vertical Industries," white paper, Feb. 2016.

[2] M. Pearce, S. Zeadally, and R. Hunt, "Virtualization: Issues, Security Threats, and Solutions," *ACM Computer Surveys*, vol. 45, no. 2, Feb. 2013, pp. 1–38.

[3] N. M. M. K. Chowdhury and R. Boutaba, "A Survey of Network Virtualization," *Computer Networks*, vol. 54, no. 5, Apr. 2010, pp. 862–76

[4] ONF TR-521, "SDN Architecture," Feb. 2016.

[5] ONF TR-526, "Applying SDN Architecture to 5G Slicing," Apr. 2016.

[6] ETSI GS NFV 002, "Network Functions Virtualization (NFV); Architectural Framework," v. 1.1.1, Dec. 2014.

[7] ONF TR-518, "Relationship of SDN and NFV," Oct. 2015

[8] ETSI GS NFV-EVE 005, "Network Functions Virtualisation (NFV); Ecosystem; Report on SDN Usage in NFV Architectural Framework," v. 1.1.1, Dec. 2015.

[9] ETSI GS NFV-MAN 001, "Network Functions Virtualisation (NFV); Management and Orchestration," v. 1.1.1, Dec. 2014.

[10] ETSI GS NFV-INF 001, "Network Functions Virtualisation (NFV); Infrastructure Overview," v. 1.1.1, Jan. 2015.

[11] ETSI GS NFV-SEC 003, "Network Functions Virtualisation (NFV); NFV Security; Security and Trust Guidance," v. 1.1.1, Dec. 2014.

[12] ETSI GS NFV 003, "Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV," v. 1.2.1, Dec. 2014.

[13] P. Andres-Maldonado *et al.*, "Virtualized MME Design for IoT Support in 5G Systems," *Sensors*, vol. 16, no. 8, Aug. 2016, pp. 1338–62.

[14] R. Harel and S. Babbage, "5G Security Recommendations Package #2: Network Slicing," NGMN Alliance, Apr. 2016.

Biographies

Jose Ordonez-Lucena (jordonez93@gmail.com) received his B.Sc. in telecommunications engineering from the University of Granada, Spain, in 2015. Currently he is a Master's student and is working in research projects with the Department of Signal Theory, Telematics and Communication of the University of Granada. His research interests are focused on network virtualization and network slicing in 5G systems.

Pablo Ameigeiras (pameigeiras@ugr.es) received his M.Sc.E.E. degree in 1999 from the University of Malaga, Spain. He performed his Master's thesis at the Chair of Communication Networks, Aachen University, Germany. In 2000 he joined the Cellular System group at Aalborg University, Denmark, where he carried out his Ph.D. thesis. After finishing his Ph.D. he worked at Optimi/Ericsson. In 2006 he joined the University of Granada, where he has been leading several projects in the field of LTE and LTE-Advanced systems. Currently his research interests include the application of the SDN and NFV paradigms for 5G systems.

Diego Lopez (diego.r.lopez@telefonica.com) joined Telefonica I+D in 2011 as a senior technology expert and is currently in charge of the Technology Exploration activities within the GCTO Unit. He is focused on network virtualization, infrastructural services, network management, new network architectures, and network security. He chairs the ETSI ISG on Network Functions Virtualization and the NFVRG within the IRTF, and he is a member of the Board of 5TONIC, the Telefonica 5G Testbed.

Juan J. Ramos-Munoz (jjramos@ugr.es) received his M.Sc. in computer sciences in 2001 from the University of Granada. Since 2009, he has held a Doctorate degree from the same university. He is a lecturer at the Department of Signals Theory, Telematics and Communications of the University of Granada. He is also a member of the Wireless and Multimedia Networking Lab. His research interests are focused on real-time multimedia streaming, quality of experience assessment, network virtualization, and network slicing for 5G.

Javier Lorca (franciscojavier.lorcahernando@telefonica.com) received an M.Sc. degree in telecommunication engineering from Universidad Politécnica de Madrid in 1998, and is currently pursuing his Ph.D. degree at the University of Granada. He is in charge of Radio Access Networks Innovation within Telefónica Global CTO. His research is focused on 5G, including virtualization, massive MIMO, mmWave, new waveforms, interference control, and advanced channel coding techniques. He has multiple patents and publications and a book chapter on 5G.

Jesús Folgueira (jesus.folgueira@telefonica.com) received his M.Sc. degree in telecommunications engineering from Universidad Politécnica de Madrid (1994) and M.Sc. in telecommunication economics in 2015 (UNED). He joined Telefónica I+D in 1997. He is currently the head of Transport and IP Networks within Telefonica Global CTO unit, in charge of network planning and technology. He is focused on optical, metro, and IP network architecture and technology, network virtualization (SDN/NFV), and advanced switching. His expertise includes broadband access, R&D management, and network deployment.

PERMIT: Network Slicing for Personalized 5G Mobile Telecommunications

Tarik Taleb, Badr Mada, Marius-Iulian Corici, Akihiro Nakao, and Hannu Flinck

The authors discuss the need for the deep customization of mobile networks at different granularity levels: per network, per application, per group of users, per individual users, and even per data of users. They also assess the potential of network slicing to provide the appropriate customization and highlight the technology challenges.

Abstract

5G mobile systems are expected to meet different strict requirements beyond the traditional operator use cases. Effectively, to accommodate needs of new industry segments such as healthcare and manufacturing, 5G systems need to accommodate elasticity, flexibility, dynamicity, scalability, manageability, agility, and customization along with different levels of service delivery parameters according to the service requirements. This is currently possible only by running the networks on top of the same infrastructure, the technology called network function virtualization, through this sharing of the development and infrastructure costs between the different networks. In this article, we discuss the need for the deep customization of mobile networks at different granularity levels: per network, per application, per group of users, per individual users, and even per data of users. The article also assesses the potential of network slicing to provide the appropriate customization and highlights the technology challenges. Finally, a high-level architectural solution is proposed, addressing a massive multi-slice environment.

Introduction

Mobile networks are nowadays architected to serve all mobile users, ensuring some degree of service-level differentiation but with no specific tailoring of the functioning to the specific user needs. However, statistics demonstrate that users do not behave all in the same way: 53 percent are light mobile phone users, 24 percent exhibit medium usage behavior, and the remaining 23 percent are heavy mobile phone users [1]. Even among heavy mobile users, usage patterns of data-intensive mobile applications, that is, those related to social, news, and video, vary considerably [2, 3]. From these statistics and others, it becomes apparent that having the same mobile network architecture serving all mobile users, let alone all mobile applications, despite the diversity they exhibit in their attitudinal response to mobile services, have to be rethought.

Furthermore, a mobile user usually subscribes to a single mobile operator that provides the delivery of all the mobile services. In addition, a single mobile network usually ensures communication for all service types, regardless of the suitability of its available functionality to deliver these services with acceptable quality of experience (QoE) and network efficiency. Due to the uniformity of the network, all users of the network are charged based on the same bandwidth consumption model, which fails to capture the specifics of the usage of applications with large overhead and makes too expensive the network for individual large-scale sensor deployments.

Last but not least, current mobile core networks are offering a uniform ubiquitous service for all the connected devices. Even if a mobile user moves far away from the mobile core network infrastructure, he/she remains serviced by the same core network, even in the case of a highly decentralized mobile system [4]. This feature may impact numerous emerging advanced mobile services with strict latency and jitter requirements (e.g., augmented reality and self-driving vehicles). High latency and jitter degrade such mobile services, rendering their respective devices unusable, ultimately turning users away and impacting revenues. Also, a large number of users do not move from a specific network area, their mobility support being only an additional overhead on the network.

To cope with the above, this article advocates the need for customizing mobile telco services, through providing a functional differentiation for the different user requirements. Herein, the intention is to leverage the emerging technologies in the areas of network function virtualization (NFV) [5], software defined networking (SDN), and cloud and edge computing for providing a single infrastructure on top of which multiple versions of the same software, generically named slices, customized to have specific behaviors are running, through this removing the overhead of the uniform network service. Furthermore, in order to account for different layers of granularity, this article underlines the need and the technical possibilities to support a very large number of slices as well as their appropriate deployment according to the momentary location of the subscribers.

This article is organized in the following fashion. We present a quick overview on network slicing and highlights its utility for the Personalized Mobile Telecom (PERMIT) vision. The overall PERMIT framework is portrayed. We introduce the PERMIT slice orchestration system and discusses its challenges. The article then concludes.

Network Slicing

The term "network slicing" has captured much attention within research communities and the industry, as well as standards development organizations (SDOs), such as the Next Generation

Digital Object Identifier: 10.1109/MCOM.2017.1600947

Tarik Taleb is with Aalto University and Sejong University; Badr Mada is with Aalto University; Marius-Iulian Corici3 is with Fraunhofer Fokus; Akihiro Nakao is with the University of Tokyo; Hannu Flinck is with Nokia Bell Labs.

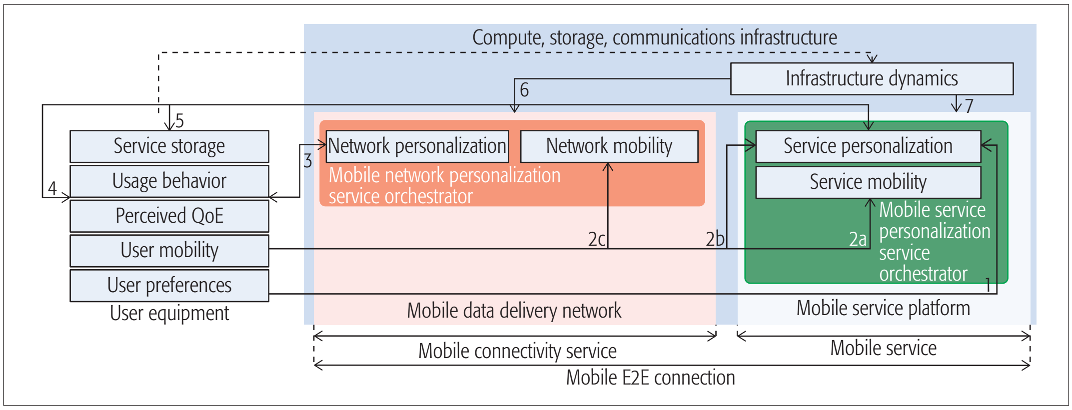


Figure 1. PERMIT virtual mobile network architecture.

These lightweight VMNs are expected to have the flexibility to serve up to a granularity of a single service for one individual user, multiple services for one individual user, or a group of users, although due to scalability reasons, most probably users with same services and behavior will be grouped into the same slice.

Mobile Network Alliance (NGMN), Third Generation Partnership Project (3GPP), and International Telecommunication Union — Telecommunication Standardization Sector (ITU-T). Although the definition of network slicing is still under heavy discussion, it generally means an isolated collection of resources and functions implemented through software programs on top of the resources, flexibly allocated on demand in order to enable quality of service (QoS) guarantee for the network requirements as well as in-network processing along end-to-end communications.

In PERMIT, network slicing (slicing hereafter) is considered to be one of the most important concepts to realize personalization of mobile networks for users. Although slicing in the mobile networking context has often been addressed for enabling different classes of communications (e.g., enhanced mobile broadband, massive machine-type communications, and ultra reliable and low latency communications), in PERMIT, slices can be instantiated per user, and in the most extreme case, per device and/or per application.

It is a well-known fact that the concept of slice in networking was first introduced in the overlay network research efforts, such as PlanetLab, in 2002. At that time, a slice was defined as an isolated static set of resources allocated for a group of users who "program" network functions and services over their overlay network, overlaid across "the planet." Since various network virtualization testbed efforts (e.g., GENI, VNode, FLARE, and Fed-4Fire) have inherited the concept of slices as a set of programmable, dynamically allocated resources to tailor new network services and protocols, it is quite natural to use the slice concept in PERMIT to further personalize the access and in-network edge processing for mobile network users.

The PERMIT Framework

Figure 1 schematically depicts the main components of the PERMIT framework. It mainly consists of two orchestrators, the mobile network personalization service orchestrator (MNP-SO) and the mobile service personalization service orchestrator (MSP-SO). PERMIT also envisions some changes to the user equipment as detailed below. The MNP-SO and MSP-SO entities can run separately or jointly on dedicated hardware or as software on virtual machines (VMs) with adequate characteristics. These two entities incorporate all necessary intelligence for mobile service personalization and mobile network personalization, respectively. They are decision making entities that decide how mobile services and the lightweight virtual mobile network (VMN) [6, 7] to transport them shall be personalized to the current and anticipated needs of a mobile user or a group of mobile users. In their decision making procedures, both MNP-SO and MSP-SO take into account the underlying infrastructure dynamics, processing and storage capabilities of user equipment, and users' contextual information like mobility and resource usage patterns.

In the envisioned PERMIT framework, lightweight VMNs are expected to run as a slice on one or multiple instantiated virtual resources [6, 8]. These lightweight VMNs are expected to have the flexibility to serve up to a granularity of a single service for one individual user, multiple services for one individual user, or a group of users, although due to scalability reasons, most probably users with same services and behavior will be grouped into the same slice. This flexibility can be attained by composing the lightweight VMN of uncorrelated building blocks that can be freely and dynamically combined or separated as per the requirements of the target mobile services and the needs of the serviced users [6].

Indeed, in PERMIT, a mobile network component (a network function) is defined in terms of its application logic and data, as depicted in Fig. 2. The application logic is decomposed into compute "blocks," including one basic block that provides minimal core network functionalities and multiple added value blocks, extending the basic components to provide additional network services such as communication reliability, QoS, and mobility support or accounting. The blocks run either at the network edge or in the cloud, depending on latency, bandwidth, resilience, and security requirements.

The composition of the blocks should follow the latest frameworks in service composition such as the ones based on micro-service application programming interfaces (APIs) [9] or on event bus communication. Both of these ensure the loose coupling between the main functionality and the other modules, permitting the addition of new functions on demand and even during runtime, as well as a comprehensive separation of the liabilities. The PERMIT architecture will further optimize these mechanisms to address the end-to-end delay while processing a request through multiple compute blocks.

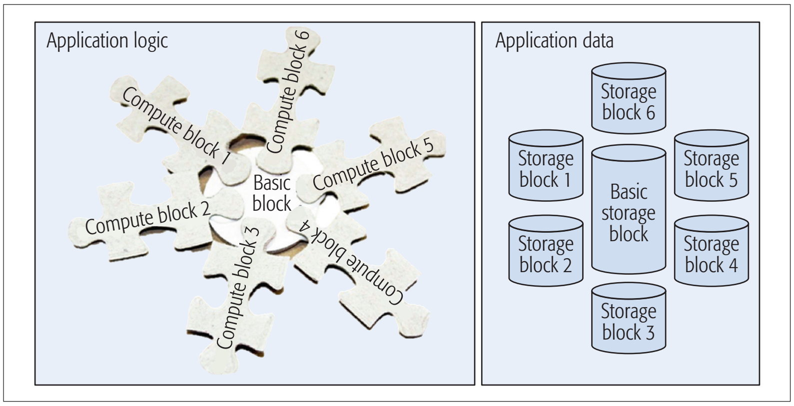


Figure 2. Network functions consisting of application logic and application data.

The PERMIT architecture enables the configuration of the in-service parameters related to the network functions as well as to the composition of the network functions within end-to-end services. This approach enables the creation of lightweight, customizable, and truly elastic mobile networks with network services/blocks that can be adapted to the users' needs. It will also enable seamless decomposition of application logic and data, allowing them to be moved to more convenient network locations. It is worth noting that careful attention shall be paid to how to gracefully compose the different functionalities within the network functions, especially considering the performance of multi-vendor software components, in order to provide a retraceable service level as well as to separate liabilities in case of failures.

At the protocol level, the lightweight feature of VMNs can also be achieved by simplifying a number of procedures typically required for mobile networks like authentication, authorization, and accounting (AAA) and charging functions. Dependencies on data anchoring (i.e., the packet data network gateway in the Evolved Packet System, EPS) and mobility anchoring (i.e., the serving gateway in EPS) concepts shall be relaxed if not completely replaced, as in the follow me cloud concept [10, 11]. With customizable VMNs, currently unsupported communication modes that mimic connectionless communication over shared media become possible over mobile networks.

In PERMIT, the personalization of mobile networks for a mobile user or a group of mobile users is achieved by anticipating the needs of the mobile services of this user or this group of users. Indeed, once the needs of a mobile service or a set of mobile services received by an individual user or a group of users are anticipated, the right VMN with the right characteristics (e.g., composing building blocks, total number of VMs involved, the locations of their respective DCs, their respective CPUs/memory/storage) can be identified so that VMNs can scale up and down as per the assessed needs. Another aspect of VMN personalization consists of its mobility to a different data center when required. For this purpose, it is possible to leverage different algorithms and mechanisms [12] that decide on and enforce the VMN mobility as per the mobility patterns of the served mobile users and/or the dynamics of the underlying infrastructure, in such a way that the "mobile network," serving a user or a group of users, follows their mobility. This decision may be based on several possibly conflicting attributes/criteria such as the mobile service type (e.g., delay-sensitive), the perceived quality of experience (QoE), the migration cost, the activity level of the users, the usage behavioral patterns, the mobility patterns, and the dynamics of the underlying communication infrastructure (e.g., for load balancing). Inputs, used for VMN personalization and VMN mobility, and relevant to users' mobile service usage behavior, perceived QoE, a user's mobility, and dynamics of underlying communications infrastructure are schematically depicted in Fig. 1 through arrows 2c, 3, and 6.

The personalization of the mobile service depends first of all on the user preferences for the service delivery (arrow 1 in Fig. 1) and on its mobility (arrow 2). Based on insights on the user behavior and on its perceived QoE, the network is customized according to the user needs (arrow 3). As multiple users may have the same network requirements, the customization can be seen as a user classification problem toward the appropriate cluster of users that have the optimal handling of the communication requirements. Similar to network customization, a service customization may also be executed (arrow 4). As the users may come from different customized networks, mainly due to their multi-application terminals, the customization of the applications should consider the customization of the network as a given parameter. A further step in the customization is the distribution of the service data (arrow 5) to the user equipment (UE) when needed or when the network conditions are appropriate, depending on the specific user behavior.

Finally, the customization highly depends on the availability of the infrastructure resources for the specific customization (arrows 6 and 7). The decisions of both the MNP-SO and MSO-SO depend on the possibility of the infrastructure to support their needs at the specific location. As the PERMIT architecture assumes a very large number of slices (i.e., up to one for each network user), it is possible that the network infrastructure will not have enough momentary resources to handle the subscriber communication. In this case, the subscribers should be classified in a default communication class for which the processing is handled by a central data center within a default slice similar to the current network infrastructure.

The PERMIT VMN Slice Orchestration System

PERMIT aims to achieve elasticity, flexibility, dynamicity, scalability, manageability, and efficiency beyond the current network level by building on-demand VMN slices, customized to the service requirements, thus much reducing the network overhead.

The PERMIT architecture is envisioned to function for different slices ranging from slices of individual users to Internet of Things (IoT) application slices as well as for verticals such as industrial control systems, autonomous driving, virtual reality, or video streaming. The resulting system will consist of numerous network slices, running in parallel and composed for end-to-end service delivery (Fig. 3).

Each slice consists of a set of virtual network functions (VNFs) within both the control and data

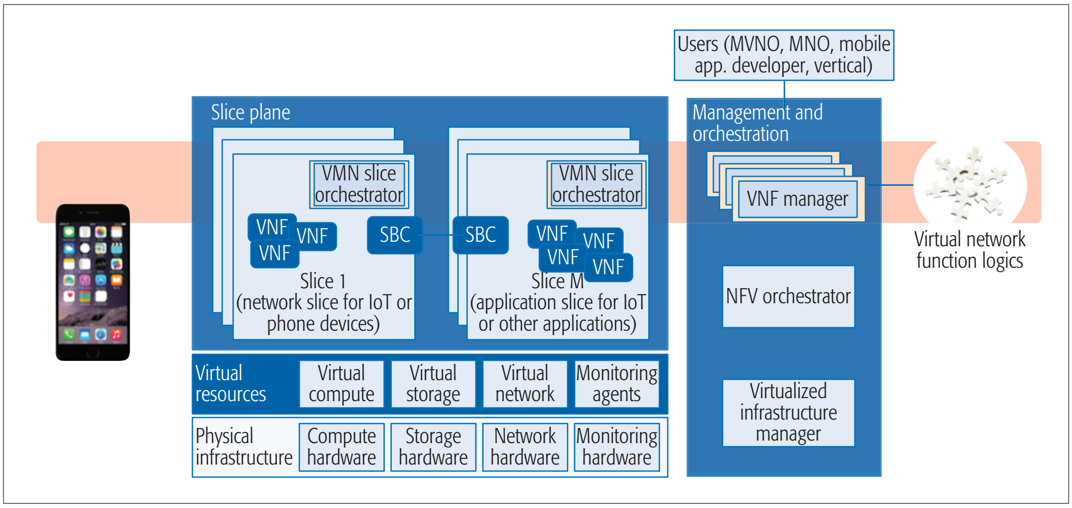


Figure 3. PERMIT VMN Slice Orchestration System: diverse VMN slices running in parallel and serving diverse verticals or individ-ual users.

The role of the VNF manager is to transmit network function placement [12] and scaling requirements to the NFV orchestrator, as well as to transmit to the VNFs the dependency parameters in order to enable the communication between the different VNF Components within a slice.

planes, customizable to the particular service types or vertical market needs or personalized to the individual end user, as presented earlier. From the perspective of the orchestration, the VNF components of each slice are seen as software functions that may be composed and customized, thus transparent to the functions they handle.

The VNFs accommodate the intrinsic features of the slices and their changing requirements, such as scaling up to match sudden growth in their traffic or smooth mobility to another network location. Also taking into account the requirements of the service delivery in terms of latency, reliability, and security together with the large number of slices, resource control becomes highly complicated. With the deployment of a large number of software services across multiple cloud and edge data centers, the complexity of the system increases beyond the capabilities of a single orchestration node. Moreover, a centralized network orchestrator may not be able to make the appropriate decisions and enforce them in due time, as it needs to handle a large amount of runtime operations, especially related to the sharing of the common data path environment. These delay and scalability limitations can be overcome through a distributed orchestration system where parts of the orchestration functionality are delegated to the edge nodes [13]. Data path sharing between the different slices and the decisions that require immediate response, such as network function failures, are particularly suitable for such delegation.

Figure 3 shows a high-level architecture of the PERMIT VMN slice orchestration system. The physical infrastructure consists of hardware for computing, storage, networking, and monitoring. These equipments can be administrated by the same entity or could belong to different domains. The slice orchestration plane of the architecture include images of VNFs, which represent the software version of existing network equipment. These VNFs could consist of building blocks (Fig. 2) designed in a clean-slate fashion or as components of existing network equipment. The VNF slice orchestration system is the main component of the architecture. It creates slices of VNFs for an individual user or a group of end users of a vertical. These slices can be created following pre-defined blueprints or in a fine-grained fashion, taking into account inputs relevant to end users' mobile service usage behavior, perceived QoE, and mobility, as discussed earlier. The slice orchestration system can be owned by a cloud provider, a mobile operator, or a new stakeholder. The users of the system can be vertical providers (e.g., automotive and IoT service provider), a mobile application developer, or an individual end user wishing for personalized mobile telecommunication service. Users can communicate to the slice orchestration system via well defined northbound interfaces (e.g., following Open Mobile Alliance, OMA, guidelines).

In order to separate the orchestration concerns, the following PERMIT orchestration levels are considered. First, a basic NFV resource orchestrator is considered, able to broker the available virtual resources, as provided by a virtualized infrastructure manager (VIM), to the different slices. The NFV orchestrator receives resource allocation requests from the VNF managers, one for each slice, which are aware of the specific slice logic. The role of the VNF manager is to transmit network function placement [12] and scaling requirements to the NFV orchestrator, as well as to transmit to the VNFs the dependency parameters in order to enable the communication between the different VNF Components within a slice.

To reduce the complexity and to appropriately manage the slice-specific operations, a VMN slice orchestrator is added to the architecture having the role of managing the functionality within the specific slice, including fault, configuration, accounting, performance, and security (FCAPS),[1] adapted to the dynamic resource environment. This includes the acquisition of monitored data on the specific service agents within the slice, a composition logic for the VNF components, enabling the appropriate processing flow allocation according to the momentary available resources as well as the interaction with other slices within the system. Another important role of the VMN slice orchestrator could also be offering flexible service function chaining (SFC) as a service, indicating the forwarding graph/path that a set of VNFs should be following within the respective slice. For this, the management system as well as the communication plane are extended with a slice border

[1] ISO/IEC 10040, 1998 — Information Technology — Open Systems Interconnection Systems Management Overview.

From the users' perspectives, PERMIT will facilitate a fully personalized and elastic end-to-end mobile connection service, which will provide mobile users with easy and efficient access to advanced mobile services. Indeed, with PERMIT, a mobile user may have his mobile network fully personalized to his current and anticipated needs or to the requirements of his mobile services.

control (SBC), which is similar to the session border controllers in the current architecture, and enables the filtering and the classification of both the inbound and outbound data traffic, the application-level firewall, as well as the appropriate forwarding to the components within the slice or to the SBC of the peer slice.

Through this separation of concerns within multiple functions, the possible policy conflicts are mitigated, although due to the shared resources, it may happen that the exact requirements of a specific slice would be partially fulfilled by the infrastructure. Considering the large number of slices, the alternative to provide the information on available resources to each VNF manager is highly complex.

From the users' perspectives, PERMIT will facilitate a fully personalized and elastic end-to-end mobile connection service, which will provide mobile users with easy and efficient access to advanced mobile services. Indeed, with PERMIT, a mobile user may have his/her mobile network fully personalized to his/her current and anticipated needs or to the requirements of his/her mobile services. True service elasticity will be attained: services of heavy mobile phone users shall never be throttled. Fair charging models can also be achieved: light and medium mobile users will have their respective VMNs running on smaller VMs, which will enable them to be fairly charged only for what they have indeed consumed. A group of users of a vertical or receiving a particular service may have a mobile network fully customized to their needs and the requirements of their mobile service. This shall enable the much desired service-tailored mobile networking concept. Furthermore, different VMNs with the right processing features can be created for different services as per the specifications of each service. Instead of being "locked in" the same mobile network for the delivery of all service types, users will then have the flexibility of subscribing to the most suitable VMN slice to receive a particular service type. Accordingly, subscription to multiple VMNs for multiple service types becomes possible. This requires UEs to have the capability to simultaneously connect to and steer mobile traffic across multiple VMN slices, optimally created for a set of services (Fig. 4). A default connectivity slice will always be assumed for each user, available for applications that use non-customized service delivery. A UE shall be able to discover existing VMN slices and request the creation of a new one. The creation of a new VMN slice for a particular service can be UE-initiated (following a set of rules and policies) or network-controlled when discovering that the current slice used for the application is inefficient for the delivery of the service to the user or to a group of users. A UE shall also have the ability to leave a VMN slice, join an existing one, or upgrade an ongoing one, either upon a UE request or upon a network-based classification of the UE.

The slice concept can be extended further to support the slicing of resources within a single UE. In the current smartphone market, it is quite common that UEs come with application vending facilities such as Google Play on Android phones, iTunes App Store for iPhones, iPads, and iPods, and proprietary application marketplaces on other smartphones. Applications on top of UEs must be examined and approved by their respective vendors and must run in the sandboxes they have prepared. We posit that the application marketplaces and the sandboxes as execution environments should be implemented, isolated within separate slices, so that one can have multiple, different, personalized execution environments and available application suites within a single UE. With this envisioned UE slicing concept, one may have multiple personalized containers within a single UE, so one may have different security and privacy contexts such as private usage and corporate usage as well as automatic personalized updates directly from the software developers. Considering data contamination and privacy breaches often observed in the mixed use of a single UE for private and public matters, it makes a lot of sense to introduce isolation between different usages. For example, one may want to isolate the applications with personal data (e.g., medical record and bank account information), those with confidential data (e.g., corporate confidential information), and those with public data (e.g., general web browsing applications) exposed to rather wild and rogue environments where security breach is often observed. If multiple UEs are carried to avoid such mishaps, it is reasonable to consolidate them into a single one using slicing techniques. Alternatively, one may have multiple UEs with different operating systems, so one may benefit from different application suites and execution environments. Embedded operating system virtualization and network virtualization technologies have already advanced to support such a concept of UE slicing [3, 16]. However, to the best of the authors' knowledge, PERMIT is the first to consider end-to-end slicing, including UE slicing, mobile network slicing, edge computing, and cloud computing. The authors are also already aware of the challenges of defining the granularity of slicing, such as the necessity of individual UE-level slicing and the feasibility of slicing in mobile operators. Although compelling application use cases drive such decisions, the authors have already conducted a feasibility study of application-specific slicing concepts [14, 15]. We plan to address these challenges in our future research work.

In PERMIT, a user can have his/her personalized VMN and his/her personalized mobile services constantly following him/her. In this way, PERMIT will support a wide gamut of high-quality services, customized to users' preferences, behavior, and mobility features. Emerging devices, such as Microsoft HoloLens, will largely benefit from the PERMIT approach, particularly when their users are moving in dense smart cities onboard smart connected vehicles.

PERMIT shall also define novel and promising business opportunities for cloud providers, especially in the area of providing value-added services beyond the basic infrastructure sharing. It also represents an innovative and ambitious solution to open up mobile networks and revolutionize the mobile networking principles, going from large-scale ubiquitous uniform connectivity service to highly efficient service support tailored to the specific needs of individuals. In PERMIT, a new set of business stakeholders may also emerge, orchestrating the mobile service and mobile network personalization for individuals and groups of users as well as orchestrating the interaction between the mobile service slices.

Conclusion and Future Research Directions

In this article, we propose the PERMIT approach, which is expected to act as a catalyst for structural changes to the current communication system configuration, whereby both the mobile delivery network and the mobile services it supports are personalized for each individual user, or alternatively customized for groups of users/verticals.

In PERMIT, service personalization goes beyond the traditional approaches whereby service personalization is based on the classical users' preferences, explicitly indicated by the users or deducted through collaborative filtering from other means (e.g., social networks), resulting in a specific parametrization of the uniform network. In PERMIT, users' mobility patterns, their mobile service usage behavioral patterns, and the dynamics of the underlying communications infrastructure are all taken into consideration for acquiring both network and service personalization, treating service personalization and networking customization as two flexibility-enabling and complementary components. To always ensure short response times for emerging advanced mobile services, the mobility of mobile services and the personalized mobile networks are both enabled toward the proximity of the respective mobile users across the overall mobile service area, as per the mobility features of mobile users, and in a seamless and cost-efficient manner. This will make both the mobile delivery network and the mobile services — after being personalized — constantly follow their respective mobile users.

With such an approach, service personalization and network personalization can take place dynamically and interactively. This also enables the much desired service-tailored mobile networking concept, achieving a so far unprecedented level of flexibility in service-specific optimizations, fine-grained network resource slicing, and transparent capacity scaling. In this regard, scalable programming of data plane and data paths and fine-grained mobility management of various services, considering both centralized and distributed approaches, are needed. This defines a promising research area that will stimulate the relevant community of researchers.

With such an approach, service personalization and network personalization can take place dynamically and interactively. This also enables the much-desired service-tailored mobile networking concept, achieving a so far unprecedented level of flexibility in service-specific optimizations, fine-grained network resource slicing and transparent capacity scaling.

Acknowledgments

This work was partially supported by the TAKE 5 project funded by the Finnish Funding Agency for Technology and Innovation (TEKES) and in part by the Finnish Ministry of Employment and the Economy. It is also partially supported by the European Union's Horizon 2020 research and innovation programme under the 5G!Pagoda project with grant agreement no. 723172.

References

[1] Y. Jin *et al.*, "Characterizing Data Usage Patterns in a Large Cellular Network," *Proc. SIGCOMM CellNet Wksp.*, Helsinki, Finland, Aug. 2012.

[2] P. Du and A. Nakao, "Application Specific Mobile Edge Computing through Network Softwarization," *IEEE CloudNet '16*, Pisa, Italy, Oct. 2016.

[3] A. Nakao, P. Du, and T. Iwai, "Application Specific Slicing for MVNO through Software-Defined Data Plane Enhancing SDN," *IEICE Trans. Commun.*, vol. E98-B, no. 11, 2015, pp. 2111–20.

[4] T. Taleb, K. Samdanis, and A. Ksentini, "Supporting Highly Mobile Users in Cost-Effective Decentralized Mobile Operator Networks," *IEEE Trans. Vehic. Tech.*, vol. 63, no. 7, Sept. 2014, pp. 3381–96.

[5] ETSI GS, "Network Function Virtualization (NFV) Management and Orchestration," NFV-MAN 001 v. 0.8.1, Nov. 2014.

[6] T. Taleb *et al.*, "EASE: EPC as a Service to Ease Mobile Core Network," *IEEE Network*, vol. 29, no. 2, Mar. 2015, pp. 78–88.

[7] T. Taleb, "Towards Carrier Cloud: Potential, Challenges, & Solutions," *IEEE Wireless Commun.*, vol. 21, no. 3, June 2014. pp. 80–91.

[8] T. Taleb, A. Ksentini, and A. Kobbane, "Lightweight Mobile Core Networks for Machine Type Communications," *IEEE Access*, vol 2, Oct. 2014, pp. 1128–37.

[9] I. Nadareishvili *et al.*, *Microservice Architecture: Aligning Principles, Practices, and Culture*, O'Reilly, 2016.

[10] T. Taleb and A. Ksentini, "Follow Me Cloud: Interworking Federated Clouds and Distributed Mobile Networks," *IEEE Network*, vol. 27, no. 5, Sept./Oct. 2013. pp. 12–19.

[11] A. Ksentini, T. Taleb, and F. Messaoudi, "A LISP-Based Implementation of Follow Me Cloud," *IEEE Access*, vol. 2, Oct. 2014. pp. 1340–47.

[12] T. Taleb, M. Bagaa, and A. Ksentini, "User Mobility-Aware Virtual Network Function Placement for Virtual 5G Network Infrastructure," *Proc. IEEE ICC 2015*, London, U.K., June 2015.

[13] A. Aissioui *et al.*, "Elastic and Distributed SDN Controllers for Follow-Me Cloud," *IEEE Access*, DOI 10.1109/ACCESS.2015.2489930, vol. 3, Nov. 2015.

[14] C. Dall *et al.*, "The Design, Implementation, and Evaluation of Cells: A Virtual Smartphone Architecture," *ACM Trans. Comp. Sys.*, vol. 30, no. 3, Aug. 2012.

[15] J. Andrus *et al.*, "Cells: A Virtual Mobile Smartphone Architecture," *Proc. 23rd ACM Symp. Operating Systems Principles*, Oct. 2011.

[16] A. Nakao and P. Du, "Application and Device Specific Slicing for MVNO," *1st Int'l. Science and Tech. Conf. SDN & NFV*, Moscow, Russia, Oct. 2014.

Additional Reading

[1] A. Ksentini *et al.*, "On Using Bargaining Game for Optimal Placement of SDN Controllers," *IEEE ICC '16*, Kuala Lumpur, Malaysia, May 2016.

Biographies

Tarik Taleb (tarik.taleb@aalto.fi) is currently a professor at Aalto University, Finland, leading the MOSA!C Lab. He worked as a senior researcher at NEC Europe Ltd until April 2015. Prior to that, he worked as an assistant professor at Tohoku University, Japan. He received his B.E. degree in information engineering with distinction, and his M.Sc. and Ph.D. degrees in information sciences from Tohoku University in 2001, 2003, and 2005, respectively.

Badr Mada (badr.mada@aalto.fi) is currently conducting his doctoral studies at Aalto University. His research focuses on mobile edge computing and open source networking. He received his Bachelor degree in mathematical and computer science and his Master degree in software engineering from University Mohammed V, Rabat, Morocco, in 2014 and 2017, respectively.

Marius Corici (marius-iulian.corici@fokus.fraunhofer.de) has been a senior researcher in Fraunhofer FOKUS's NGNI department for 12 years. He is currently deputy director of the Software Networks division in charge of research and innovation in the areas of 5G, NFV, SDN, and massive IoT, and the development of the correspondent software toolkits, sustaining the industry and academia R&D to obtain and demonstrate meaningful results with high impact toward standardization.

Akihiro Nakao (nakao@iii.u-tokyo.ac.jp) received his B.S. (1991) in physics and M.E. (1994) in information engineering from the University of Tokyo. He was at IBM Yamato Laboratory, Tokyo Research Laboratory, and IBM Texas Austin from 1994 to 2005. He received his M.S. (2001) and Ph.D. (2005) in computer science from Princeton University. He is a full professor and also a department chair at the Interfaculty Initiative in Information Studies, Graduate School of Interdisciplinary Information Studies, University of Tokyo.

Hannu Flinck (hannu.flinck@nokia-bell-labs.com) is a research manager at Nokia Bell Labs, Espoo, Finland. He received his M.Sc. degree (1986) and Lic. Tech. degree (1993) in computer science and communication systems from Helsinki University of Technology, Finland. His current research agenda includes cloud technology, SDN, and content delivery in fifth generation mobile networks.

Network Slicing in 5G: Survey and Challenges

Xenofon Foukas, Georgios Patounas, Ahmed Elmokashfi, and Mahesh K. Marina

The authors review the state of the art in 5G network slicing and present a framework for bringing together and discussing existing work in a holistic manner. Using this framework, they evaluate the maturity of current proposals and identify a number of open research questions.

ABSTRACT

5G is envisioned to be a multi-service network supporting a wide range of verticals with a diverse set of performance and service requirements. Slicing a single physical network into multiple isolated logical networks has emerged as a key to realizing this vision. This article is meant to act as a survey, the first to the authors' knowledge, on this topic of prime interest. We begin by reviewing the state of the art in 5G network slicing and present a framework for bringing together and discussing existing work in a holistic manner. Using this framework, we evaluate the maturity of current proposals and identify a number of open research questions.

INTRODUCTION

Mobile devices have become an essential part of our daily lives and, as such, the mobile network infrastructure that connects them has become critical. It is set to take on an even bigger role with the fifth generation (5G) mobile systems envisioned to support a wide array of services and devices. In this article, we consider the architectural aspect of mobile networks looking toward 5G. Examining the evolution of mobile networks until now suggests that the changes across generations have been driven largely by the need to support faster data oriented services. For instance, spectral efficiency in the radio access network (RAN) has increased by a factor of 30 from 2G to 4G. On the core network (CN) front, the packet switched component (IP) introduced initially in the 2.5G (general packet radio service, GPRS) system eventually supplanted the legacy circuit switched component in 4G systems.

What 5G systems are going to be has yet to be determined. However, it is conceivable that the eventual 5G system will be a convergence of two complementary views that are currently driving the research and industrial activity on 5G. One is an *evolutionary* view focusing on significantly scaling up and improving the efficiency of mobile networks (e.g., 1000× traffic volume, 100× devices, and 100× throughput). Much of the research focused around this view is on the radio access side looking at novel technologies and spectrum bands (e.g., massive miultiple-input multiple-output, MIMO; millimeter-wave).

The other *service-oriented* view envisions 5G systems catering to a wide range of services differing in their requirements and types of devices, and going beyond the traditional human-type communications to include various types of machine-type communications. This requires the network to take different forms depending on the service in question, leading naturally to the notion of *slicing* the network on a per-service basis, the focus of this article. Realizing this service-oriented view requires a radical rethink of the mobile network architecture to turn it into a more flexible and programmable fabric, leveraging technologies like software defined networking (SDN) and network functions virtualization (NFV), which can be used to simultaneously provide a multitude of diverse services over a common underlying physical infrastructure. We take this view in this article as it is intertwined with the 5G mobile network architecture, although the evolutionary view also has architectural implications.

We aim to survey the existing work on network slicing in the 5G context and identify challenges remaining to be addressed to make the service-oriented 5G vision a reality. This article is, to the authors' knowledge, the first survey on this important topic. We start by outlining representative 5G architectural proposals that highlight the crucial role network slicing is expected to play to meet the widely different requirements of various use cases. We then present a generic 5G architectural framework made up of infrastructure, network function, and service layers as well as the cross-cutting aspect of service management and orchestration (MANO). With respect to this framework, we discuss the state of the art in network slicing in the mobile/5G context. This leads us to identify some key outstanding challenges to realize a slice-able, softwarized 5G mobile network architecture.

NETWORK SLICING IN THE 5G ARCHITECTURE

This section highlights the need for a flexible 5G architecture to accommodate diverse use cases; outlines representative architectural proposals that indicate the crucial role network slicing is expected to play; and presents a generic framework that broadly represents various proposals and is used as the reference for our network slicing literature review in following sections.

USE CASES AND REQUIREMENTS

The 5G network is expected to be the basis for a range of verticals and use cases. For example, the International Telecommunication Union (ITU) and Fifth Generation Public Private Partnership (5G-PPP) have identified three broad use case families: enhanced mobile broadband, massive machine-type communications, and critical communications. Within those, it is possible to define several specific use cases [1] ranging from general broadband access with global coverage to spe-

Digital Object Identifier: 10.1109/MCOM.2017.1600951

Xenofon Foukas, Mahesh K. Marina are with the University of Edinburgh; Georgios Patounas and Ahmed Elmokashfi are with Simula Research Laboratory.

 0163-6804/17/$25.00 © 2017 IEEE

cialized networks for sensors or extreme mobility. The stark differences between these use cases translates to a set of heterogeneous requirements (Fig. 1) that cannot be satisfied by a one-size-fits-all architecture. With this in mind, alternative architectural proposals for 5G have emerged recently to accommodate use cases with diverse requirements; we outline two such proposals next.

Architecture

The Next Generation Mobile Network Alliance's (NGMN's) architectural vision [1] advocates a flexible softwarized network approach. This views network slicing as a necessary means for allowing the coexistence of different verticals over the same physical infrastructure. Initial proposals were limited to slicing the CN, but NGMN has argued for an end-to-end (E2E) scope that encompasses both the RAN and CN. To realize this and provide the needed context awareness, both parts need to be flexibly sliced into several overlaid instances serving different types of users, devices, and use cases. This whole process needs to be orchestrated by an E2E MANO entity that has a central role in the architecture.

The overall NGMN architecture is split into three layers: infrastructure resource, business enablement, and business application. Realizing a service in this proposal follows a top-down approach via a network slice blueprint that describes the structure, configuration, and workflows for instantiating and controlling the network slice instance for the service during its life cycle. The service/slice instance created based on the blueprint may be composed of several subnetwork instances, each in turn comprising a set of network functions and resources to meet the requirements stipulated by the service in question.

5G-PPP's architectural vision [3] offers a more elaborate examination of the roles and relationships between different parts of the 5G network. Overall, 5G-PPP shares the NGMN view that a potential 5G architecture must support softwarization natively and leverage slicing for supporting diverse use cases. 5G-PPP's architectural proposal is divided into five layers: infrastructure, network function, orchestration, business function, and service layers. Relating this to the NGMN proposal, while both are built on infrastructure and network function (business enablement) layers, there are a couple of differences: the orchestration/MANO is viewed as a separate layer in the 5G-PPP proposal; and the business application layer in the NGMN proposal is divided into two layers (business function and service) in the 5G-PPP case.

More generally, there seems to be a broad consensus on the need for native support for softwarization and network slicing as a means of realizing widely different services in 5G. Moreover, various 5G architectural proposals can be broadly mapped to a generic framework that is shown in Fig. 2. This framework is composed of three main layers: the infrastructure layer, the network function layer, and the service (or business) layer. It also consists of a MANO entity that translates use cases and service models into network slices by chaining network functions, mapping them to infrastructure resources, and configuring and monitoring each slice during its life cycle.

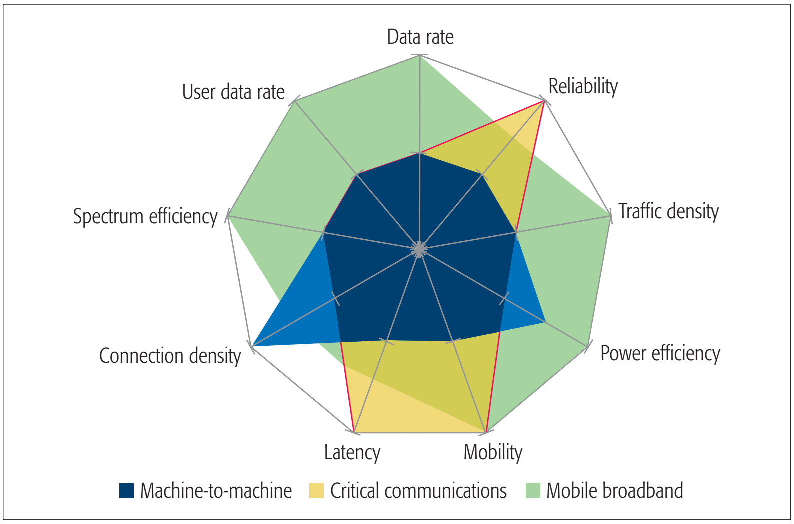


Figure 1. Key 5G use cases and their requirements. In this illustration, the further the distance of a requirement from the center, the more important it is to the corresponding use case. It is inspired by a similar illustration from ITU [2]. Diverse use cases need to be mapped to suitably tailored network structures. It is therefore vital for a 5G architecture to be flexible to realize different structures as needed.

State of the Art in 5G Network Slicing

In this section, we discuss the existing work on network slicing for 5G using the generic framework presented in the previous section (Fig. 2) as a reference. Table 1 summarizes this discussion. Figure 3 gives an overview of the various research issues related to network slicing, and indicates where future research should be focused.

Infrastructure Layer

Scope: The infrastructure layer broadly refers to the physical network infrastructure spanning both the RAN and the CN. It also includes deployment, control, and management of the infrastructure; the allocation of resources (computing, storage, network, radio) to slices; and the way that these resources are revealed to and can be managed by the higher layers.

Existing work: The related work focuses on two main subjects: the composition of the network infrastructure and its virtualization.

Composition of Network Infrastructure: It has been advocated that in order to realize slicing, we need to move towards the infrastructure as a service (IaaS) paradigm [4], where different infrastructural elements covering different requirements can be leased to accommodate the needs of the various slices. This paradigm is well known in the context of cloud computing, but it needs to be further adapted for the 5G context.

Considering the CN, there is a broad consensus in favor of using generic hardware infrastructure for the deployment of virtual network functions [5–7]. However, due to the differing constraints between various services deployed over the network, a simple centrally positioned cloud infrastructure might not be suitable for all the slices. For example, a sub-millisecond latency requirement of a tactile service such as remote surgery deployed over a dedicated network slice cannot be accommodated if the cloud infrastructure is located far away from the RAN. Therefore,

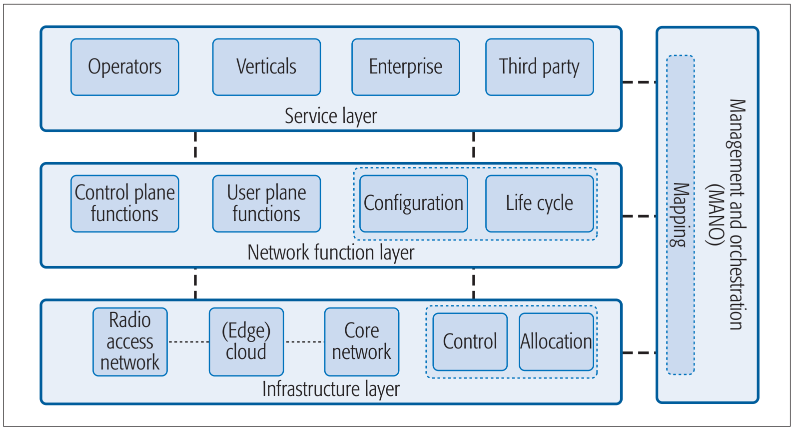


Figure 2. Generic framework representing various 5G architectural proposals. We review and appraise the 5G network slicing literature with respect to this framework.

some network slicing architectures [5, 7] propose a mix of central and edge cloud computing infrastructures where resources can be allocated to either of them, depending on the slice requirements.

On the other hand, the RAN comprising of multiple base stations is expected to span diverse radio access technologies (RATs) including LTE and Wi-Fi. Moreover, since the slices are expected to be created dynamically with their service requirements not known a priori, the RAN infrastructure needs to be flexible enough to provide support for various RATs on the fly, following a RAN as a service (RaaS) paradigm. This is why a large number of architectural proposals [4, 6, 7] for network slicing expect the deployment of generic software-defined base stations composed of centralized baseband processing units and remote radio heads as the logical next step.

Infrastructure Virtualization: The ability to virtualize the underlying infrastructure and provide isolation among services is essential for network slicing. This not only means virtualization and full isolation of the underlying resources (processing, storage, network, and radio) among slices but also the ability to support different types of control operations over the resources in a virtualized manner based on the service requirements. This characteristic of providing a virtualized end-to-end environment that can be potentially opened up and fully controlled by third parties is one of the key features that separates network slicing from the already existing network sharing solutions [4, 5].

Considering the virtualization of the CN infrastructure, research done in the context of cloud computing can be leveraged. Specifically, technologies like kernel-based virtual machines (KVMs) and Linux containers (LXC) can provide isolation guarantees in terms of processing, storage, and network resources at the operating system (OS) or process level. These isolation guarantees combined with the capabilities offered by platforms like OpenStack for the pooling of resources can greatly simplify the on-the-fly creation of virtualized CNs. Due to the high maturity level of the aforementioned technologies, concrete prototype implementations of slicing frameworks are already available, enabling the deployment of virtual core network functions — virtual mobility management entity (MME), virtual service gateway (SGW), and so on — over cloud infrastructures (e.g., [7]).

On the other hand, virtualization approaches for the RAN are at an early stage. Applying VM and container-based solutions in this domain does not fully address the problem as they do not deal with the additional dimension of virtualizing and isolating radio resources (spectrum and radio hardware). Existing RAN virtualization approaches that account for this dimension fall into one of two categories:

- Providing a dedicated chunk of spectrum for each virtual base station (slice) to deploy a full virtual network stack on top of it [7]
- Dynamically sharing the spectrum between different virtual base station instances (slices) by employing common underlying physical and lower medium access control (MAC) layers [8]

The dedicated spectrum approach is easier to implement, especially with dedicated radio hardware per slice, since isolation of radio resources is guaranteed through the static fragmentation of the spectrum, but it can result in inefficient use of radio resources. The other approach of fine-grained and dynamic spectrum sharing has the opposite problem of making isolation between slices challenging.

Network Function Layer

Scope: The network function layer encapsulates all the operations that are related to the configuration and life cycle management of the network functions that, after being optimally placed over the (virtual) infrastructure and chained together, offer an end to end service that meets certain constraints and requirements described in the service design of the network slice.

Existing work: The research interest in this layer mainly revolves around the technologies that can act as enablers for the deployment and management of network functions, as well as around issues regarding the granularity and type of the deployed functions.

Enabling Technologies: There already seems to be a consensus among researchers and the industry about the role of SDN and NFV [5–7, 9]. NFV is an ideal technology for the life cycle management and orchestration of the network functions, while SDN can inherently act as an enabler of NFV by allowing the configuration and control of the routing and forwarding planes of the underlying infrastructure through standardized protocols (e.g., Openflow).

Granularity of Network Functions: One particularly interesting aspect of this layer that is thoroughly discussed in various relevant works is the granularity (scope) of the available virtual network functions [5, 10]. On one end, we have coarse grained functions, where each one is responsible for a large portion of the network's operations (e.g., individual functions for LTE eNodeBs, MMEs, S-GWs). On the other end, we have functions with very fine granularity, where each of the coarse-grained functions mentioned above is divided further into many sub-functions. For example, in [11] the LTE Enhanced Packet Core (EPC) is broken down into functions responsible for mobility and forwarding traffic (MME, S-GW,

		Advantages	Disadvantages
Virtualization of radio resources	Dedicated resources	Natrually ensures resource isolation, ease of realizing virtual base station stacks and supporting multiple RATs	Inefficient utilization of radio resources
	Shared resources	More efficient use of radio resources	Requires more sophisticated techniques to ensure isolation of radio resources
Granularity of network functions	Coarse-grained	Easier deployment and management of network functions	Less flexible and adaptive to changes in underlying network conditions
	Fine-grained	More flexible and easier to conform to SLAs	Service chaining and interoperability of functions challenging
Service description	Human-readable format	Easier to express service requirements	MANO role challenging in mapping requirements to network components
	Set of functions and network components	Non-intuitive way to express service requirements	Simpler realization of network slices

Table 1. Approaches for addressing different aspects of network slicing and their (dis)advantages.

packet gateway, P-GW), which in turn are further decomposed into sub-functional entities including signaling load balancers, mobility managers, and functions dedicated to the forwarding of either control or data plane traffic.

The coarse-grained approach offers a more simplified way of placing and managing the network functions of a slice. However, this comes at the expense of a slice that is less flexible and less adaptive to the changes of the underlying network conditions, something that can be critical when the slice needs to conform to specific service level agreements (SLAs) [10]. For example, the radio resource scheduler in a slice might need to be swapped for another one with a different scheduling policy when a large number of mobile devices appear concentrated in a specific location in order to avoid violating the slice's SLA. If the scheduler is tightly coupled and packed as a single function with the rest of the eNodeB, performing the swapping operation can become a challenge. However, it has also been argued that, despite its benefits on the adaptation of a slice to the network conditions, fine granularity can be problematic for the interfacing and chaining of the network functions since the more network functions exist, the more interfaces need to be defined for their inter-communication [5]. This is particularly an issue when virtual network functions are made available by third parties through some kind of a network function store [7], since without common interfaces, their interoperability is not guaranteed. As a workaround, the use of a container-based protocol that will wrap the interface of the contained functions has been proposed [5]; however, there is no concrete description as to how to achieve this.

Service Layer and MANO

Scope: Perhaps the most important element that distinguishes network slicing in the context of 5G from other forms of slicing that have been considered in the past (e.g., cloud computing) is its end-to-end nature and the requirement to express a service through a high-level description and to flexibly map it to the appropriate infrastructural elements and network functions. This observation regarding the operation of slicing in the context of 5G naturally leads to two new high-level concepts:

- A service layer that is directly linked to the business model behind the creation of a network slice
- Network slice orchestration for the hypervision of a slice's life cycle.

Existing work: Due to the novelty that this layer introduces in terms of concepts and ideas, the related research in this domain naturally focuses on answering fundamental questions regarding network slicing architectures. More specifically, the topics considered are related to the way services should be described and how they should be mapped to the underlying network components, and the architecture of network slicing managers and orchestrators.

Service Description: Regarding the service layer and the way that the business model of a service should be described in high-level terms, there are two different proposals. In one approach, the service level description (manifest) is simply a set of traffic characteristics, SLA requirements (e.g., for performance related aspects like throughput and latency), and additional services (e.g., localization service) [6]. In the second approach, the service description is more detailed in the sense that it can identify specific functions or RATs that are bundled together and should be used for the creation of the slice [7] that provides a specific service (slice as an application). The main difference lies in the way that the network slice will be generated. In the first case, the slice orchestrator will be assigned the more complex task of identifying the appropriate functions and technologies that will guarantee the fulfillment of the requirements described in the slice's manifest, while in the second case things are more simplified since the required building blocks of the slice are already identified in its description. However, the second approach can be less efficient as it leaves less flexibility to the slice orchestrator to tune the components of the slice.

MANO architecture: The exact form that the network slicing MANO entity should have is still unclear with different works presenting different ideas. Some proposals envision that network slicing will come through an evolution of the current Third Generation Partnership Project (3GPP) standards and therefore propose enhancements in terms of interfaces and functionalities for the existing mobile architecture [12]. Others envision

An important issue is how to map and stitch together the components that are available to the various layers of the architecture in order to compose an end-to-end slice. Two types of mapping have been considered: the functional/SLA mapping of the service requirements to network functions and infrastructure types; and the mapping of network functions and infrastructure types to vendor implementations.

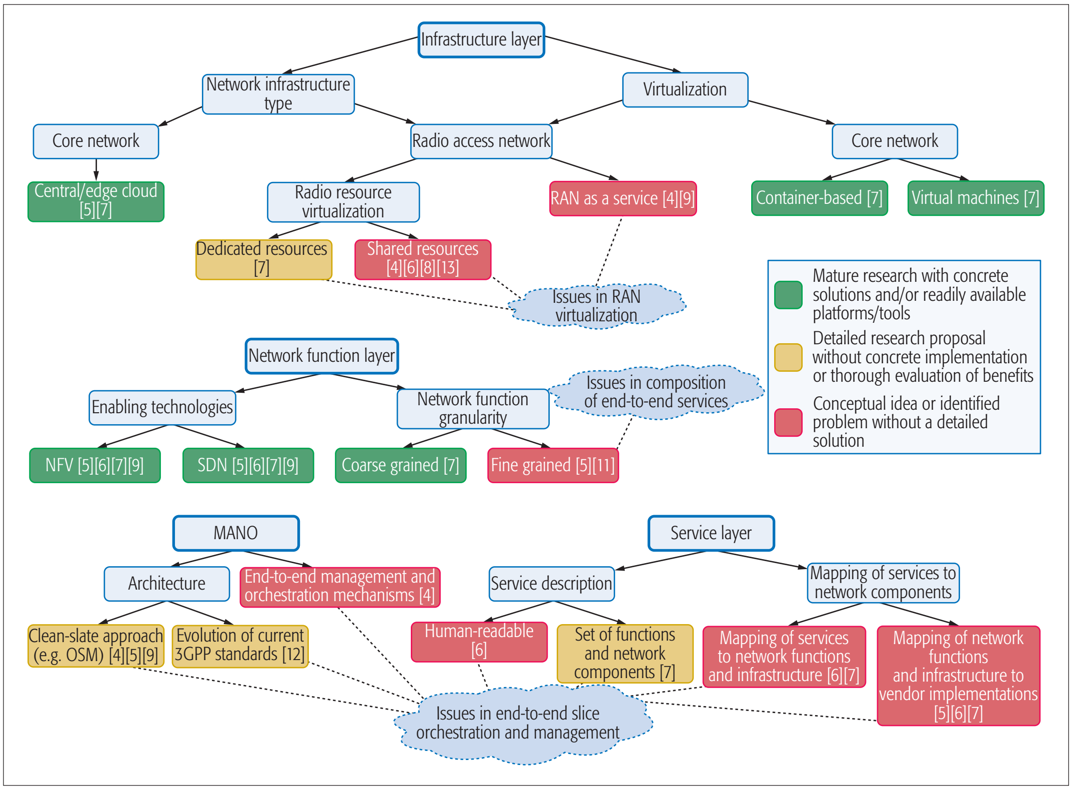


Figure 3. Maturity level of various aspects of 5G network slicing research.

a more radical clean-slate approach where the slice management and orchestration will be implemented as an application over an SDN controller, which will oversee both the wired and wireless domains [4, 5, 9]. Concrete implementations of MANO reference frameworks such as Open Source MANO[1] (OSM) are already appearing, which enable experimental studies on end-to-end 5G network slicing.

Mapping of Services to Network Components: Another very important issue is how to map and stitch together the components that are available to the various layers of the architecture in order to compose an end-to-end slice. Two types of mapping have been considered:

- The functional/SLA mapping of the service requirements to network functions and infrastructure types
- The mapping of network functions and infrastructure types to vendor implementations [6, 7]

The first type of mapping refers to the way that MANO chooses appropriate high-level network elements which are required to create a slice for a given service in order to meet its functional requirements and SLA. For example, if a slice has a need to cover devices over a wide area without any capacity concerns, choosing an LTE deployment with macrocells might be a good option. For this mapping, it has been proposed that the available infrastructural elements and network functions should reveal their capabilities to the MANO in a form of meta-data, describing the types of services that they can support [6].

Once the type of functions and infrastructural elements required for the slice have been identified, there is a need for a further mapping of these elements to concrete vendor implementations. Depending on the implementation of a function by a vendor, different levels of services can be offered. For example, alternative software implementations of an LTE eNodeB could provide support for a different number of users, with different performance guarantees or even different capabilities (e.g., flexible modification of the MAC scheduler). Here too, the high-level solution to this problem seems to be the use of meta-data in the elements provided by the vendors of the functions and the infrastructure. Such meta-data could describe both the capabilities of the vendor-specific functions and hardware [5, 6] as well as their deployment and operational requirements (connectivity, supported interfaces, and infrastructural key performance indicator [KPI] requirements) [6, 7], providing the MANO with sufficient information to perform the best possible configuration for the slice.

Challenges

From the last section, it is apparent that 5G network slicing has already received a fair amount of attention from the research community and industry. At the same time, there are several aspects key to end-to-end network slicing that are not well understood, as captured by the illustration in Fig. 3. With this in mind, we now elaborate on several significant outstanding challenges that need to be addressed to fully realize the vision of network-slicing-based multi-service softwarized 5G mobile network architecture.

RAN Virtualization

As already discussed earlier, the main challenges for infrastructure virtualization lie in the RAN. Solutions that pre-allocate distinct spectrum chunks to virtual base station instances (slices)

[1] https://osm.etsi.org/

are straightforward to realize and provide radio resource isolation but have the downside of inefficient use of radio resources. The alternative dynamic and fine-grained spectrum-sharing-based RAN virtualization approach does not have this limitation and therefore is desirable. However, ensuring radio resource isolation is a challenge for this approach. This can potentially be addressed by adapting SD-RAN controllers like [13]. As 5G networks are expected to span multiple RATs (including emerging technologies, e.g., 5G new radio and narrowband Internet of Things, NB-IoT), it is vital for RAN virtualization solutions to be able to accommodate multiple RATs. This presents an additional outstanding challenge as it is unclear whether multiple RATs can be multiplexed over the same possibly specialized hardware, or each needs its own dedicated hardware; the answer to this question might depend on the set of RATs under consideration.

From the RAN virtualization viewpoint, realizing the RaaS paradigm is another major challenge over and above the ones outlined above. This is a significant step over the notion of RAN sharing that involves sharing of radio resources among tenants, such as mobile virtual network operators (MVNOs), via a physical mobile network operator; various solutions for RAN sharing exist (e.g., [8, 13]). The RaaS paradigm requires going beyond radio resource and physical infrastructure sharing [4, 9] to have the capability to create virtual RAN instances on the fly with tailored sets of virtualized control functions (e.g., scheduling, mobility management) to suit individual slice/service requirements while at the same time ensuring isolation between different slices (virtual RAN instances).

Service Composition with Fine-Grained Network Functions

Ease of composing a service out of the available network functions is, as discussed earlier, directly dependent on the granularity of these functions. Coarse-grained functions are easy to compose as fewer interfaces need to be defined to chain them together, but this comes at the cost of reduced flexibility for the slices to be adaptable and meet their service requirements. Fine-grained network functions do not have this limitation and are more desirable. However, we lack a scalable and interoperable means for service composition with fine-grained functions that could be implemented by different vendors. The straightforward approach of defining new standardized interfaces for each new function is not scalable as the functions increase in number and the granularity becomes finer.

End-to-End Slice Orchestration and Management

A significant challenge for the realization of a network slice is how to go from high-level description of the service to the concrete slice in terms of infrastructure and network functions. The problem of describing services has already been identified in the literature but without satisfactory resolution. A good approach to address this void is to develop domain-specific description languages that allow the expression of service characteristics, KPIs, and network element capabilities and requirements in a comprehensive manner while retaining a simple and intuitive syntax (e.g., in the philosophy of [14]). Two important features that such languages should inherently provide are the flexibility/extensibility to accommodate new network elements that may appear in the future (e.g., new network functions, new RATs) and the applicability to be used in multi-vendor environments. A desirable feature would also be the capability to compose complex rules and expressions out of simpler ones, introducing abstraction layers in the expression of service requirements.

As noted earlier, concrete MANO frameworks like OSM have emerged in recent years. While such platforms are essential to flexibly realize network slices end-to-end and as needed, there's a more significant challenge that is only starting to be addressed. This concerns holistic orchestration of different slices so that each meets its service/SLA requirements while at the same time efficiently utilizing underlying resources. This calls for a sophisticated end-to-end orchestration and management plane. Such a plane should not be limited to trivial slice generation that does mapping of slices to network components and statically allocates them resources. Instead, it should be adaptive, ensuring that the performance and resiliency requirements of the deployed services are met. To achieve this, it should efficiently and holistically manage resources by making decisions based on the current state of slices as well as their predicted state/demands in the near future [4].

Such issues have been thoroughly investigated in the context of cloud computing and data centers, where many concrete solutions have already been proposed (e.g., [15]). While underlying principles from these other contexts can be leveraged, mechanisms targeting 5G network slicing should be suitably adapted and extended considering additional types of resources. Specifically, not just the resources found in cloud environments (memory, storage, network), but also radio resources need to be included, considering their correlation and how adjusting one resource type could have a direct effect on the efficiency of another, and therefore on the overall service quality. The problem of meeting requirements of different services while efficiently managing underlying network resources in the 5G network slicing context is also somewhat analogous to quality of service (QoS) provisioning in the Internet.

Summary

We have presented what we believe to be the first survey of the state of the art in 5G network slicing. To this end, we have presented a common framework for bringing together and discussing existing work in a holistic and concise manner. This framework essentially groups existing slicing proposals according to the architectural layer they target; that is, the infrastructure, network function, and service layers along with the MANO entity. With respect to this framework, we have evaluated the maturity of current proposals and identified remaining gaps. While several aspects of network slicing at the infrastructure and network function layers are quickly maturing, issues such as virtualization in the RAN are unresolved. Also, approaches for realizing, orchestrating, and managing slices are still in their infancy with many open research questions.

While several aspects of network slicing at the infrastructure and network functions layers are quickly maturing, issues such as virtualization in the RAN are unresolved. Also, approaches for realizing, orchestrating, and managing slices are still in their infancy with many open research questions.

REFERENCES

[1] NGMN Alliance, 5G White Paper, Feb 2015.

[2] ITU-R, "IMT Vision — Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," Sept. 2015.

[3] 5G PPP Architecture Working Group, "View on 5G Architecture," July 2016.

[4] I. F. Akyildiz, P. Wang, and S. Lin, " SoftAir: A Software Defined Networking Architecture for 5G Wireless Systems," *Computer Networks*, vol. 85, no. C, 2015.

[5] P. Rost *et al.*, "Mobile Network Architecture Evolution Toward 5G," *IEEE Commun. Mag.*, vol. 54, no. 5, May 2016.

[6] X. Zhou *et al.*, "Network Slicing As A Service: Enabling Enterprises' Own Software-Defined Cellular Networks," *IEEE Commun. Mag.*, vol. 54, no. 7, July 2016.

[7] N. Nikaein *et al.*, "Network Store: Exploring Slicing in Future 5G Networks," Proc. *10th ACM Int'l. Wksp. Mobility in the Evolving Internet Architecture*, Sept. 2015.

[8] X. Costa-Pérez *et al.*, "Radio Access Network Virtualization for Future Mobile Carrier Networks," *IEEE Commun. Mag.*, vol. 51, no. 7, July 2013.

[9] A. Banchs *et al.*, "A Novel Radio Multiservice Adaptive Network Architecture for 5G Networks," *Proc. 81st IEEE VTC-Spring '15*, May 2015.

[10] M. Sama *et al.*, "Reshaping the Mobile Core Network via Function Decomposition and Network Slicing for the 5G Eram," *Proc. IEEE WCNC*, Apr 2016.

[11] F. Z. Yousaf *et al.*, "Cost Analysis of Initial Deployment Strategies for Virtualized Mobile Core Network Functions," *IEEE Commun. Mag.*, vol. 53, no. 12, Dec. 2015.

[12] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From Network Sharing to Multi-Tenancy: The 5G Network Slice Broker," *IEEE Commun. Mag.*, vol. 54, no. 7, July 2016.

[13] X. Foukas *et al.*, "FlexRAN: A Flexible and Programmable Platform for Software-Defined Radio Access Networks," *Proc. ACM CoNEXT*, Dec. 2016.

[14] Q. Zhou *et al.*, "Network Virtualization and Resource Description in Software-Defined Wireless Networks," *IEEE Commun. Mag.*, vol. 53, no. 11, Nov. 2015.

[15] D. Shue, M. J. Freedman, and A. Shaikh, "Performance Isolation and Fairness for Multi-Tenant Cloud Storage," *Proc. 10th USENIX Symp. Operating Systems Design and Implementation*, Oct. 2012.

BIOGRAPHIES

XENOFON FOUKAS (x.foukas@sms.ed.ac.uk) is a Ph.D. student in computer science at the University of Edinburgh. He received his B.Sc. degree from Athens University of Business and Economics (AUEB) in 2012 and his M.Sc. degree in advanced computing from the Department of Computing at Imperial College London in 2013. His research interests are in the areas of mobile/wireless networking and distributed systems. The focus of his current work is on next-generation and software-defined mobile networks.

GEORGIOS PATOUNAS (gpatounas@simula.no) is a Ph.D. candidate in computer networks with Simula Research Laboratory, Oslo, Norway. He received his B.Sc. in computer systems engineering from the University of Applied Sciences of Piraeus, Greece, and his M.Sc. in distributed networks from the University of Oslo. His current work focuses on resilience, scalability, and evolution of wired as well as wireless and mobile networks.

AHMED ELMOKASHFI (ahmed@simula.no) is a senior research scientist at Simula Research Laboratory. He received his Ph.D. degree from the University of Oslo in 2011, his M.Sc. from Blekinge Institute of Technology in 2007, and his B.Sc. degree in electrical and electronics engineering from the University of Khartoum in 2003. His research focuses on the measurement and quantification of robustness in mobile broadband networks; the resilience, scalability, and evolution of the Internet infrastructure; and the understanding of dynamical complex systems.

MAHESH K. MARINA (mahesh@ed.ac.uk) is a reader in the School of Informatics at the University of Edinburgh. Before joining Edinburgh, he had a two-year postdoctoral stint at the University of California Los Angeles Computer Science Department. He received his Ph.D. in computer science from the State University of New York at Stony Brook. More information about him and his current research can be found at http://homepages.inf.ed.ac.uk/mmarina/.

5G-ICN: Delivering ICN Services over 5G Using Network Slicing

Ravishankar Ravindran, Asit Chakraborti, Syed Obaid Amin, Aytac Azgin, and Guoqiang Wang

Abstract

The challenging requirements of 5G, from both the applications and architecture perspectives, motivate the need to explore the feasibility of delivering services over new network architectures. As 5G proposes application-centric network slicing, which enables the use of new data planes realizable over a programmable compute, storage, and transport infrastructure, we consider information-centric networking as a candidate network architecture to realize 5G objectives. This can coexist with end-to-end IP services that are offered today. To this effect, we first propose a 5G-ICN architecture and compare its benefits (i.e., innovative services offered by leveraging ICN features) to current 3GPP-based mobile architectures. We then introduce a general application-driven framework that emphasizes the flexibility afforded by network functions virtualization and software defined networking over which 5G-ICN can be realized. We specifically focus on the issue of how mobility as a service (MaaS) can be realized as a 5G-ICN slice, and give an in-depth overview on resource provisioning and inter-dependencies and coordination among functional 5G-ICN slices to meet the MaaS objectives. The article tries to show the flexibility of delivering services over ICN where virtualization of control and data plane can be used by applications to meet complex service logic execution while creating value to its end users.

Introduction

The key driving factors for the fifth generation (5G), which have been laid out in [1], include:

- Support for high-density Internet of Things (IoT) devices and services with very stringent end-to-end requirements (e.g., latency of 1–10 ms)
- Support for very high throughput, with the average being 50 Mb/s in all urban conditions, and peaking at 1–10 Gb/s in ideal conditions
- Support for a new service class consisting of tactile applications that simultaneously carry low latency and high reliability requirements

Accordingly, a significant differentiator operators seek in 5G is the transition toward a service-centric infrastructure[1] that is also capable of fostering new business models between operators and the popular over-the-top (OTT) providers. These factors along with a shift in communication patterns from connecting hosts to efficient dissemination of information [2] considering security and mobility requirements motivate the need to evaluate new network architectures (other than the currently used IP networking).

Current research efforts on 5G network architecture adopt two different views.

The first view, as considered in the 5G Public Private Partnership (5GPPP) project [3], proposes a 5G architecture with focus on the evolved radio access network (RAN), while preserving 4G's core network architecture but over a flexible network functions virtualization/software defined networking (NFV/SDN)-based infrastructure. However, adopting 4G network architecture to 5G also means to inherit the drawbacks of the current IP architecture, with respect to:

- Complex core networking based on tunnelling technology to support mobility
- Security challenges leading to high signaling costs
- Lack of multihoming support
- A networking infrastructure that does not leverage the agility of cheap computing and storage resources in the transport infrastructure

The second view, as considered by the International Telecommunication Union's (ITU's) 5G focus group FG-IMT2020's Phase-1 [4] and 5G-America's Phase-2 [5] works, acknowledging the heterogeneous service requirements, discusses the benefits of architectures like information-centric networking (ICN) with inherent support for features like name-based networking, in-network storage, edge computing, security, and mobility in 5G. This is made feasible in 5G by the proposed network softwarization and the ability to slice the endpoints, access and core transport, and compute and storage resources, among multiple services [1] with heterogenous service guarantees. Within the context of such 5G architecture that is driven by a network slicing framework, ICN can be realized as a slice comprising physical and virtual resources over which the services can be delivered.

In this article, adopting the second view, we discuss a 5G-ICN architecture based on the NFV/SDN framework to realize a top-down service-centric platform, in which the ICN-based service delivery platform becomes a natural extension of the cloud into the infrastructure. This is made possible as:

- ICN merges compute, storage, and network virtualization on the same resource platform.
- An ICN-based service delivery can orchestrate complex service logic execution by service function placement and content processing at

The challenging requirements of 5G, from both the applications and architecture perspectives, motivate the need to explore the feasibility of delivering services over new network architectures. As 5G proposes application-centric network slicing, which enables the use of new data planes realizable over a programmable compute, storage, and transport infrastructure, the authors consider information-centric networking as a candidate network architecture to realize 5G objectives.

[1] By this, we mean an infrastructure that is operated in a top-down manner using a service aware management, control, and data plane.

The authors are with Huawei Research Center.

Digital Object Identifier: 10.1109/MCOM.2017.1600938

ICN allows different application-centric naming schemas, such as human-readable, self-certified or a hybrid one. Self-certified names offer another desirable property, that is, authentication of hosts, services, devices or contents with minimal signalling cost [13]. Such state can be managed at the BS to authenticate upstream or downstream transmissions.

the extreme edges of the network (e.g., base stations [BSs]) while being extendible to commodity utilities (e.g., lampposts or traffic lights). Furthermore, as ICN has proven its usefulness in constrained and ad hoc infrastructures [6, 7], it also represents an ideal platform to deliver unified IoT services [8] over the 5G framework. We exemplify the benefits of such a service platform by considering the case of delivering mobility as a service (MaaS) over a converged programmable infrastructure that enables network slicing.

The remaining sections are laid out as follows. We provide a brief introduction to ICN, followed by a discussion on a 5G-ICN architecture, with focus on accommodating IoT services and applications with high bandwidth requirements. Here, we explain the features enabled by this architecture considering the current 3GPP systems, and various 5G-ICN deployment scenarios. We introduce a generalized network slicing framework, over which both IP and 5G-ICN services can be delivered. We present a use case study of realizing the 5G-ICN architecture over a network slicing framework, where we discuss how MaaS can be realized. We emphasize how to bootstrap different ICN network and service slices and their interactions to achieve the MaaS objectives. We then present our final remarks.

Information-Centric Networking

ICN [2] is a result of various future network architecture research studies pursued in various parts of the world that enable features such as:

- Name-based networking of resources corresponding to contents, services, devices, and network domains
- Session-less transport through per-hop name resolution (of the requested resource), which also enables 5G-targeted features such as mobility, multicasting, and multi-homing
- Exploiting transport embedded compute-storage resources that are virtualizable among heterogenous services
- Network layer security, which allows one to authenticate user requests and the returned content objects, thereby allowing location-independent caching and computing as desired by ICN applications and infrastructure providers
- Suitability to both ad hoc and infrastructure-based IoT environments, where the information-centric nature of IoT applications matches what the ICN infrastructure offers

Although ICN has been an active area of research, many research challenges still remain [9] in various aspects of the architecture and business feasibility.

We next discuss an ICN-based 5G architecture capable of leveraging these features to enable a service-oriented network architecture.

5G-ICN Architecture

5G presents a great opportunity for introducing new network architectures to address service requirements that are difficult to satisfy with current IP networking. The need for a new network architecture can be justified based on the following important benefits:

- To address the issue of having a single protocol that can handle mobility and security instead of having a diverse set of IP-based 3GPP protocols (as is the case for the current cellular systems)
- To serve as a unifying platform with the same layer 3 (L3) application programming interfaces (APIs) to integrate heterogenous radios (e.g., Wifi, LTE, 3G) and wired interfaces over which devices and services connect to the network
- To converge computing, storage, and networking over a single platform, which improves the flexibility of enabling virtualized service logic and caching functions anywhere in the network (especially for the access segment)

All of these benefits can be achieved by a 5G architecture based on ICN (i.e., 5G-ICN).

In the data plane, 5G-ICN is capable of realizing a flat architecture without specialized gateways, as shown in Fig. 1, where applications and devices seek connectivity through the RAN to ICN gateways (resulting in a service-enabled RAN or SE-RAN). The SE-RAN is optimized to deliver services with both high bandwidth and strict latency requirements. This requires an enhanced cloud RAN implementation to support 5G-ICN architectural features, called next generation cloud RAN (NG-RAN). In Fig. 1, we refer to the edge ICN routers as ICN service routers (ICN-SRs), as these nodes are equipped with additional compute, storage, and bandwidth resources to be shared among services. The inherent ICN capabilities (e.g., in-network caching and computing, and multi-homing features) enable 5G-ICN to naturally offer support for high bandwidth applications. Similarly, the suitability of ICN to support IoT applications [10] (due to exploiting, for instance, naming, device-to-device communications, contextual networking, and self-X features such as configuration, management, and healing) allows IoT services to be efficiently delivered using the same protocol infrastructure. The 5G-ICN segment for IoT services can be supported with distributed middleware service functions over ICN (e.g., device and service discovery, naming, context processing, and publish/subscribe service) features required for the IoT systems.

In control and service planes, considering the service-oriented networking requirement as stated by [1], 5G-ICN naturally lends itself to service virtualization through a global service orchestrator. This can be realized through a logically centralized service plane, which abstracts resources from domain-level orchestrators that monitor, manage, and abstract ICN infrastructure resources within each domain. To support data plane virtualization, we consider sharing cache, compute, and networking resources within ICN routers among multiple services using compute virtualization such as virtual network functions (VNFs),[2] a P4 framework [11], or logical resource partitioning of a physical or software ICN router (similar to VPN technologies over IP today).

Additionally, 5G-ICN can meet several requirements that do not exist in current cellular network architectures such as LTE [12], and these include:

Naming: Applications today conflate IP addresses as identifiers; hence, it becomes difficult to support session mobility or achieve multihoming in an IP architecture. On the other hand, ICN applications bind to persistent names that are used to identify hosts, contents, or services. Naming resources

[2] Virtual network functions are virtual machines (VMs) or containers required to support specific logical functions of the network such as IP/ICN forwarding, fronthaul/backhaul RAN processing, and generic middlebox functions.

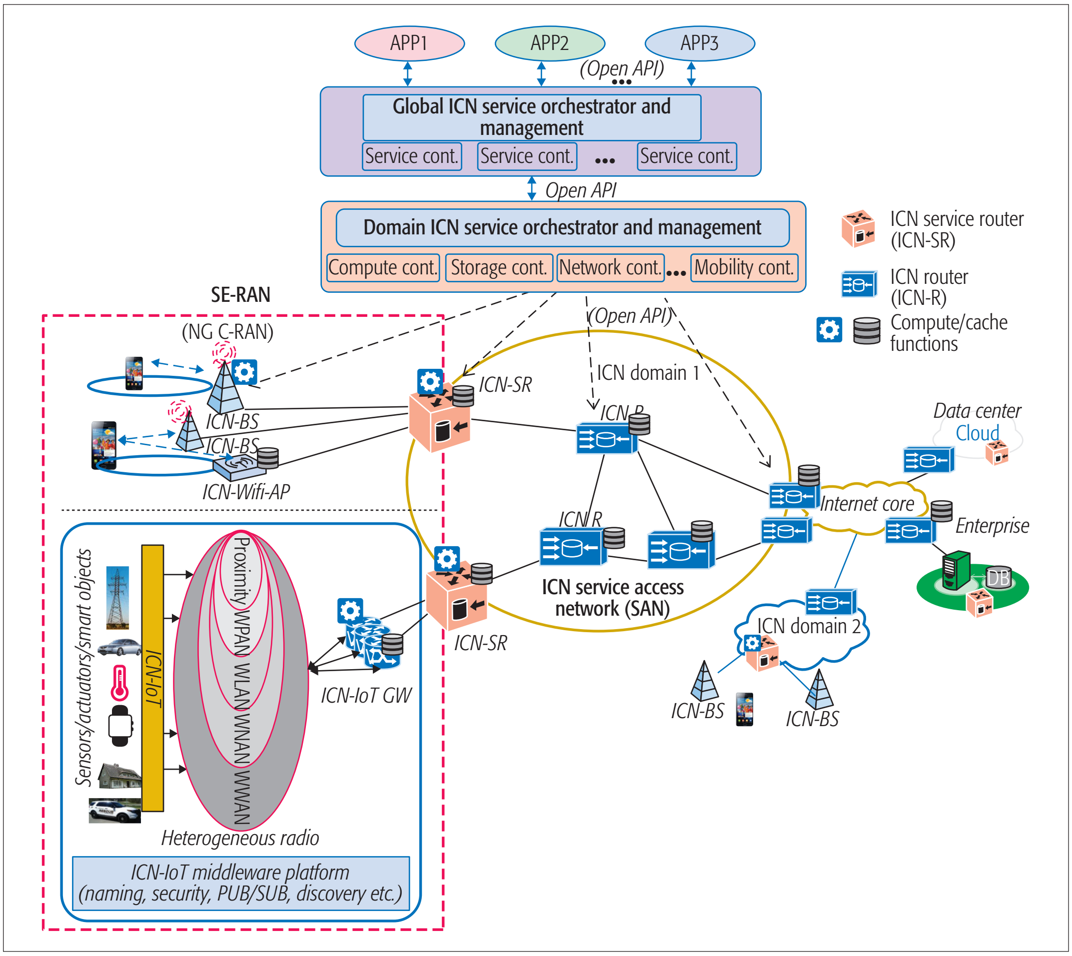


Figure 1. 5G-ICN architecture.

insulates applications from any kind of host mobility or even service mobility, as the ICN layer handles the mapping from the high-layer application identifiers to network identfiers. ICN allows different application-centric naming schemas, such as human-readable, self-certified, or a hybrid one. Self-certified names offer another desirable property, that of authenticating hosts, services, devices, or contents with minimal signaling cost [13]. Such state can be managed at the BS to authenticate upstream or downstream transmissions.

Mobility: As shown in Fig. 1, ICN enables a flat architecture, where mobility can be handled in a distributed manner by the point of attachment (PoA) nodes, which in our architecture can be the ICN base station (ICN-BS) or the ICN-SR integrating multiple radios. On the other hand, mobility in LTE [12] is handled by an orthogonal set of protocols. Specifically, in LTE, a per-user bearer tunnel state is created between the LTE-BS (eNodeB) and the Evolved Packet Core (EPC) containing the service gateway (S-GW) and the packet data network (PDN) gateway (P-GW), through which the user equipment's (UE's) incoming/outgoing traffic is tunneled, as the UE hands over from one eNodeB to another. As the amount of tunnel state required to handle mobility in the data plane is proportional to the number of UEs, signaling overhead increases with host dynamicity.

Security: In current LTE systems, it typically takes 60 ms for a device to go from idle state to active state before sending or receiving any data [14], which is due mostly to UE authentication and signaling of the bearer paths — between eNodeB, S-GW, and P-GW — for the UE traffic. In the case of ICN, application APIs for Interest/Data traffic[3] bind identity information to enable security features such as content integrity and provenance validation. Also, the primitives associated with ICN traffic can be contextualized with additional security attributes such as device or user identity, which can be subjected to in-network security verification.

Reliability: The features inherently offered by ICN, such as sessionless store-and-forward operation, per-hop name resolution,[4] per-hop congestion control, multi-homing, replicated caching, and multi-path routing, allow quick and painless recovery from congestion scenarios or link failures in manners not achievable with IP networking.

Efficiency: ICN offers efficiency at every level, from the data plane to the control and management planes. Data plane efficiency is achieved through replicated computing, caching, and storage in the network, thereby reducing costs associated with upstream bandwidth, latency, storage, and computation uses. Control plane efficiency is achieved through name-based networking that allows different modes of control plane techniques based on the networking environment, such as the flooding mechanism in the case of device-to-device communications for ad hoc scenarios, or leveraging traditional routing mechanisms to enable unicast, multicast, or anycast in the infrastructure. Management plane efficiency is achieved by sev-

The features inherently offered by ICN, such as sessionless store-and-forward operation, per-hop name resolution,4 per-hop congestion control, multi-homing, replicated caching, and multi-path routing, allow quick and painless recovery from congestion scenarios or link failures in manners not achievable with IP networking.

[3] Note that the Interest/Data primitives belong to CCN/NDN [15], but are similar to that of other ICN protocols such as MobilityFirst [13].

[4] This applies to both Interest and Data.

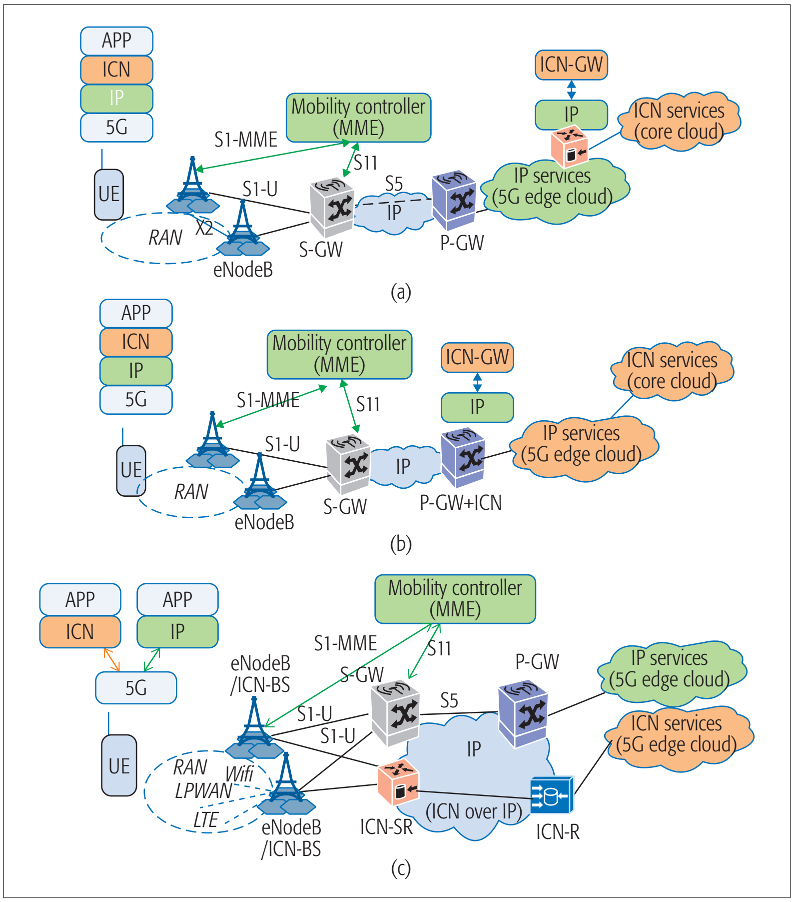


Figure 2. ICN deployment models: a) overlay model; b) integrated model; c) flat model.

eral Self-X features that ICN provides, such as per-hop congestion control, multi-path routing, and minimal overhead for bootstrapping.

Contextual Communication: ICN APIs are service-centric and contextualized by nature. For instance, consumers can make requests for a content using optional contextual metadata (i.e., location, device, or criticality of information request or response), which can then be subjected to in-network processing at the ICN routers or through overlaid virtual service functions (VSFs)[5] for further processing. In doing so, these contextualized requests can be satisfied at the network edge, which can improve UEs' quality of experience (QoE) to meet the service objectives.

5G-ICN Deployment Models

We consider three possible 5G-ICN deployment scenarios, which are shown in Fig. 2 and discussed next using the LTE [16] architecture as a reference.

The overlay model is shown in Fig. 2a, in which ICN becomes an overlaid service over the current IP infrastructure. Despite being overlaid, ICN can still be realized as a service platform that is managed by the operator, while offering caching and compute benefits to heterogenous applications through the edge and core cloud infrastructure (which can serve a large regional geography). .

The integrated model is shown in Fig. 2b, in which ICN becomes tightly integrated with the core mobile network infrastructure. This model assumes an explicit control and management plane to relay ICN packet data units (PDUs) from UE over 5G to the P-GW, which hosts the ICN router. Hence, for an ICN service slice, ICN service flows can choose to route to different ICN P-GWs based on the service requirements. In this scenario, mobility is handled by the underlay 5G protocol, but the benefits of caching and computing are distributed within the core infrastructure and closer to UE. In this model the ICN channel could be enabled using the same control and signaling infrastructure of the LTE.

The flat model is shown in Fig. 2c, in which the 5G-ICN architecture (discussed earlier) is integrated within the network to take advantage of all the benefits offered by 5G-ICN. As ICN integrates security and anchorless mobility, other overlay protocols are not required in the ICN infrastructure. Even if control functions such as mobile management entity (MME), home subscriber service (HSS), or policy and charging rule functions (PCRFs) are required, they can be adapted to the ICN network after accounting for the features enabled by ICN in the network layer.

With the above view of the 5G-ICN architecture, we discuss how 5G-ICN can be realized in a framework capable of offering a network slicing service.

Network Slicing Architecture

Slicing a 5G network on an end-to-end basis, which spans multiple technology domains and includes the UE resources, aims to support a diverse set of applications with different service requirements (e.g., latency, bandwidth, and reliability) using a common resource pool consisting of compute, storage, and bandwidth resources. Figure 3 shows a generic network slicing architecture with capability to create IP and 5G-ICN service slices. The framework has the following five functional planes (FPs):

`FP1`— **Service Business Plane** forms the interface between the external 5G service users and the softwarized infrastructure that helps realize connectivity and user-centric services in a dynamic manner. `FP1` exposes various service APIs, which the network is capable of delivering through a formal intent model, along with service management and monitoring functions. The business plane APIs can be realized with a high-level abstraction language, through which the service expresses:

- Its objectives (i.e., what it wants to accomplish)
- The network services required to accomplish it

For instance, service input can include service type, demand patterns, and requirements on service level agreement (SLA)/quality of service (QoS)/QoE and network services such as reachability, security, mobility, multicasting, and storage. `FP1` converts these requirements into information-and-data models as required by `FP2`.

`FP2` — **Service Orchestration and Management Plane**, upon receiving the service requests from `FP1` — with explicit information on the narrow waist to use for service delivery — communicates the service requirements to the respective IP/ICN global service orchestrators for their execution.

`FP3` — **IP/ICN Global Orchestrator** realizes IP and ICN services by leveraging the already existing slices (if necessary). In Fig. 3, IP and ICN service orchestrators are logically separated, as the network and services operate on different data, control, and service plane APIs. `FP3` interfaces with

[5] Virtual service functions are responsible for executing specific service logics related to the services, or generic ones to aid services such as with service discovery and naming services.

domain controllers to virtualize compute, storage, and network resources to meet the service requirements, with the help of the following functions:

- Translating service requirements to resource requirements in the data plane, and identifying different VNF/VSFs required to support the given service, while generating a slice context for service, control, and data plane management
- Monitoring compute, storage, and network resources at the edge and core clouds, and transport segments[6]
- Keeping an abstract view of the physical (topology) resources and its mapping in the context of multiple slices
- Interfacing with technology-specific domain controllers (in FP4), to enforce the rules determined by the service orchestrator
- Handling global life cycle management of the VNF/VSFs, failure management, and network reliability based on the service layer agreement (SLA) requirements

FP4 — Domain Service Orchestration and Management support orchestration of IP and ICN services within domains. As end-to-end network segments will comprise multiple domains with differing technologies, ranging from 4G/5G RAN to optical/multiprotocol label switching (MPLS) transport domains, and from edge to central cloud resources, each of these domains will be governed by its own local network, compute, and storage controllers. ICN controllers are realized as sub-controllers in the ICN relevant domains. Functions such as domain slice SLA management, VNF/VSFs life cycle, and failure management are also handled by the domain controllers in coordination with the FP3 function.

FP5 — Infrastructure Plane distributed among multiple domains and managed by FP4, enables the service rules in an end-to-end manner, spanning the UEs, RAN, heterogenous transport segments, edge, and central clouds.

In the context of multiplexed IP/ICN flows, the transport plane should be able to differentiate among these flows to provide the appropriate resource guarantees. In the case of ICN, finer flow-level service differentiation depends on the ability of the network to understand the ICN primitives (i.e., name-based flows, resources to manage the multicast state,[7] request forwarding and software-defined cache management policies, in-network computing and context processing rules, and, QoS and queue management policies).

Software-defined radio (SDR) allows the realization of flexible and elastic MAC/PHY layers to cater to a diverse set of services, such as low-power IoT and high-bandwidth video applications. Various middlebox, data plane, control and service functions can be realized over the generic Intel-x86 infrastructure in the form of VNF/VSFs. Data plane can be based on virtual network overlays or deeply programmable hardware based on the P4 or OpenFlow technologies to multiplex IP/ICN service flows to achieve QoS isolation and line-rate switching. UE programmability enables mapping IP/ICN flows (e.g., using a host virtual switch) to appropriate radio slices and then handing over the service flows at the BS to the appropriate service slice.

Having provided a broad discussion of the network slicing framework capable of delivering both IP and ICN services, we next discuss the realization of MaaS as a 5G-ICN slice and its functional interaction with other 5G-ICN service slices.

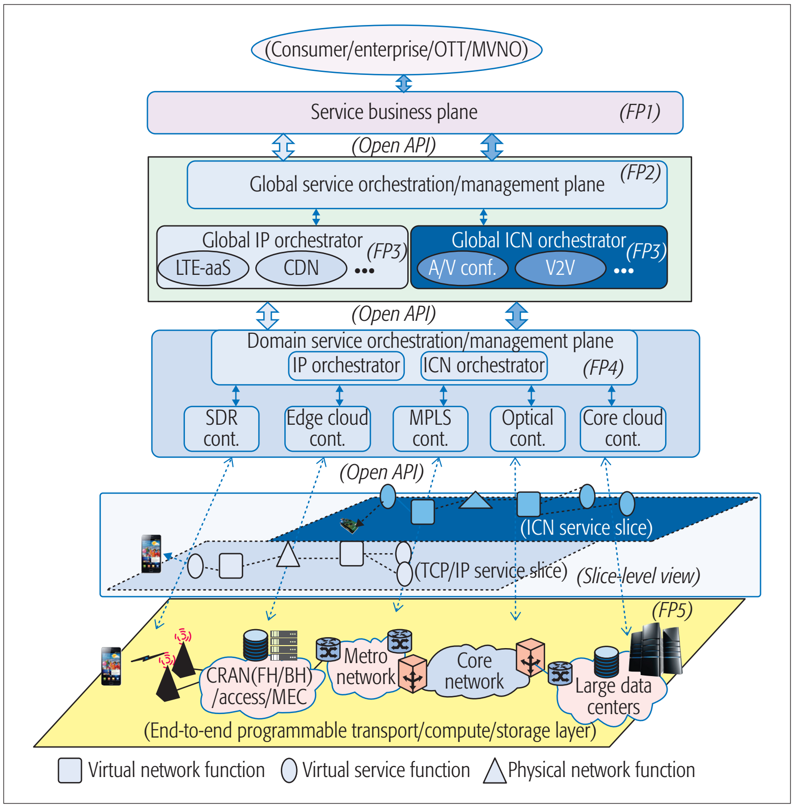


Figure 3. Network slicing framework.

5G-ICN Mobility as a Service

In this section, as a case study of realizing 5G-ICN within a network slicing framework, we discuss MaaS and how other service slices can use the APIs exposed by MaaS to enable mobility service to its dynamic entities. We consider the following objectives for MaaS:

- On-demand mobility allows ICN resource names belonging to any slice to be dynamically (de-)registered for the mobility service, and when registered, all the flows for that name are provided with seamless mobility support.
- Minimal session disruption is needed to provide seamless mobility support to service flows within a mobility enabled network slice, as a member UE moves from one PoA to another within the same slice.

Even though the specific details of handling mobility in ICN vary depending on the protocol, the general principle remains the same, that is, separating application name binding from the topological network names or addresses, which is also referred to as the ID/locator name split and the late-binding feature [17] that allows ICN PoA to redirect flows to a UE's new PoA in a dynamic manner.

For this case study, we assume the ICN network is realized as a set of virtual entities, such as using a *Container* technology; hence, the state within a virtual ICN forwarders consists of only the service states, such as cached items and name reachability state, that only remain for the lifetime of the virtual slice instance.

[6] In the case of ICN, compute and storage resources become part of the infrastructure.

[7] This includes both multiple user requests for the same content, or pushing a content to multiple receivers.

Even though the specific details of handling mobility in ICN vary depending on the protocol, the general principal remains the same, i.e., separating application name binding from the network address, which is also referred as the ID/locator name split and late-binding feature [17] that allows ICN PoA to redirect flows to UE's new PoA in a dynamic manner.

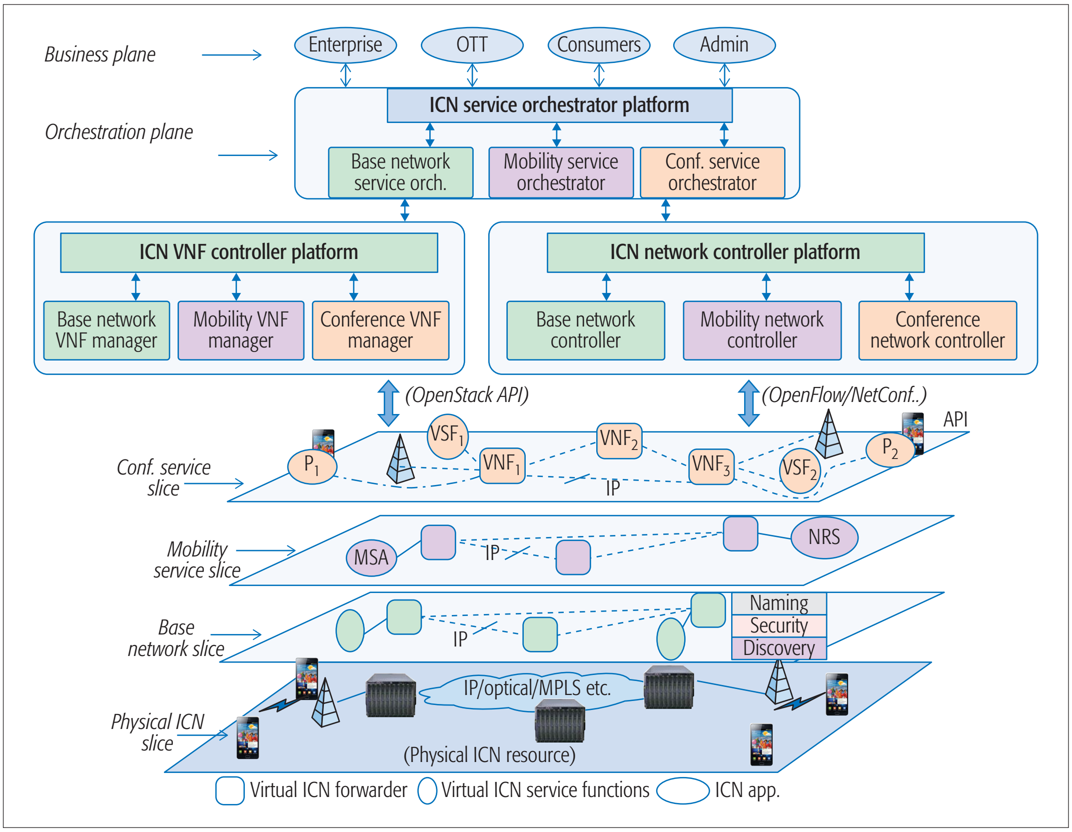


Figure 4. Mobility as a service realization.

MaaS Operation

We show the overall architecture that enables MaaS, which any service slice can leverage, in Fig. 4. This architecture is based on the discussion provided earlier. The application we assume is a videoconference service, in which the participants join the conference randomly and can solicit audio/video/text content from other desired parties dynamically. Here, the global and domain service controllers, and VNF/VSFs of each service expose the appropriate APIs that can be used by each other to achieve a service objective. We next explain how MaaS is realized considering the various stages of slice provisioning and interactions among the different slices.

Step 1: Base network slice bootstrap. To support ICN service virtualization, first, we need to bootstrap the service functions that enable UE applications to discover and name the services, provide security functions, and connect them to the appropriate service gateway. We call this the Base Network Slice, which is managed by the Base Network Slice VNF Orchestrator. ICN connectivity of this service slice among the various VNFs[8] and the UE is managed by the Base Network Slice Controller.

Step 2: Mobility network slice bootstrap. As the mobility service is bootstrapped by the Mobility Service Orchestrator, two important service functions are enabled in the corresponding slice:

- Mobility Service Agent (MSA), which exposes APIs for the name resolution
- Name Resolution Service (NRS), which maps the registered names to the corresponding locators in the network

For entities outside a domain, we assume the NRS to have APIs for inter-domain resolution. As NRS is a very critical component, the mobility compute and network controllers should ensure high availability for this service (see mobility network slice in Fig. 4).

Step 3: Creating a videoconferencing slice. As an external trigger for a videoconferencing instance arrives from the business plane to the Conference Service Orchestrator, it maps the conference requirements (i.e., location information, number of participants at a physical location, device types, etc.) to the ICN VNFs and VSFs with appropriate compute, cache, and bandwidth resources to manage the expected traffic load. The Conference Service VNF Manager provisions the requested set of virtual ICN forwarders and service functions to support the conference session. The Conference Network Controller manages the connectivity between the virtual forwarders and service functions, and maps the dynamically arriving participants and their requests to appropriate VNFs for load balancing. Also, appropriate forwarding rules are pushed into the VNF instances to handle the service flows (see the conference service slice in Fig. 4).

Step 4: UE application bootstrap. The ICN application at the UE discovers the service to connect to, through well-known APIs available for service discovery (which is provided by the Base Network Service Slice). The discovery results in the application receiving names, keys, and trust information, and connecting to the appropriate slice gateway in the conference slice. For instance, in Fig. 4, UE's application instance, P_1, connects to the gateway VNF_1 when the UE joins the conference session.

Step 5: Enabling dynamic mobility. Assume that an external trigger from the business plane

[8] The VNFs here are the virtual ICN forwarders.

requests mobility service for the participants in a given conference slice instance. This request is first received by the Conference Service Orchestrator, which invokes the service APIs provided by the Mobility Service Orchestrator that pushes the request to the Mobility Network Controller, which sets the appropriate policy state (depending on the specific ICN protocol) in the MSA and the NRS within the mobility network slice. At the same time, the Conference Service Orchestrator triggers the Conference Network Controller to register the mobile named entities of that slice to the NRS. The Conference Network Controller configures the ICN virtual forwarders to invoke resolution function to handle ICN flows bound to these mobile names. In short, for an incoming ICN request, the resolution request from the conference slice is passed to the MSA function in the mobility network slice, which then invokes the NRS for resolution to the mobile participant's current location.

Step 6: Handling seamless mobility. A late-binding mechanism (e.g., [17] for NDN) can be used by the conference slice in the data plane to handle seamless mobility of the participants and to achieve minimal session disruption for the voice/video sessions handled by the conference slice.

Note that, similar to the above scenario, mobility service can be disabled or enabled by a trigger in the business plane over any service slice.

Conclusion

In this article, we explore the feasibility of realizing future networking architectures, like ICN, under the network slicing framework proposed for 5G. We argue that while ICN simplifies the network architecture, it can help meet the heterogenous service objectives leveraging the several desirable features of ICN. We examine a potential 5G-ICN architecture, explaining the features it enables and the possible deployment models. Finally, we study how mobility as a service can be dynamically enabled as a service slice considering a 5G-ICN framework and how other service slices could leverage it to enable mobility to its resource entities.

References

[1] NGMN 5G White Paper, https://www.ngmn.org/ uploads/media/NGMN_5G_White_Paper_V1_0.pdf, accessed 02/20/2017.

[2] G. Xylomenos *et al.*, "A Survey of Information-Centric Networking Research," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 2, Feb. 2014, pp. 1024–49.

[3] 5GPPP Architecture Working Group, "View on 5G Architecutre," https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-5G-Architecture-WP-July-2016.pdf, accessed 02/20/2017.

[4] ITU FG-IMT 2020 Phase-1, "Network Standardization Requirement for 5G," http://www.itu.int/en/ITU-T/focusgroups/imt-2020/ Documents/T13-SG13-151130-TD-PLEN-0208!!MSW-E.docx, accessed 02/20/2017.

[5] 5G-Americas: Understanding Information Centric Networking and Mobile Edge Computing, http://www.5gamericas.org/files/3414/8173/2353/ Understanding_Information_Centric_Networking_and_ Mobile_Edge_Computing.pdf, accessed 02/20/2017.

[6] E. Baccelli *et al.*, "Information Centric Networking in the IoT: Experiments with NDN in the Wild," *Proc. 1st Int'l. Conf. Information-Centric Networking*, New York, NY, 2014, pp. 77–86.

[7] G. Grassi *et al.*, "Vanet Via Named Data Networking," *2014 IEEE INFOCOM Wksps.*, Apr. 2014, pp. 410–15.

[8] Y. Zhang *et al.*, "Requirements and Challenges for IoT over ICN," draft-zhang-icnrg-icniot-requirements-01.txt, IETF/IRTF/ICNRG, 2016.

[9] D. Kutscher *et al.*, "RFC 7927:ICN Research Challenges," IRTF, ICNRG, 2016.

[10] Y. Zhang *et al.*, "ICN Based Architecture for IoT," draft-zhang-icnrg-icniot-architecture-00.txt, IETF/IRTF/ICNRG, 2016.

[11] P. Bosshart *et al.*, "P4: Programming Protocol-Independent Packet Processors," *SIGCOMM Comp. Commun. Rev.*, vol. 44, no. 3, July 2014, pp. 87–95.

[12] 3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description."

[13] A. Venkataramani *et al.*, "Mobilityfirst: A Mobility-Centric and Trustworthy Internet Architecture," *SIGCOMM Comp. Commun. Rev.*, vol. 44, no. 3, July 2014, pp. 74–80.

[14] D. Singhal *et al.*, "LTE-Advanced: Handover Interruption Time Analysis for IMT-A Evaluation," *2011 Int'l. Conf. Signal Processing, Commun., Computing and Networking Technologies*, July 2011, pp. 81–85.

[15] V. Jacobson *et al.*, "Networking Named Conten," *Proc. 5th ACM Int'l. Conf. Emerging Networking Experiments and Technologies*, 2009, pp. 1–12.

[16] K. Dimou *et al.*, "Handover within 3GPP LTE: Design Principles and Performance," *2009 IEEE VTC-Fall*, Sept. 2009, pp. 1–5.

[17] A. Azgin *et al.*, "Seamless Producer Mobility as a Service in Information Centric Networks," *Proc. 3rd ACM Conf. Information-Centric Networking*, 2016, pp. 243–48.

Biographies

Ravishankar Ravindran (ravi.ravindran@huawei.com) is a principal staff researcher at Huawei Technologies. He has been conducting advanced telecommunications research for over 16 years. His current areas of research include future Internet architectures, 5G, the Internet of Things, software defined networking, and network functions virtualization. The future networking research includes NSF-funded future Internet proposals such as NDN/CCN, MobilityFirst, and XIA , which is in collaboration with external research groups, standard bodies, and academia. Prior to this role, he was part of the CTO office at Nortel, where he was part of the Advanced Technology Group focused on research areas such as control plane routing protocols related to IP/(G)MPLS, scheduling problems in 4G wireless, and end-to-end QoE/QoS engineering for multimedia. Over the course of his research, he has been part of 10 granted patents and more than 30 pending filings in various areas of networking technologies. He has more than 30 technical publications in conferences and journals. He received his Ph.D. in systems and computer engineering from Carleton University, Canada

.

Asit Chakraborti (asit.chakraborti@huawei.com) is a senior staff research engineer at Huawei Technologies. He has contributed to research work in the areas of information-centric networking, 5G, the Internet of Things, mobility, and routing. He has 10 technical publications in conferences and journals. Prior to this role, he worked in the areas of IP routing and high availability computing at Ericsson and Sun Microsystems. He received his Master's degree in computer science and engineering from the Indian Institute of Technology Madras.

Syed Obaid Amin (obaid.amin@huawei.com) is a staff researcher at Huawei Technologies. He received his Ph.D. in computer engineering from Kyung Hee University, South Korea, in 2009. His major research interests are future Internet, real-time systems, sensor networks, and network security. He is a scholarship holder of the Institute of Information Technology Assessment-South Korea and Kyung Hee University-South Korea. He is also a lifetime member of the Pakistan Engineering Council.

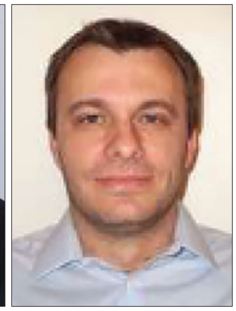
Aytac Azgin (aytac.azgin@huawei.com) is currently working as a researcher at Huawei Technologies, where he conducts research on future networking architectures. He received his B.S. degree in electrical and electronics engineering from Bilkent University, Turkey, and his M.S. and Ph.D. degrees in electrical and computer engineering from the University of Arizona, Tucson, and Georgia Institute of Technology, Atlanta, respectively. His research interests include network architectures and protocols, mobility management, scalable system design, multimedia communications and networking, and the Internet of Things.

Guo Qiang Wang (gq.wang@huawei.com) is a principal researcher at Huawei Technologies. He currently works on information-centric technologies for next generation Internet infrastructure and applications such as 5G and IoT. He is a subject matter expert with 27 years of networking experience for various products/technologies including content-centric switching/routing technologies, IP/MPLS routing and protection, DWDM switching, GMPLS control plane protocols, WiMAX mobility, and roaming.

The ICN application at the UE discovers the service to connect to, through well known APIs available for service discovery (which is provided by the Base Network Service Slice). The discovery results in application receiving names, keys and trust information, and connecting to the appropriate slice gateway in the conference slice.

TELECOMMUNICATION STANDARDS EDUCATION

Tarek S. El-Bawab



Periklis Chatzimisios

David G. Michelson

Interest in standards education has been rising in the science, technology, engineering, and mathematics (STEM) fields in recent years. Technical standards are formal documents that establish uniform criteria, methods, and practices through accredited and consensus processes in numerous areas within the STEM domain. They ensure equipment/system interoperability and are catalysts for technological innovation and global market competition.

Standards and standardization processes are not traditionally incorporated into university curricula. In some 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 advance standards education. This situation results in a knowledge gap among STEM graduates and professionals, which is impacting the ability of our workforce to face emerging global challenges.

Telecommunication is a field of substantial standardization activities. Although standards constitute a cornerstone for the telecom industry, students in telecom-related disciplines and most telecom professionals experience little knowledge, if any, of standards. Several efforts are underway today to deal with this shortcoming, to integrate standards education into university curricula, to explore innovative instruction methods that are suitable for standards education, and to develop standards education programs for professionals.

This Feature Topic provides an ensemble of standards' education projects, activities, and experiences, and opens a window of opportunity for the reader of *IEEE Communications Magazine* to learn about them. The variety of activities reported herein is inspiring to academicians and to telecom industry professionals. They present models of telecom standards education that can be followed by others.

The Feature Topic comprises seven articles. The first article, Interactive Research-Based Instruction Strategies for Standards Education" by El-Bawab and Effenberger, reports on the project Integrating Standards into Telecommunication Engineering Education (ISTEE), which is funded by the U.S. National Institute of Standards and Technology (NIST). The project developed a specialized, broad-scope university course, Telecommunication Standards and Standardization Processes, and incorporates a research-based instruction strategy where simulation of standards development organization (SDO) meetings is utilized to enhance students' learning. The second article, "Learning Mobile Communications Standards through Flexible Software Defined Radio Base Stations" by Jimenez *et al.*, focuses on modern mobile communications and the need for standards education therein. This work, which is funded by several Spanish funding agencies, describes a Flexible Radio Access Mobile Environment Defined by Software (FRAMED-SOFT) approach to standards education. The third article, "Multidisciplinary Learning through Implementation of the DVB-S2 Standard," is from Belgium and is authored by Murillo, Van den Bergh, and others. It describes a project-based approach to teaching and learning based on existing standards. In the EAGLE project, students partially implement the European Telecommunications Standards Institute (ETSI) ratified Second Generation Digital Video Broadcasting over Satellite (DVB-S2) standard in transmitter systems. The fourth article, "Teaching Communication Technologies and Standards for the Industrial IoT? Use 6TiSCH!" by Watteyne *et al.*, describes a collaborative international effort to focus on the Internet Engineering Task Force's (IETF's) 6TiSCH as an education tool to empower students with competences in computer networking, embedded systems, wireless communications, and Industrial Internet of Things (IIoT) applications.

The following two articles address the needs of the telecom industry from university programs and discuss methods to bridge the gap between theory and practice. The fifth article in this Feature Topic, "Hands-On Education about Standardization: Is That What Industry Expects?," is also international in its authors' affiliations, and has been written by Katusic *et al*. The article underlines difficulties recent engineering graduates are facing in real life due to the fact that the education programs they graduate from rarely encourage them to solve standards-based problems. The article proposes a research-based approach where students can acquire hands-on experience by getting involved in industrial projects. Our sixth article, "Teaching Telecommunication Standards: Bridging the Gap between Theory and Practice" by Gelonch-Bosch *et al.*, views telecom standards as a platform for collaboration between industry and research institutions. The article proposes a software-defined radio approach to teaching wireless communication standards that benefits both sides while fulfilling certain student learning objectives.

The approach utilizes open source software and commercial of-the-shelf computers, and can easily be adopted by other education institutions. The last article in this Feature Topic, "ITU Spectrum Management Training Program: A Comprehensive Modular Framework for Formalized Professional Education" by Medeisis, discusses the International Telecommunication Union's (ITU) professional education Spectrum Management Training Program (SMTP), currently implemented under the auspices of the ITU Academy. This program is designed to give professional certification in radio spectrum management.

We would like to thank all colleagues who submitted manuscripts to this Feature Topic and those who served as reviewers. We received a record number of submissions for *IEEE Communications Magazine*'s Education and Training Series. We are hopeful that this Feature Topic, which is the first of its kind, will be enjoyed by the magazine's readership, and will stimulate further efforts by academia and industry to collaborate in telecom standards education.

BIOGRAPHIES

TAREK S. EL-BAWAB [SM] (telbawab@ieee.org) led the Telecommunication Engineering Education movement, resulting in recognition of telecom engineering as a distinct ABET-accreditable discipline. He was the first recipient of the ComSoc Education Award. His other research interests include networking and performance analysis. Before Jackson State University, he was with Alcatel-Lucent (United States), Colorado State University, and the University of Essex (United Kingdom), and led large-scale international telecom projects. He is a member of IEEE-Eta Kappa Nu; a ComSoc Distinguished Lecturer; the Editor of Springer's Series, Textbooks in Telecommunication Engineering; and has more than 70 publications. He is a member of the IEEE Educational Activities Board and ComSoc Educational Services Board. He has served IEEE/ComSoc in numerous other technical/leadership capacities.

PERIKLIS CHATZIMISIOS [SM] (peris@it.teithe.gr) serves as an associate professor, the director of the Computing Systems, Security and Networks (CSSN) Research Lab, 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 of more than 110 peer-reviewed papers on performance evaluation and standardization of mobile/wireless communications, the Internet of Things, big data, and vehicular networking. He is a member of the IEEE ComSoc Educational Services Board and the IEEE ComSoc Standardization Programs Development Board. He received his Ph.D. from Bournemouth University, United Kingdom, in 2005 and his B.Sc. from ATEITHE in 2000.

DAVID G. MICHELSON [M] (davem@ece.ubc.ca) leads the Radio Science Lab at the University of British Columbia, where his research focuses on wireless communications. He is a member of both the IEEE Standards Association and the IEEE Antennas and Propagation Standards Committee, and serves as Vice Chair of the IEEE P.211 Working Group. He served as ComSoc's Director of Education and Training from 2012 to 13, 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).

Interactive Research-Based Instruction Strategies for Standards Education: Project ISTEE

Tarek S. El-Bawab and Frank Effenberger

Efforts are underway to advance technical standards education into science, technology, engineering, and mathematics (STEM) university programs. The authors provide an overview of Project ISTEE: Integrating Standards into Telecommunication Engineering Education. They discuss the background and motivations behind this project and describe its tasks and deliverables.

ABSTRACT

Efforts are underway to advance technical standards education into science, technology, engineering, and mathematics (STEM) university programs. This article provides an overview of Project ISTEE: Integrating Standards into Telecommunication Engineering Education. We discuss the background and motivations behind this project and describe its tasks and deliverables. The project produces a university course in telecommunication standards and standardization processes. Evaluation results of two offerings of this course are reported and discussed. Our findings confirm that interactive research-based instruction strategies are particularly suitable for standards education. A model for standards education is developed that can be adopted by other STEM disciplines.

BACKGROUND

Lack of sufficient components of standards education in science, technology, engineering, and mathematics (STEM) curricula is an issue facing the U.S. higher education system. This issue results in a workforce with a knowledge gap in standards, a critical aspect of tomorrow's science, engineering, and economy. Standards control access to virtually every market in global commerce. *Integrating Standards into Telecommunication Engineering Education* (*ISTEE*) is a project to advance and integrate knowledge of standards into STEM education in general, and into telecommunication engineering education in particular [1, 2]. ISTEE focuses on telecom as a platform, and presents a model of standards education that can be adopted by other STEM disciplines. The project is a partnership involving academia, industry, and the U.S. National Institute of Standards and Technology (NIST) to work toward achieving this objective.

Three distinct approaches to technical standards education are possible. In the first approach, some elements of standards education are introduced into existing STEM courses. In the second approach, specialized courses in standards and standardization processes are designed and included in university curricula. These can be introduced as core or elective courses, depending on the needs of the discipline. In the third approach, the first two methods can be mixed in various parts.

Project ISTEE was launched as a result of efforts underway to revisit STEM education [3–8], efforts to integrate standards thereto,[1] and the recognition of telecommunication engineering as a distinct university education discipline by the Accreditation Board for Engineering and Technology (ABET) [9]. Telecom is a field that is particularly rich in standards and standardization activities, which makes a case for specialized course(s) in telecom standards. Progress in several emerging areas such as the Internet of Things (IoT), big data, and green information and communication technologies (ICT), all of which find strong roots in telecom, actually makes a case for further post-ISTEE efforts in telecom standards education.

This article provides an overview of Project ISTEE. In the following section, we briefly describe this project. In the third section, we present our evaluation and assessment results of the course offered under ISTEE. These results are further discussed in the fourth section, where we analyze the outcomes of the project and discuss some of the lessons learned. The article is then concluded in the fifth section.

THE PROJECT

Technical standards education is challenging on two levels. They are based on STEM topics, theories, and applications. Therefore, their education embodies the same challenges facing STEM education in general [3–8]. Meanwhile, standards education materials may not always be taught like those of conventional STEM courses. Innovative nontraditional instruction strategies are necessary to engage students, get them excited, and advance standards education. Methods where students undertake research activities and interact with each other have the potential to be particularly effective in understanding standards and standardization processes [1, 2].

Project ISTEE produces a novel 3-credit-hour university course, where innovation is twofold: in course structure and content, and in instruction strategy. The course title is *Telecommunication Standards and Standardization Processes*, and it has been offered twice so far (spring 2016 and fall 2016) at Jackson State University, Mississippi, United States. It is designed for senior undergraduate and graduate students. The course comprises two interleaved components, a conventional (lecture type) component and an interactive research-based component of two workshops. The first component is mostly composed of expository

This work is sponsored in part by Federal Award Number 70NANB15H342 of the U.S. National Institute of Standards and Technology (NIST). Opinions expressed in this article are those of its authors only

[1] Several organizations and efforts address the issue of integrating standards into STEM education. For example, the U.S. DoC NIST's Standards Services Curricula Development (SSCD) program, which funds project ISTEE, focuses on advancing standards education in U.S. STEM programs; [10] is an example of similar activities in the European Union; and the International Cooperation for Education about Standardization (ICES), http://www.standards-education.org/, has been active in promoting standards education since 2006.

Digital Object Identifier: 10.1109/MCOM.2017.1601236

Tarek S. El-Bawab is with Jackson State University; Frank Effenberger is with Futurewei Technologies, Inc.

material to help students understand the background, and acquire the vocabulary, of telecom standards. Conventional classes build toward the workshops (the second component) with preparations and research assignments. Every workshop is focused on particular standards project(s). A workshop is structured into two parts. The first part is a one-hour seminar by a standards' expert. The second is a two-hour standards meeting simulation where teams of students assume the various roles played by numerous parties in the standardization process, make contributions concerning the project(s) they have researched, debate these contributions, and appreciate how standards are developed and the role they play in today's engineering, economy, and business [2].

In our model, each of the simulation workshops may be based on a specific standards development organization's (SDO's) rules/culture. Simulations are designed around simplified technical problems in order to strike a balance between basing them on technical grounds, making them telecom-centric while open to other undergraduate majors such as electrical engineering and computer engineering/science, and leaving adequate room for students to learn about nontechnical issues. Examples of topics that meetings were simulated around include Internet-based home appliances, enhancing the performance of the tin-cup telephone, networked cars, wrist-band health monitors, and networks for manufacturing plants.

The new course is designed to give students a practical perspective of telecom standards and standardization processes. The syllabus also involves discussions of topics at the crossroads of engineering, technology, economy, market, politics, and human behavior. Examples of these topics included the competition between VHS and Betamax video tapes, the rise and fall of asynchronous transfer mode (ATM), the digital subscriber line (DSL) line-code war, the quest to deploy wavelength division multiplexing (WDM) passive optical networks (PONs), and others.

Students are graded mainly based on their performance in workshops. Other grading instruments, such as creative writing (papers, standards contributions, drafts, etc.), presentations, and quizzes, have been utilized as well. These other instruments are geared toward coaching students for the workshops and enhancing their learning experience from them.

Project ISTEE attempts to help students enter the workforce or proceed to graduate study with strong understanding of the value and benefits of standards and standardization. Further details about the course we produced in this project are described in [2].

Evaluation Results

Preliminary evaluation results were reported in [2]. These were based on an after-class survey and on the first course offering in spring 2016. Some of the improvements introduced in fall 2016 were based on the data collected and results obtained during the spring evaluation process.

We have been measuring the impact of the new course on students' learning and on their appreciation of the role standards play in engineering and economy. Two types of metrics are involved in our evaluation plan. Type 1 is survey-based numerical metrics where students grade several aspects of the course along with potential avenues to improve it. This type of measurement helps us understand students' perception of the new course, and explore methods to further stimulate their interest and enhance their learning from it. Type 2 measurements are based on ABET-type metrics to assess certain student learning outcomes. Data of this second type were collected using a combination of questions added to the spring 2016 survey and new quizzes introduced in fall. The goal of the evaluation plan is to examine the impact of the project on students' education and to guide our way toward continuous improvement.

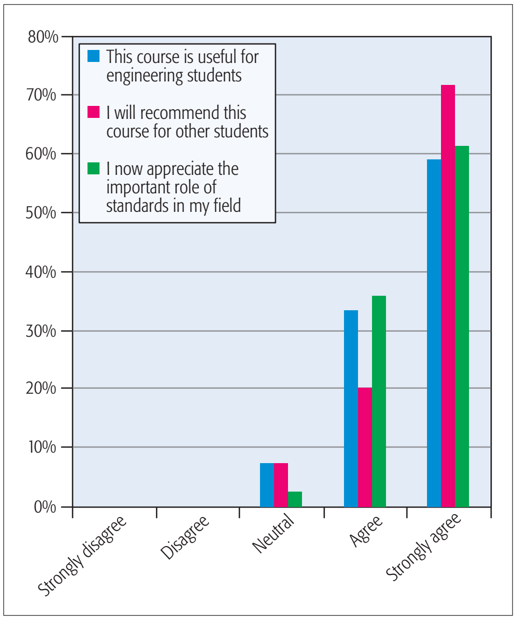


Figure 1. Students' perception of course return and usefulness.

Students' are graded mainly based on their performance in workshops. Other grading instruments, such as creative writing (papers, standards contributions, drafts, etc.), presentations, and quizzes, have been utilized as well. These other instruments are geared toward coaching students for the workshops and enhancing their learning experience from them.

The number of students enrolled in Telecommunication Standards and Standardization Processes in spring 2016 was 18. In fall 2016, 21 students took the class, bringing the total count of enrollments so far to 39 engineering students.

In this article, we report the results obtained up to the end of fall 2016, using the two types of measurements indicated above. The survey was enhanced by adding educational-outcome related questions to the spring version. Besides their role as a grading instrument, the quizzes were articulated in such a way as to complement survey questions and to test the effectiveness of the course in meeting certain program educational objectives. The results reported here are cumulative in their coverage of the inputs and performances of 39 students over two semesters. They are also expanded in terms of their consideration of some metrics we were not able to measure earlier with sufficient credibility.

Figure 1 depicts students' general perception of the impact of the new course on their education, how it fits within their program of study, and whether they better appreciate the role standards play in their discipline after taking this course. The trends observed in our earlier measurements continue. Only 8% of the students are neutral in terms of their recognition of the course as a beneficial addition to their program of study, and in

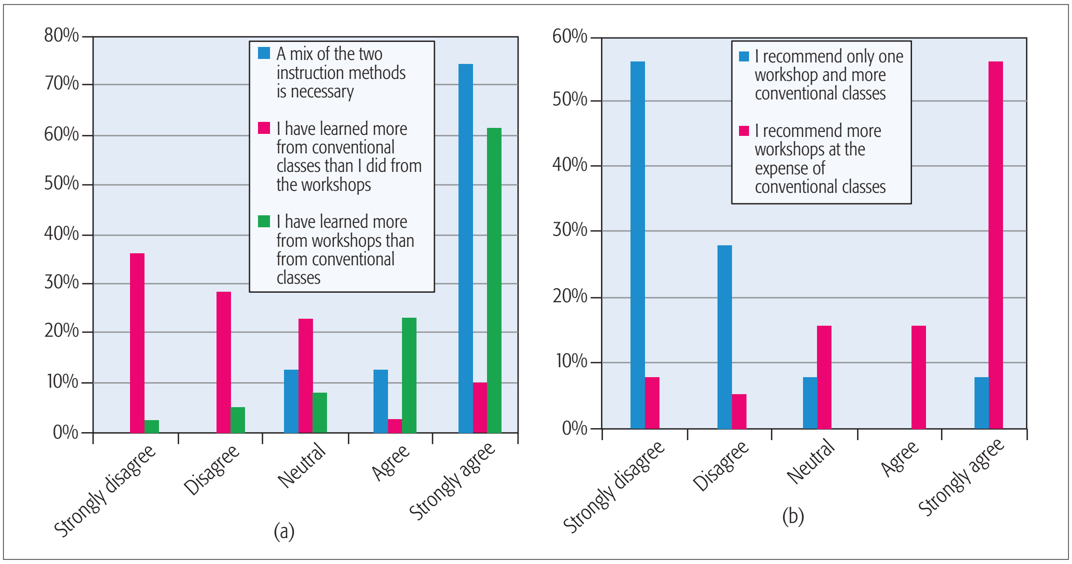


Figure 2. Students' perception of the two instructional components of the course: a) students' response to the two instruction methods; b) students' recommendation to have more/less workshops.

Course part	Fall 2016 topic	Spring 2016 topic
Part I	Introduction to Standards and Standardization: A Telecom Perspective	Introduction to Standards and Standardization: A Telecom Perspective
Parts II and/or III	Part II: International, Regional, and National Standards Organizations	Part III: Telecom Standards Development Organizations (SDOs)
	Part III: Academic and Industrial Standards Organizations	
Part III		Standardization Processes
Part IV	The Standardization Process	Standards Relation to Technology Life Cycle (TLC)
Part V	Standards in Relation to Technology Life Cycle	Standards and Telecom Market Forces
Part VI	Telecom Market Forces	Miscellaneous Topics

Table 1. Course parts.

terms of their intention to recommend it to other students. Only 3% are not sure they appreciate, or do not appreciate, the role of standards in their fields. None of the students described the course as not useful or suggested they have not acquired some knowledge of the role of standards.

Figure 2 focuses on students' perception of the two instruction methods involved in the new course. 87% of the students agree that a mix of conventional instruction and interactive research-based workshops is necessary for these kinds of university courses.

Interactive workshops enable students to apply knowledge and skills they acquire in class into simulations based on simplified case studies. Three types of standards meetings have been simulated: based on an International Telecommunication Union Telecommunication Standardization Sector (ITU-T) model, based on an Institute of Electrical and Electronics Engineers Standards Association (IEEE-SA) model, and based on a generic (not SDO-specific) countries' model. In these simulations, students experience critical thinking and team work. They get an opportunity to appreciate how standards projects are initiated, debated, and negotiated; how real-life standards are developed, and what impact they can have on engineering, market, and business. Out of the 39 enrolled students, about 85% believe that they learned more from workshops than they did from conventional classes. About 13% believe the other way around. Approximately 71% of the students want to see more workshops in this course at the expense of conventional classes, whereas 8% prefer to have only one workshop instead of two. Based on our experience, however, it can be difficult to organize more than two or three workshops of this type in a semester, especially if industrial guest experts are to participate in their design and administration.

Along with Table 1, Fig. 3 illustrates how students graded the topics of the course *Telecommunication Standards and Standardization Processes*. As part of our effort to improve the syllabus for fall 2016, the course was restructured, some part(s) were added, and other parts renamed, as described in the table. The figure shows how students graded all six topics/chapters over the two semesters on a scale ranging from 0 to 10, where 0 is the lowest grade and 10 is the highest. The horizontal bars depict the average grade given to each topic. The cumulated averages are strikingly close to preliminary results we reported earlier for spring 2016, and verify them. A new Part VI (Miscellaneous Topics), which was introduced in fall, received the highest score so far. This part came up as a fruitful addition to the course. It covers the relationship between standards and patents, and discusses competition in the arena of telecom standards.

Project ISTEE and standards education offer an opportunity to measure certain student educational outcomes typically sought in ABET's engineering self-studies.[2] Usually, these measurements are not particularly easy to perform. Examples of the outcomes we are able to measure using the new course include the students' ability to design within real-life constraints, to function on multi-disciplinary teams, to understand professional and ethical responsibili-

[2] Studies prepared by programs/institutions seeking ABET accreditation.

ties, to know about contemporary issues and their impact on engineering work, and to appreciate the need for life-long learning. Figure 4 gives the percentages of students who were able to demonstrate evidence of acquiring or improving these abilities through their enrollment in Telecommunication Standards and Standardization Processes.

The data in Fig. 4 were collected in fall 2016. As designed, the course is ideally suited to blending technical, economic, social, and political issues, and provides an excellent platform for students to see how these issues interplay in real life. This interplay and the standards resulting thereto, coupled with continuing progress in science, technology, and engineering, resulted in overwhelming recognition by students of the importance of lifelong learning, of acquiring multidisciplinary skills, and of designing within real-life constraints. The students, the vast majority of whom come from underserved minority backgrounds [3–8], have also demonstrated compelling evidence of their recognition of professional and ethical responsibilities and the impact of contemporary social, economic, and political issues on their engineering work. It is worth noting that the picture drawn herein is not of the percentage of students who acquired these five specific skills. Figure 4 merely reflects the percentages of students who were able to demonstrate evidence that they did so.

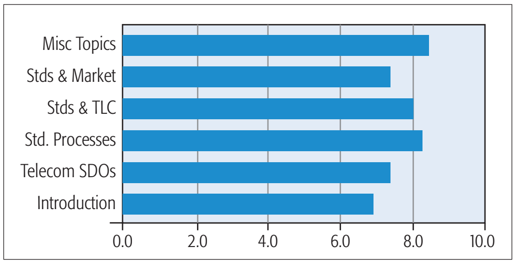


Figure 3. Students' reaction to course topics.

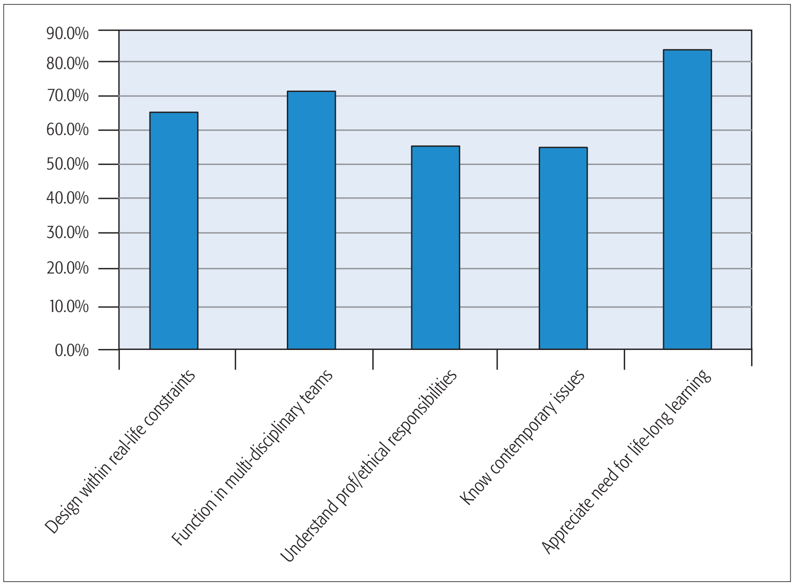


Figure 4. Assessment of course outcomes in terms of some ABET criteria.

Discussion

As project ISTEE is getting close to an end, the deliverables are fairly aligned with project goals and objectives. We have developed and delivered a first-of-its-kind syllabus and university-course material to help advance standards education in telecommunication engineering, and articulated a model of standards education for other STEM disciplines, departments, and universities to follow. Telecom is an area of engineering and technology where a case can be made for a specialized standards course. In other disciplines, standards education can be approached by integrating smaller, adequate, syllabic components into existing courses. In these cases, the ISTEE model provides a wealth of guidelines and examples to follow as well.

The new course introduces students to telecom standards and enables them to appreciate the combination of technical, regulatory, market, political, and other factors involved in standards and standardization. Telecom SDOs are classified and discussed. Various standardization processes and cultures are explored. The relationship between standards and technology life cycle (TLC) is examined; so is the impact of the telecom market forces on standards. The course also offers students a unique opportunity to learn about patents and patenting, their relation to standards, how industry may cooperate in standards while competing in research and product development.

Instruction methods are key to attracting students to topics of academic study, and standards are no exception. Results of project ISTEE demonstrate that a mix of conventional classes and research-based interactive classes is the most appropriate formula for standards education. Our experience proves that research-based instruction strategies are ultimately the most effective approach to standards education because they actively engage students and get them excited into real-life simulations. Our results are also in line with discipline-based education research (DBER), where similar findings have been observed in science and engineering education in general [11, 12]. Our findings, however, indicate that the impact of research-based strategies on standards education in particular is strong and can be transformational. Courses in economics, policies, and standards can be unpopular among some engineering students who are used to and prefer certain analytical and pedagogical approaches. Interactive research-based approaches to standards education can resonate with the engineering method to acquire knowledge and can overcome some of the challenges facing standards education in university programs.

DBER is an emerging interdisciplinary research enterprise that combines the expertise of scientists and engineers with methods and theories that explain learning. This enterprise investigates teaching and learning in a discipline from a perspective that reflects the discipline's priorities, views, knowledge, and practices. Informed by and complementary to research on learning and cognition, DBER generates insights that can be used to improve students' learning in science and engineering [11, 12]. We believe that these insights are particularly helpful in advancing technical standards education and agree with our findings in this regard. Research-based and interactive instruction strategies are more effective than traditional lecturing in improving knowledge acquisition and students' attitude toward standards learning. These strategies involve a range of approaches, including the use of real-life simulation exercises and making students work in groups, as is evident by the success of the workshops we designed and implemented. Other approaches include making lectures more interactive, and incorporating authentic problems and activities, which are methods we utilized as well. Several tools are developed today and can be utilized in this regard [11–14].

Project ISTEE has, of course, faced some difficulties, and other lessons can be learned thereto. The logistics and administrative tasks involved in launching a new course that incorporates two workshops where industrial speakers take organizational roles can consume considerable time and energy. These matters are budget-dependent. Therefore, institutional support (by universities) of a project like ISTEE, and of standards education in principle, is critical. Release time for faculty member(s) responsible for putting these workshops together and undertaking project activities is essential for the timely implementation and success of this kind of project. Online tools that have been developed in recent years can help overcome some of these problems [13, 14].

Standard education is necessary for future engineers who will practice their profession in a global space. More efforts are needed to communicate the need for curricular materials on the topic and to promote their incorporation in STEM programs everywhere. The instruction material of the course produced by project ISTEE will be published and made available to accredited university programs, industrial educators, and the general public. Progress in the field of telecom and telecom-related standards suggests that further projects are desirable. ISTEE is a catalyst project that paves the way for more efforts to advance and modernize standards education. Areas of future interest and focus include the evolving standards and standardization efforts in IoT, big data, and green ICT.

Progress in the field of telecom and telecom-related standards suggests that further projects are desirable. ISTEE is a catalyst project that paves the way for more efforts to advance and modernize standards education. Areas of future interest and focus include the evolving standards and standardization efforts in IoT, big data, and green ICT.

Conclusion

Science, technology, engineering, and mathematics (STEM) education is re-examined today, in terms of its objectives, curricula, majors, and instruction methods. Several changes are anticipated, including new majors such as Telecommunication Engineering and new instruction strategies. The goal of these efforts is to better prepare our workforce for the global challenges of the 21st century. As part of these trends, efforts are underway to advance technical standards education and integrate it into STEM programs. Project ISTEE is a project to advance technical standards education in U.S. university programs. The project has developed a first-of-its-kind university course in telecom standards and standardization processes, and presents a model for other STEM disciplines, programs, departments, and universities to follow. We have discussed the motivation and rationale of the project, described the tasks involved therein, discussed the results obtained so far, and explored future possibilities to expand this work and further advance its cause and goals.

Acknowledgments

We would like to acknowledge the support of the National Institute of Standards and Technology (NIST) throughout the duration of this work. In particular, we are grateful for the help and encouragement of our Program Officer Patricia Harris, and Eric Puskar. We thank Jeffrey Strauss, the Associate Director of the Center for Technology & Innovation Management at Northwestern University, Illinois, United States, for providing the material used for designing one of the two fall workshops.

References

[1] Tarek S. El-Bawab (PI), "Advancing and Integrating Knowledge of Telecommunication Standards into STEM Education," Federal Award Number 70NANB15H342, 2015-NIST-SSCD-01, the U.S. DoC's NIST Standards Services Curricula Development (SSCD) Program, Sept. 2015–Feb. 2017.

[2] T. S. El-Bawab and F. Effenberger, "Project ISTEE: Integrating Standards into Telecommunication Engineering Education: The Quest to Advance Standards Education in STEM," *2016 IEEE Conf. Standards for Commun. and Net. (CSCN)*, 2016, pp. 123–38.

[3] Nat'l. Science Board, "Moving Forward to Improve Engineering Education," NSB-07-122, NSF, Nov. 19, 2007.

[4] "Rising Above the Gathering Storm, Revisited: Rapidly Approaching Category 5," Nat'l. Academy of Sciences, ISBN: 978-0-309-16097-1, 2010.

[5] "Report to the President, Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics," Executive Office of the President, President's Council of Advisors on Science and Technology, Feb. 2012.

[6] "Federal Science, Technology, Engineering, and Mathematics (STEM) Education 5-Year Strategic Plan," Report from the Committee on STEM Education National Science and Technology Council, Executive Office of the President of the United States, May 2013.

[7] "Investing in Science, Engineering, and Education for the Nation's Future," nsf14043, NSF, Strategic Plan for 2014–2018, Mar. 2014.

[8] C. L. Fry, Ed., *Achieving Systemic Change: A Sourcebook for Advancing and Funding Undergraduate STEM Education*, The Coalition for Reform of Undergraduate STEM Education, The Association of American Colleges and Universities, 2014.

[9] 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.

[10] S. A. Bøgh, Ed., *A World Built on Standards- A Textbook for Higher Education*, Danish Stds. Foundation 2015, DS Handbook 181, project no. M298297, ISBN: 978-87-7310-964-9.

[11] *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering*, Nat'l. Academies, © The Nati'l. Academy of Sciences, ISBN 978-0-309-25411-3, 2012.

[12] *Reaching Students: What Research Says About Effective Instruction in Undergraduate Science and Engineering*, Nat'l Academies, © The Nat'l. Academy of Sciences, ISBN 978-0-309-30043-8, 2014.

[13] Eric Puskar, "A Game About Standards," Pub. of NIST, Sept. 4, 2012.

[14] IEEE Stds. Univ., "Mars Space Colony: A Game of Standardization," http://www.standardsuniversity.org/game/, accessed Feb. 2017.

Biographies

Tarek El-Bawab [SM] (telbawab@ieee.org) led the Telecommunication Engineering Education movement, resulting in recognition of telecom engineering as a distinct ABET-accreditable discipline. He is the first recipient of the ComSoc Education Award. His other research interests include networking and performance analysis. Before Jackson State University, he was with Alcatel-Lucent (United States), Colorado State University, and the University of Essex, United Kingdom, and led large-scale international telecom projects. He is a member of Eta Kappa Nu; a ComSoc Distinguished Lecturer; the Editor of Springer's Series: Textbooks in Telecommunication Engineering; and has more than 70 publications. He is a member of the IEEE Educational Activities Board and ComSoc Educational Services Board. He has served IEEE/ComSoc in numerous other technical/leadership capacities.

Frank Effenberger [F] (frank.effenberger@huawei.com) has worked in the optical access field at Bellcore, Quantum Bridge Communications (Motorola), and Futurewei Technologies, where he is now the vice president of the fixed access network lab. His team works on forward-looking fiber, copper, and coax access technologies, with several "world's first" prototypes and trials. He is a Fellow of the OSA and Huawei, and holds over 80 U.S. patents.

Learning Mobile Communications Standards through Flexible Software Defined Radio Base Stations

Víctor P. Gil Jiménez, Alejandro Lancho Serrano, Borja Genovés Guzmán, and Ana García Armada

The authors describe a new learning approach based on a new flexible hardware/software platform (FRAMED-SOFT), which is also detailed. Although the authors focus on two wireless standards, GSM and UMTS, the work discussed in this article can easily be extended to other standards of interest, such as LTE and beyond, WiFi, and WiMAX.

This work has been partially funded by the Spanish MECD FPU fellowship program granted to the second and third authors, and the Spanish national projects 'GRE3N-SYST' (id. TEC2011-29006-C03-03), 'ELISA' (id. TEC2014-59255-C3-3-R) and 'OTOSiS' (id. TEC2013-41718-R), and the Comunidad de Madrid (project 'CASI-CAM-CM,' id. S2013/ICE-2845).

Digital Object Identifier: 10.1109/MCOM.2017.1601219

Abstract

Mobile communications are today widespread and contribute to the development of our society. Every day new devices include some means of wireless transmission, which is becoming ubiquitous with the Internet of Things. These systems are standardized by international organizations such as the IEEE, 3GPP, and ETSI, among others. Even though knowledge of wireless standards is key to the understanding of these systems, wireless communications are quite often taught in engineering degrees in a traditional way, without much emphasis on the standardization. Moreover, strong focus is often placed on the theoretical performance analysis rather than on practical implementation aspects. In contrast, most of the current applications make extensive use of mobile data, and the global users' satisfaction is highly correlated with the mobile data throughput. Thus, modern wireless engineers need to have deep insight on the standards that define the mobile transmission systems, and this knowledge is not acquired following the traditional theoretical teaching schemes. In this article, a new learning approach is described. This novel paradigm is based on a new flexible hardware/software platform (FRAMED-SOFT), which is also detailed. Although the article is focused on two wireless standards, GSM and UMTS, the work discussed in this article can easily be extended to other standards of interest, such as LTE and beyond, WiFi, and WiMAX.

Introduction

Mobile communications are nowadays deployed worldwide, and their applications are continually growing. In fact, our economy and everyday life are conceived today around mobile communications. The Internet of Things (IoT) and the new generation of smartphones, which include access to social networks, are just two examples of how ubiquitous communications have become, and this trend is foreseen to continue in the future.

However, a reliable but efficient and universal wireless technology is not easy to design, deploy, and manage. In this heterogeneous scenario, many different wireless systems will cooperate and interact. Indeed, the mobile engineers of the future will have to start from the traditional and current communications standards and evolve them toward a new generation. For this purpose, they need to have deep knowledge and insight on wireless standards and their practical implementation. Specifically, in this article we focus on two well established wireless standards, Global System for Mobile Communications (GSM) and Universal Mobile Telecommunications System (UMTS). Some of the work that we discuss can easily be extended to Long Term Evolution (LTE), and even to other wireless systems such as WiFi and WiMAX.

According to our experience, it is highly demanded that new communication engineers have good experience in practical aspects of wireless communications, besides a very strong theoretical knowledge of these technologies. This insight can only be acquired through the laboratory work and practical experience. In this sense, software defined radio (SDR) can be a good option for students to approach the real world in the industry, and prepare them for the professional requirements they will face after finishing their studies. This requires a new learning paradigm where flexible equipment is used by engineering students and instructors, including cooperative education, which ensures that the students will be able to perform team work.

Some previously published papers highlight the benefits of teaching practical communications with SDR platforms. In [1] the work developed in six U.S. universities is presented showing how they have integrated SDR in their curricula, most of them including a significant laboratory component. The authors of [2] focus on the practical challenges found in designing and implementing wireless communications that are overlooked when focusing only on theory. They show how their students have improved their skills when working with the Universal Software Radio Peripherals (USRPs) of National Instruments (NI). Reference [3] describes the development of a hands-on open courseware using very low-cost SDR devices such as the RTL-SDR, which can be used in combination with Matlab to enhance the teaching of the principles and applications of digital signal processing and communications theory. Reference [4] presents a summary and comparison of hardware and software alternatives by examining four case studies of application of SDR in industry, academia, and government. All these works show experiences of using SDR in

The authors are with University Carlos III of Madrid. Alejandro Lancho Serrano is also with Gregorio Marañón Health Research Institute.

teaching focusing on the principles and fundamentals of communications. Also, recently, some commercial communications design frameworks are starting to include some libraries that cover standards such as GSM and LTE. However, their focus is on commercial system development and not on learning. Therefore, many important practical aspects are already solved, and they are not illustrative for learning purposes.

Complementary to these approaches, in this article we present our experience, where we have prepared and exercised with students some practical work based on well-known wireless communications standards. After discussing the available platforms and recent work, we explain our proposal: a novel SDR-based platform denoted as Flexible Radio Access Mobile Environment Defined by Software (FRAMED-SOFT). We provide a detailed description of the hardware, the developed software, and the contents of the practical work performed by the students. The article finishes with some conclusions about the learning methodology and the students' perception of the experience.

Hardware/Software Platforms and Recent Work

The use of SDR is rapidly expanding, and these devices have become a very useful tool to prototype new algorithms and evaluate them in a real-world environment [4]. The following SDR platforms are in widespread use: Ettus USRP N200/N210, ZedBoard with Xilinx Zynq-7000 FPGA and AD-FMCOMMS5-EBZ, NooElec NESDR Mini SDR USB Stick, and Ettus USRP E300. A detailed comparison of them can be found in [4]. Besides these platforms, there is a programming system called RF Network-On-Chip (RFNoC) developed at Ettus Research to ease large SDR designs implemented in field programming gate arrays (FPGAs).

NI offers the NI USRP-2920, which is functionally equivalent to Ettus USRP. It is equipped with two antenna ports. The digital-to-analog converters (DACs) and analog-to-digital converters (ADCs) are 400 Msamples/s with 16 bits/sample and 20 MHz of bandwidth. The operational frequency ranges from 50 MHz up to 2.2 GHz. These values enable the implementation of transmitters and/or receivers with enough quality and fidelity for many of today's communications standards.

The USRP needs to be handled with the help of the Laboratory Virtual Instrumentation Engineering Workbench (LabVIEW) software, which integrates all the tools that engineers and scientists need for the implementation of a wide range of applications. It is a graphical development software that easily allows the description and implementation of a complete system.

Within the LabVIEW interface, the NI LabVIEW MathScript RT allows the use of Matlab code directly in LabVIEW designs. This is especially useful when students are not very familiar with LabVIEW, since they are generally familiar with Matlab.

Alternatively, open source software tools for SDR prototyping are currently also used because they potentiate quick development, leveraging a high community of users. They include OSSIE [5], CubicSDR, ALOE [6], and the GNU Radio project [7]. However, due to our large experience with NI LabVIEW, their extensive range of products, and their comprehensive and timely support, we selected LabVIEW to develop our platform.

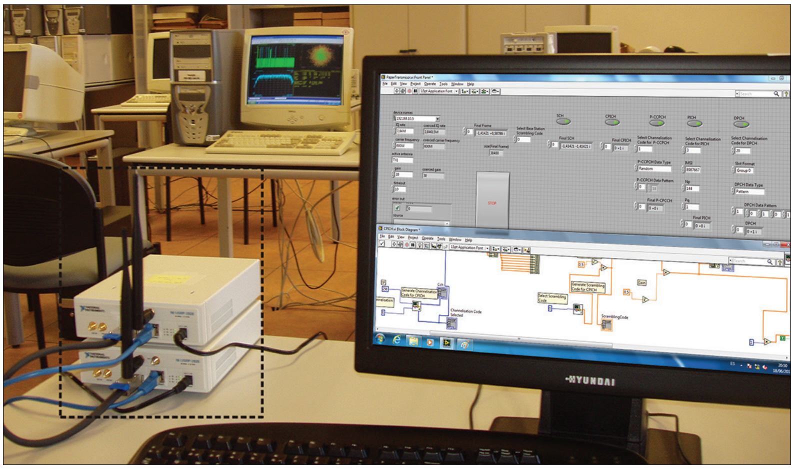


Figure 1. The developed BTS including the hardware part with NI USRP (dashed box) and software part composed of the computer and the LabVIEW design.

Using these types of SDR hardware and software, several universities and some research institutions are including practical teaching for their students to better understand the communications concepts.

• Instructors at the University of Strathclyde propose in [3] open courseware for SDR using the cost-effective RTL-SDR. With this receiver they can scan the RF spectrum and digitize I/Q signals within the range 25 MHz to 1.75 GHz. The software used is Matlab and Simulink.

• El-Hajjar *et al.* in [2] propose a practical session using USRP and LabVIEW, where the students must build a differential quadrature phase shift keying (DQPSK) modulation receiver. Besides, they present additional work that consists of building a PSK using pilot symbol assisted modulation (PSAM).

• Prof. Robert Heath uses NI USRPs and LabVIEW to teach digital wireless communication concepts at the University of Texas at Austin.

• Prof. Joseph Camp designed a course based on FPGAs and the FPGA-based hardware platform Wireless Open-Access Research Platform (WARP), for the School of Engineering at Southern Methodist University.

• Reference [8] shows the procedure to provide some wireless communication knowledge to engineering students with backgrounds in other areas. It is based on USRP and LabVIEW. In fact, they explain that they are replacing the WARP boards by NI USRP due to their enormous development in the last few years.

All these constitute successful experiences of teaching the principles of wireless communications in a practical way. They present the main building blocks of a generic wireless system, without particular focus on a given standard. Against this background, the proposed FRAMED-SOFT offers the possibility of working with real communication standards and real base stations, as shown later. This is indeed the innovation of our

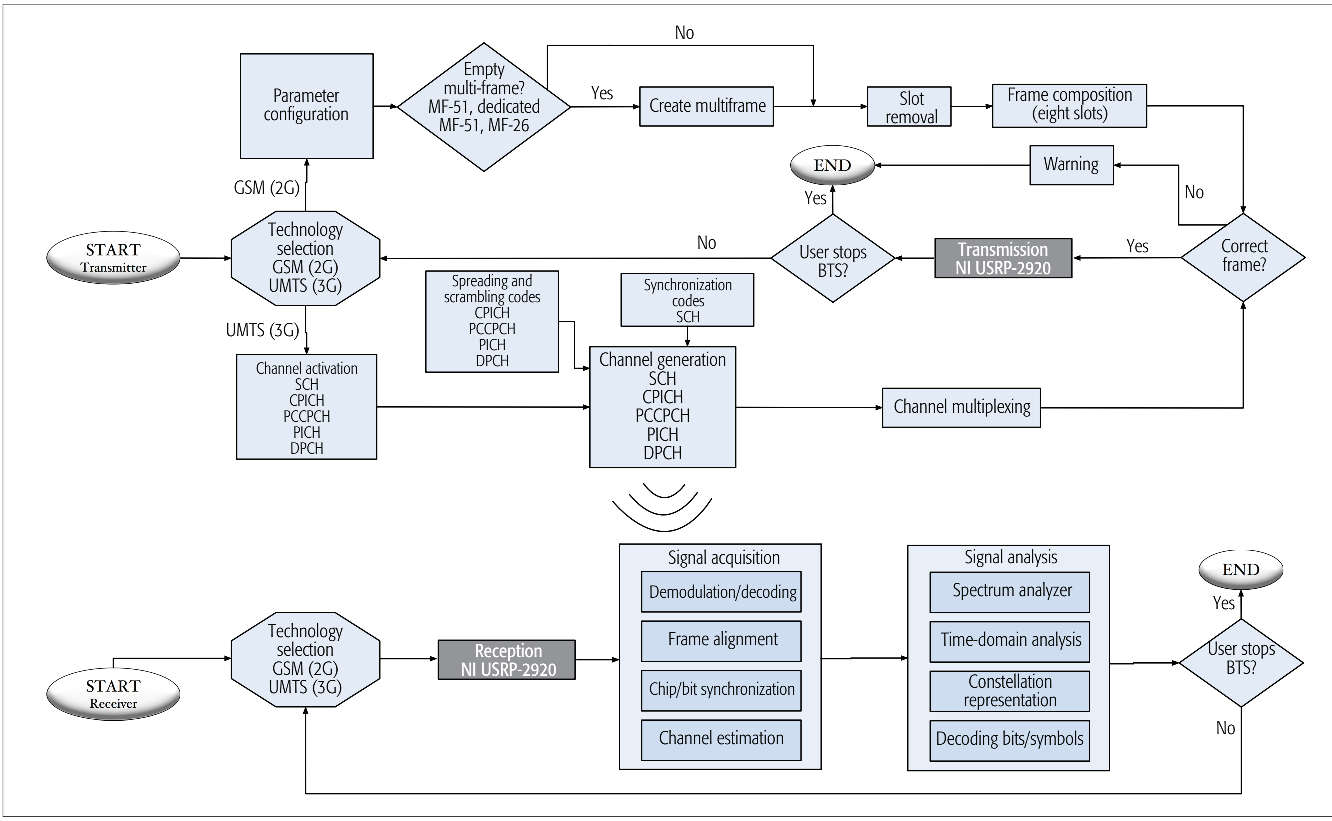


Figure 2. Simplified FRAME-SOFT diagram for the GSM and UMTS laboratory.

approach: the focus on learning the telecommunication standards by experience. Our platform integrates a complete communications chain, including the bit generation, framing, modulation, coding, and transmission of two representative examples of current mobile communications standards, GSM and UMTS. It is a flexible and modular platform that promotes cooperative learning among students and is scalable to include additional building blocks of the system. Moreover, the student contributions are included in the platform to be used in following years, promoting continuous improvement of the platform.

Working with real standards prepares the students for industrial experience. It also motivates them because they can practice with devices that they use in their everyday lives. The course is oriented to undergraduate students. It shows them some of the recent improvements in the evolution of the wireless communication systems, encouraging them to explore their own ideas. FRAMED-SOFT uses the NI USRP-2920 and LabVIEW as the hardware and software components.

The article is focused on learning the portion of the GSM and UMTS standards related to the physical layer. However, higher-level procedures established by the standards could also be envisaged to be dealt with in our platform.

Concepts Learned with FRAMED-SOFT

As mentioned before, engineers need to have deep insight on telecommunications standards because they will probably need to implement, improve, and manage systems and services based on them. However, the standards documents are not easy to read and/or understand. More specifically, the portion of the standards related to the physical layer is hard to learn, not only because of the complexity of the specifications, but also because there are many external and internal factors involved and multiple disciplines required. Interference, handovers, coexistence among standards and technologies, signal processing, communication theory, and hardware design are key ingredients to be considered.

Practical knowledge of the standards is often not included in the curricula of engineering. In fact, the traditional way of teaching the telecommunications standards relies on describing, with higher or lower detail, how the frame structure is formed and the modulation used for transmitting the signal. Besides, the channel coding and other procedures such as authentication, ranging, paging, and resource allocation are also theoretically explained. All these explanations are necessary. However, often the knowledge is neither assimilated nor remembered by the students because of the high complexity and abstraction level of the topics.

When there is a chance to illustrate telecommunications standards with practical classes or laboratory sessions, showing how the signal is built and transmitted, the learning of these concepts is highly improved. Indeed, based on our experience, if these concepts are supported by current standards, students pay more attention. In order to consolidate this knowledge, realistic practical experience needs to be offered to future engineers. Moreover,the industry is demanding engineers with these practical skills, good knowledge of standards, and understanding of implementation issues [8].

In our practical work, students become familiar with the specifications of the standards and learn how to browse throughout them. They are able to find and understand the parts of the documentation describing what they want to implement. Also, the students understand the importance of standardization in telecommunications experiencing two practical situations: they can implement a standard-compliant receiver that can be used with a working system, and, because of the interoperability allowed by standards, they realize it is straightforward to integrate different parts of a system performed by different students.

The FRAME-SOFT platform allows understanding of the main concepts related to implementation according to the standards of wireless communications systems as follows.

Transmission and reception: This includes all aspects related to, among others, the concepts of pulse shaping, sampling, interpolation, bandwidth, amplifier design, and automatic gain controllers. Working close to a real implementation, students need to deal with these concepts and design algorithms to carry out the required tasks. They can check what happens when changing their parameters and gain more insight into real physical aspects related to transmission.

Multiplexing and multiple access: In the design of multiplexing/demultiplexing, students deal with the frame structure, frame detection, and alignment and slot duration. According to our experience, this part is usually very time-consuming for students due to the variety of schemes and algorithms that are involved.

Modulation and coding: This deals not only with the modulation scheme, but also with other signal processing procedures associated with transmission (e.g., spreading and scrambling in the case of UMTS). Block coding, turbo coding, and advanced coding schemes for the new mobile standards must be considered. When designing each of these blocks, they are aware of their complexity and the interactions among blocks, as well as memory issues, overflows, and so on.

Receiver algorithms: Synchronization and channel estimation algorithms need to be designed and implemented to align signals in time and compensate the effects of the propagation channel. Although there are several algorithms already implemented in the platform, they can always improve them or design new ones. Delay is another important parameter, and students should realize here how it can compromise performance and quality of experience.

Signaling: There are procedures in wireless mobile communications including channel request, channel assignment, authentication, control signaling transmission and reception, resource allocation/management, and so on that students will use and implement in the FRAME-SOFT platform.

Thus, after finishing our practices, the students are aware of the importance of standards and are better prepared to join the industry.

The FRAMED-SOFT Platform

In order to support the new concepts and pedagogy described above to learn the wireless standards and in particular the physical layer, new tools and platforms need to be developed. These architectures must be flexible to allow fast implementation of different standards but powerful enough to be able to implement either the base stations or the mobile terminals.

SDR platforms such as the USRP-2920 from NI [9] possess those characteristics and thus are perfect for our purpose. Two of these devices can be seen in Fig. 1, where the developed base station is shown (highlighted in a dashed box).

By using the above described USRPs, the FRAMED-SOFT platform has been designed and implemented (Fig. 2).

Although the same USRP could work simultaneously as transmitter and receiver (one antenna for transmission and the other one for reception), our design uses two different USRPs connected by the provided interconnection wire. In this way, the design is simpler, it allows the use of less powerful computers for controlling the USRP, and it is not highly demanding for each USRP. Thus, it allows us to modify or optimize the transmission/reception parts independently.

The software implementing the base station allows one to select the GSM or UMTS standard. Depending on which one is chosen, the adequate LabVIEW design is executed. It could also be extended to further standards in the near future.

Since both standards are completely different from the physical layer point of view, two different designs have been implemented. However, from the logical point of view, both of them have several similarities. For example, both have a collection of specific channels for synchronization and channel estimation purposes, or for transmitting information or control tasks.[1] All of these channels have been included and can be enabled or disabled. Since the platform is focused on the physical layer, the vocoder has not been implemented and is emulated by random bits. However, since the platform is modular and flexible, higher-layer procedures could also easily be included.

For the GSM standard [10], FRAMED-SOFT includes the formation of the different multiframes (MF-51, Dedicated MF-51, and MF-26) following the instructions of burst composition and its corresponding information. The following channels are developed: FCCH, SCH, BCCH, AGCH, PCH, TCH/FS, and SACCH/TF.

To build the main functionality of the UMTS base station, the following physical channels are implemented: DPCH, P-CCPCH, CPICH, PICH, and SCH. As UMTS is based on wideband code-division multiple acces (WCDMA), the concatenation of specific spreading (channelization) and scrambling codes is implemented for each physical channel according to [11, 12]. Additionally, the SCH is generated to contain the synchronization codes [11].

In Fig. 3, the front panel of FRAMED-SOFT is captured, where it is shown that in GSM, the multiplexing/demultiplexing is performed in the time domain (upper part of Fig. 3), while in UMTS it is carried out in the code domain. The different channels for UMTS that can be fully configured are shown in the lower part of Fig. 3.

Learning and Getting Excited about the Physical layer of Telecommunication Standards

This section explains how the FRAMED-SOFT platform is used for introducing a new paradigm in learning about telecommunication standards. It is a complement and reinforcement to traditional

All of these channels have been included and can be enabled or disabled. Since the platform is focused on the physical layer, the vocoder has not been implemented and it is emulated by random bits. However, since the platform is modular and flexible, higher layer procedures could also be easily included.

[1] Besides, there are many procedures, such as paging, resource allocation, synchronization, or handover, among others, that are very similar between technologies, and all of them can be implemented in FRAME-SOFT.

The variety of laboratory sessions is very rich. Some students can work with the transmitter while other students can deal with the receiver. They can choose between GSM or UMTS for their experiments. Moreover, once they have selected the element to work on, they can even focus more on which aspect to tackle.

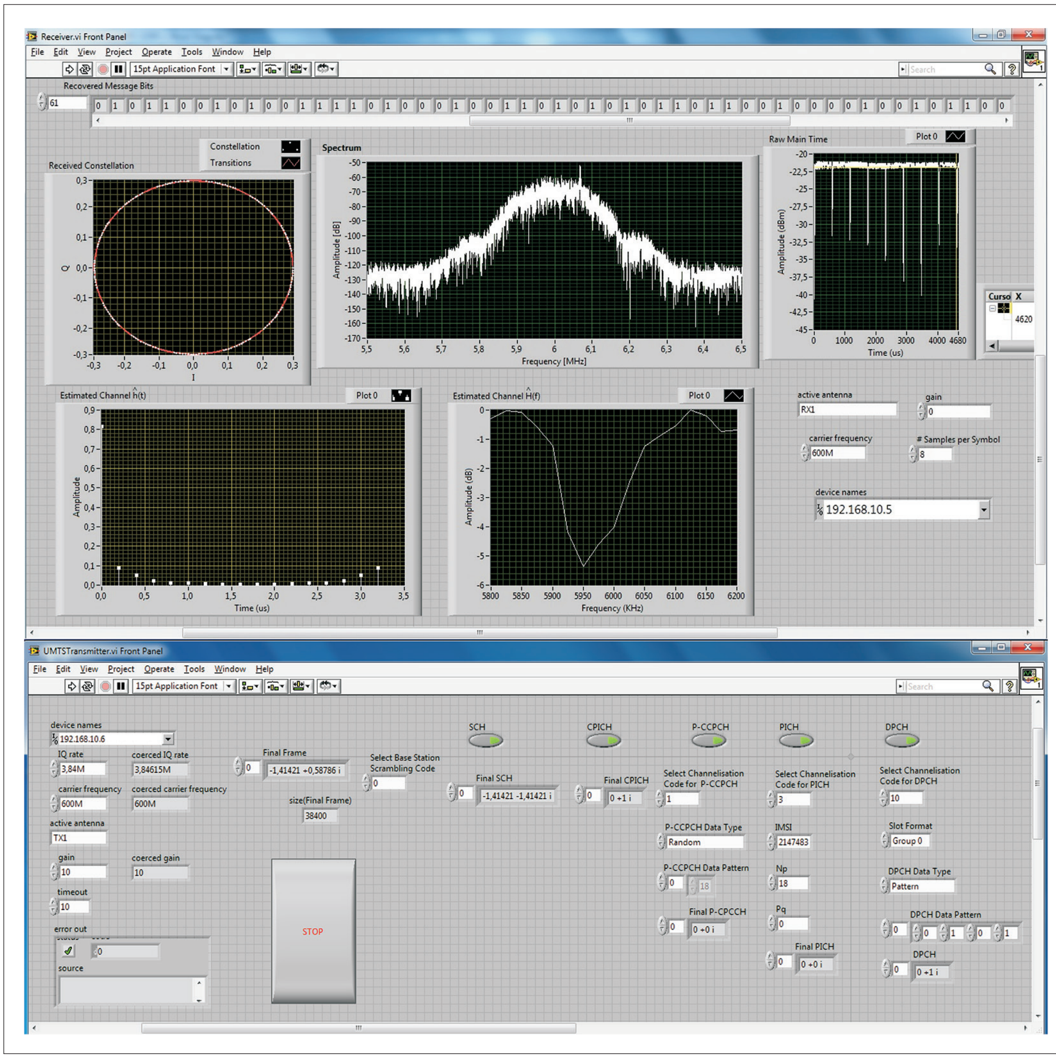


Figure 3. Two examples of the front panel for the FRAMED-SOFT tuned to 600 MHz. Upper part: GSM receiver showing constellation, frequency and time signal, channel estimation, and received decoded bits. Lower part: UMTS base station showing the main transmission parameters that can be configured.

teaching. The students practice with physical layer implementation, experience the concepts that they have previously seen in theoretical classes, and establish connections among concepts that were distant before.

The main idea is to involve the students as much as possible in their learning process. They are provided with knowledge and tools, and they are given some challenges to solve while acquiring realistic and practical insight. At the same time, we have seen that this methodology helps them enjoy while they learn.

Students involved in the course have full access to the complete LabVIEW version because of the existing agreements with National Instruments [13]. Thus, they can install the software at home. In this way, the students can implement their solutions at home and come to the laboratory to test their prototypes in the hardware. The platform also allows saving or loading signals from files. This allows, for example, the group working on the receiver to have access to a signal saved in electronic format at the transmitter side and be able to check their progress.

The variety of laboratory sessions is very rich. Some students can work with the transmitter while other students deal with the receiver. They can choose between GSM or UMTS for their experiments. Moreover, once they have selected the element to work on, they can focus even more on which aspect to tackle.

Cooperative Problem Solving

Students can work on a standalone problem. For example, one student can deal with the channel estimation and equalization at the UMTS base station, whereas another student can decide to improve the rake receiver. Both of them work separately at the beginning and cooperate at the end to obtain an improved equalizer. In addition, they can cooperate with another student who may have chosen to cope with synchronization, obtaining in that way a much better receiver. The platform is flexible and modular to give these opportunities.

Cooperative Challenge

If the instructor and the class agree, they can organize a competition to see who is able to better improve the system: the class is divided into groups of several students, and then one of the standards is chosen (GSM or UMTS). The different groups work on the diverse parts of the system

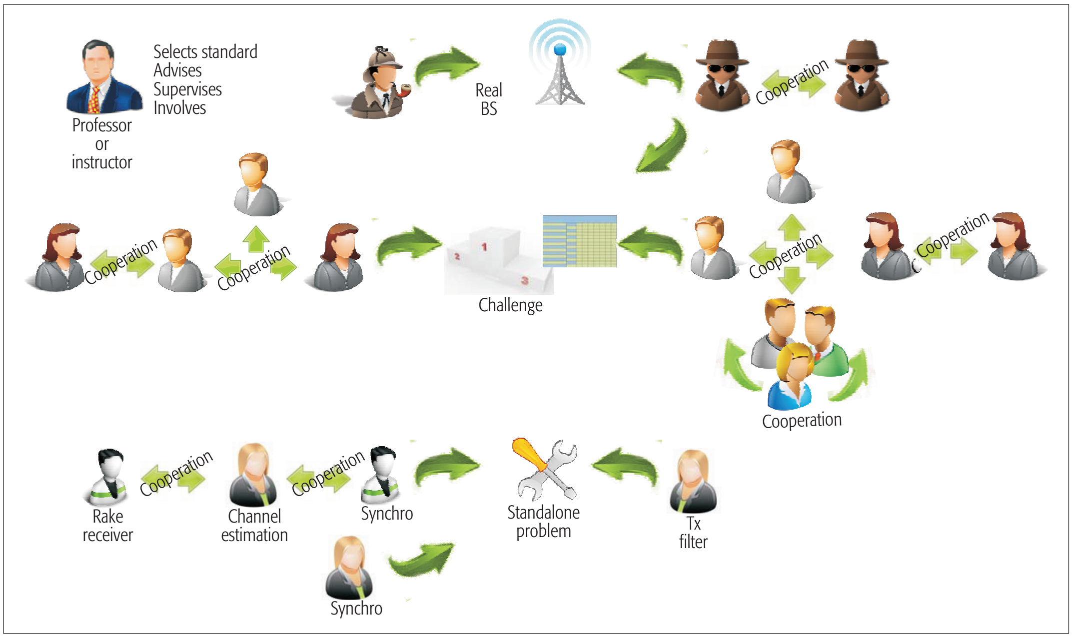


Figure 4. Examples of class configurations, roles, scenarios, and interactions among students.

> The platform offers the possibility of implementing part of a realistic receiver. The advantage of using designs based on real standards is that the student can implement a receiver to detect a signal that is transmitted by a real base station from a telecommunications operator. They can even extract additional knowledge from the signal.

in order to enhance them. They must cooperate with each other because at the end, the winning team will be the one that produces the whole system with the best performance, measured in terms of either error probability or throughput. This ranking will be considered for grades. The results from the current and previous years can be accessible to students. Thus, they can know how close they are to the previous or current teams.

Real Base Station

The platform offers the possibility of implementing part of a realistic receiver. The advantage of using designs based on real standards is that the student can implement a receiver to detect a signal that is transmitted by a real base station from a telecommunications operator. They can even extract additional knowledge from the signal. For example in UMTS, the scrambling code being used and the channelization codes can be found, or the information in the broadcast channels can be decoded. In GSM, they can obtain information about the organization of the cell and the cell size by obtaining the timing advance parameter from the broadcast channels, the number of neighboring cells, or even the standards being used, whether GSM, general packet radio service (GPRS), or enhanced data rate for global evolution (EDGE), as well as the exact localization by measuring the received power and signal quality at different points. The improvements at the receiver in channel estimation, synchronization, diversity, and detection will play an interesting role when working with these real signals.

Some examples of these configurations can be seen in Fig. 4 where, as explained before, all of them can be simultaneously given in the same class.

Organization and Scheduling of the Laboratory Work

In the following, data obtained during academic years 2013–2014 and 2014–2015 in the Mobile Communications course [14] within the degree in engineering in communications systems at the University Carlos III of Madrid, is briefly described. The laboratory contents are summarized as follows.

When focusing on GSM, the purpose of the session is the generation of several GSM control channels along with a traffic channel. The main aim is that the students learn about the whole procedure starting from the source information bits until the information is transmitted. Concretely, two control channels and one traffic channel have been examined in detail follows these steps:

- Generation of two control channels:
  - –Frequency correction channel (FCCH).
  - –Synchronization channel (SCH): to generate the encrypted bits (78 bits). Each group has different parameters.
- Generation of one TCH (traffic channel): to create the end information bits (114 bits).
- Creation of bursts: The idea is to generate a module that receives the bits obtained before and conforms the bursts by including a training sequence and some tail bits (148 bits).
- Integration of all the modules created previously and transmission of the generated signals with a USRP transceiver; afterward, reception of these signals with another USRP transceiver and analysis of the frequency spectrum, the time domain signal, and the constellation.
- Preparation of a written report with the explanation of all the steps followed, the achieved results, conclusions, and difficulties found.

When the focus is on UMTS, the goal is the generation of very specific UMTS control channels, including:

- Generation of the SCH by using LabVIEW, including the SCH codes. Synchronization data are different for each group.
- Generation of the common pilot channel (CPICH) including the spreading and the scrambling codes for the specific cell.
- Multiplexing of both channels and transmission by the USRPs.

The main outcome of the proposed methodology using FRAMED-SOFT is that the students get involved in their learning process and enjoy the physical layer of mobile communications based on standards. Indeed, the study of standards that is usually considered boring by students becomes exciting because they can understand and implement them.

- Signal reception by an additional USRP and analysis of the quality of the received signal.
- Preparation of a written report with the explanation of all the steps followed, the obtained results, conclusions, and difficulties found.

To speed up the learning curve of LabVIEW and USRPs for the students, a tutorial has been elaborated and an introductory laboratory training session has been scheduled. As [4] states, an open-ended design experience improves the students' learning. Thus, several challenges are proposed such as the creation of a synchronizer, equalizer, and channel estimator, and the extraction of base station information. The students, organized in groups, design and implement a part of the communication system from scratch during several laboratory sessions. Detailed material is available in [14].

Project Outcomes

After the first experiences with the new teaching methodology and the developed platform, several important lessons have been learned. The most important one is that this new practical work requires more effort from both sides, students and teachers. From the students, because they have not been exposed to LabVIEW before, they are not familiar with the tool, and the way of describing algorithms and procedures in a real prototype is new to them. From the teachers, more effort is required because they also have to get used to the tool and because of the wide range of questions and troubleshooting that are often encountered when dealing with real systems. Besides, we have to consider the amount of work devoted to the design of the practical work in the first year.

According to the students, who have answered a survey, their experience was very positive (95 percent satisfaction). They agreed that the laboratory sessions with LabVIEW and the USRPs allowed them to learn both a completely different programming language and how to use new hardware devices, which together allowed them to understand better the real problems associated with the implementation of communications standards. Now they are ready to tackle the development of standard-compliant communications systems, and also, they are aware of the importance of standardization. We would like to highlight that most of the students considered this practical work more difficult than the traditional work based on simulations with Matlab (80 percent). They evaluated as very important the initial tutorial taught at the beginning of the course (100 percent), although the practical work implied more workload compared to the traditional methods.

From the instructors' point of view, the development of the new practical work entailed significant effort. From now on, this effort will be lower as only some continuous improvement and updating of the practical work will be needed.

Another important and interesting outcome is that after the introduction of this practical work into the Mobile Communications course, many students have decided to carry out their Bachelor's thesis by using this platform. This is of great interest because their results will help to improve the existing practical work and develop new laboratory sessions. In fact, a simple LTE version is being developed.

Conclusions

In this article the architecture and functionalities of a new flexible platform, called FRAMED-SOFT, have been explained. By using this platform, new concepts and learning methodology have been explored, giving as a result an improved way of learning telecommunication standards. We have focused on two of the main cellular standards nowadays. Besides, these possibilities can be extended to also cover other wireless systems that are currently being deployed. The main outcome of the proposed methodology using FRAMED-SOFT is that the students get involved in their learning process and enjoy the physical layer of mobile communications based on standards. Indeed, the study of standards that is usually found boring by students becomes exciting because they can understand and implement them. The fact that they can work in realistic scenarios adds value to their learning. Moreover, several students have decided to carry out their Bachelor's and Master's theses in the field of wireless communications specifically using and improving our platform. Other students had interesting ideas on how to improve the standards. With FRAMED-SOFT and the practical work carried out, students learned about some of the most important aspects in these standards, such as transmission and reception, multiplexing, modulation and coding, synchronization, and channel estimation. They have done so using current standards, which has allowed them to experience their implementations with real-world signals. Any instructor may use our platform upon request as explained on our website [14].

References

[1] S. G. Bilèn *et al.*, "Software-Defined Radio: A New Paradigm for Integrated Curriculum Delivery," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 184–93.
[2] M. El-Hajjar *et al.*, "Demonstrating the Practical Challenges of Wireless Communications Using USRP," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 194–201.
[3] R. W. Stewart *et al.*, "A Low-Cost Desktop Software Defined Radio Design Environment Using MAtlab, SImulink, and the RTL-SDR," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 64–71.
[4] A. M. Wyglinski *et al.*, "Revolutionizing Software Defined Radio: Case Studies in Hardware, Software, and Education," *IEEE Commun. Mag.*, vol. 54, no. 1, Jan. 2016, pp. 68–75.
[5] M. R. *et al.*, "OSSIE: Open Source SCA for Researchers," *Proc. SDR Forum Tech. Conf.*, vol. 47, 2004.
[6] I. Gomez, V. Marojevic, and A. Gelonch, "Aloe: An Open-Source SDR Execution Environment with Cognitive Computing Resource Management Capabilities," *IEEE Commun. Mag.*, vol. 49, no. 9, Sept. 2011, pp. 76–83.
[7] E. Blossom, "GNU Radio: Tools for Exploring the Radio Frequency Spectrum," *Linux Journal*, no. 122, 2004, p. 4.
[8] M. Petrova, A. Achtzehn, and P. Mähönen, "System-Oriented Communications Engineering Curriculum: Teaching Design Concepts with SDR Platforms," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 202–09.
[9] NI, "Universal Software Radio Peripheral (USRP)," http://www.ni.com/usrp/esa/, accessed February 7, 2017.
[10] ETSI, "Mobile Technologies: GSM," http://www.etsi.org/technologies-clusters/technologies/mobile/gsm, accessed February 6, 2017.
[11] 3GPP, "Spreading and Modulation (FDD)," TS 25.213, tech. rep. v. 7.7.0, Mar. 2007.
[12] —, "Physical Channels and Mapping of Transport Channels onto Physical Channels (FDD) (Release 5)," TS 25.211, tech. rep. v. 5.0.0, Mar. 2002.
[13] NI, "Student and Professor Program of National Instruments Spain," http://spain.ni.com/academia/programa-estudiantes-faq, accessed Feb. 7, 2017.

[14] GCOM, "Mobile Communications Course's Web Page," http://gcom.tsc.uc3m.es/en/teaching/mobile-communications/, accessed Feb. 6, 2017.

Biographies

Víctor P. Gil Jiménez [S'00, AM'02, M'03, SM'12] received his B.S. and M.S. degrees in telecommunication in 1998 and 2001 from the University of Alcalá and University Carlos III of Madrid, Spain, respectively, and his Ph.D. degree in 2005 from the University Carlos III of Madrid, all of them with Honors. He is with the Department of Signal Theory as an associate professor. He has also led several private and national Spanish projects, and has participated in several European and international projects. He holds one patent. He received the Master Thesis and Ph.D. Thesis Awards from the Professional Association of Telecommunication Engineers of Spain in 1998 and 2006, respectively. He has published over 50 journal/conference papers and 5 book chapters. His interests are in the field of advanced multicarrier systems for wireless radio and visible light communications.

Alejandro Lancho Serrano [S'14] received his B.Sc. and M.Sc. degrees in electrical engineering from the University Carlos III of Madrid in 2013 and 2014, respectively. He is currently a Ph.D. student in the Signal Processing Group at the University Carlos III of Madrid. He is a recipient of an FPU scholarship from the Spanish Ministerio de Educación, Cultura y Deporte. His current research focuses on finite blocklength information theory and wireless communications.

Borja Genovés Guzmán [S'14] received his B.Sc. and double M.Sc. degrees in electrical engineering from the University Carlos III of Madrid and the Institut Mines-Télécom, France, in 2013 and 2015, respectively. He is currently a Ph.D. candidate in the Communications Research Group at the University Carlos III of Madrid. He is a recipient of an FPU scholarship from the Spanish Ministry of Education, Culture and Sport. His current research focuses on new techniques to improve the efficiency of optical wireless communications systems.

Ana García Armada [S'96, A'98, M'00, SM'08]received her Ph.D. degree in electrical engineering from the Polytechnical University of Madrid in 1998. She is currently a professor at University Carlos III de Madrid, where she leads the Communications Research Group. She has participated in more than 60 national and international research projects and contracts with the industry. She has published around 150 papers in international journals and conference proceedings, and she holds 4 patents. She has contributed to international standards organizations and is a member of the expert group of the European 5G PPP. She has received a Young Researchers Excellence Award and an Award for Best Practices in Teaching, both from University Carlos III of Madrid. Her main interests are signal processing techniques applied to wireless communications.

Multidisciplinary Learning through Implementation of the DVB-S2 Standard

Yuri Murillo, Bertold Van den Bergh, Jona Beysens, Alexander Bertrand, Wim Dehaene, Panagiotis Patrinos, Tinne Tuytelaars, Ruth Vazquez Sabariego, Marian Verhelst, Patrick Wambacq, and Sofie Pollin

We give an overview of the objectives of the EAGLE project and the relevance of the DVB-S2 standard, the design and implementation task that the students need to accomplish, and how they finally learn to master the complexity of a real communication standard, VHDL coding, and system verification and integration.

Yuri Murillo, Berthold Van den Bergh, Jona Beysens, and Sofie Pollin are the primary authors of this article.

Digital Object Identifier: 10.1109/MCOM.2017.1700007

ABSTRACT

Telecommunication standards are documents that contain consolidated knowledge about communication systems and implementation best practices. They are created based on long consensus processes in order to meet practical constraints. This article describes how the DVB-S2 standard is used in the electrical engineering curricula at KU Leuven within a design and implementation course called EAGLE. The goal of the course is to teach third-year Bachelor students how to apply the abstract knowledge gathered in the theoretical courses (control theory, communication theory, software, and hardware) to the design of a complex and real engineering project. Based on their progress during the academic year, we illustrate how standards can be used to teach a broad skill set and act as an accelerator to abstract system implementation complexity. We give an overview of the objectives of the EAGLE project and the relevance of the DVB-S2 standard, the design and implementation task that the students need to accomplish, and how they finally learn to master the complexity of a real communication standard, VHDL coding, and system verification and integration.

INTRODUCTION

The training of electrical engineering students should be broad and multidisciplinary, providing strong theoretical and practical foundations. However, it is challenging to design an engineering curriculum or even a single course that forces students to simultaneously think at the system level, develop theoretical insight on the underlying topics, and master practical implementation. As standards are documents culminating years of research efforts and must be written in coherent and unambiguous terms, they can potentially be a unique means of teaching students how theoretical knowledge is translated into real-world applications. Moreover, it can help them identify industry leaders and understand the business aspects that rule standardization before entering the market once their education is completed.

Contrary to this idea, engineering education at the Bachelor level classically omits the study of standards as they are often considered to be too specific for one particular technology or to contain too many practical details. As a result, courses cover only abstract knowledge and understanding of common technologies, and application is postponed to the Master level. During the first years of their education, students obtain a skewed view of the engineering field, making them unaware of the real competencies needed for engineering, which are covered at the Master level [1, 2].

In this article we explain how we used the second generation of the digital video broadcasting over satellite (DVB-S2) standard [3] in the design of the EAGLE course, taught to third-year Bachelor students. This standard is used to implement a wireless video link for a drone within a total time budget of around 50 hours using a lightweight and widely used field programmable gate array (FPGA) board, the ZyBo [4]. The main task is the partial implementation of a DVB-S2 transmitter in the very high-speed integrated circuit (VHSIC) Hardware Description Language (VHDL). Aware of the importance of introducing project-based learning at the Bachelor level [5], we decided to avoid designing a completely made up project and exploited the benefits of using an already existing standard: students are able to understand its importance faster and become more motivated thanks to its connection to the real world [6].

We now summarize the main arguments in favor of our choice of standard.

- As the task that the students need to fulfill is the development of a wireless video link, it makes sense to select one of the main technologies used nowadays. DVB-S2 is currently the main satellite digital television broadcast standard, used worldwide to broadcast standard definition television (SDTV) and high definition television (HDTV) services. The first part of [7] gives a clear overview of the advantages of DVB-S2 over legacy systems. More advanced standards could have been used (e.g., the terrestrial extension known as DVB-T2), but they were too complex for the technical background of third-year Bachelor students.
- It covers content that is already studied or will be studied for the degree.
- Off-the-shelf components can be used for testing the implemented code, such as the ZyBo FPGA board for the transmission and a standard-compliant DVB-S2 receiver for the reception of the video signal. Additionally, as these components are relatively cheap, students are not intimidated by them and feel free to experiment without the fear of breaking an expensive device.
- While DVB-S2 is a common and widely used standard, no open source, complete VHDL implementation is available online. This means that students get the opportunity to play a central role

The authors are with KU Leuven

in the implementation of the standard, while also learning a new programming language in the process.

The structure of this article is as follows. First, more details of the EAGLE course are given. Next, we give in-depth information about the wireless video link task. We discuss the conclusions we can draw after running the course for one year. Finally, we conclude this article, where we give examples of other accents that could be given with the same material. The whole project gives the instructor a wide palette of possible study items on which to elaborate.

Top-Level Overview of the EAGLE Course

In the third year of the electrical engineering Bachelor program at KU Leuven, there is an important design and implementation course aiming to teach students how to bring this broad theory into practice, called EAGLE. We introduce this project and give an overview of the different modules to be implemented by the students, while we describe the wireless video link task related to the DVB-S2 standard. Finally, we elaborate on the different resources available to the students for completing the course and its evaluation criteria.

The EAGLE Mission

The EAGLE problem solving and design course carries a total study load of 270 hours per student, representing nine European Credit Transfer System (ECTS) credits out of the 60 envisioned for the academic year [8]. In the current EAGLE course, nine teams of ten students have to design a smart drone able to fly autonomously and execute a complex task: the EAGLE mission. Note that a different drone per team should be designed. In order to do so, a total of six modules need to be implemented:

1. Hardware design of the electronic control system of the engines
2. Vision processing to recognize the environment
3. Cryptography to secure the mission
4. Autonomous navigation and control
5. Wireless power transfer
6. Wireless video link for transmitting the video stream obtained from a camera to a central controller

This last module represents 50 out of the 270 working hours dedicated to the project.

The EAGLE mission is demonstrated during the final demo session of the course. A safety net ensures physical separation between the drone and the students, and a 5 × 4 grid of quick response (QR) codes is placed on the floor, as well as a landing platform. This mission takes place in an indoor scenario and is divided into three different phases, as can be seen in Fig. 1. An increasing number of modules is needed for successfully completing each phase, with all of them correctly performing to ensure that the mission is finally accomplished.

In the first phase students need to manually fly the drone toward the first QR code coordinate using a remote controller and based on the camera signal sent through the wireless video link to a receiver PC. The steering of the motors and

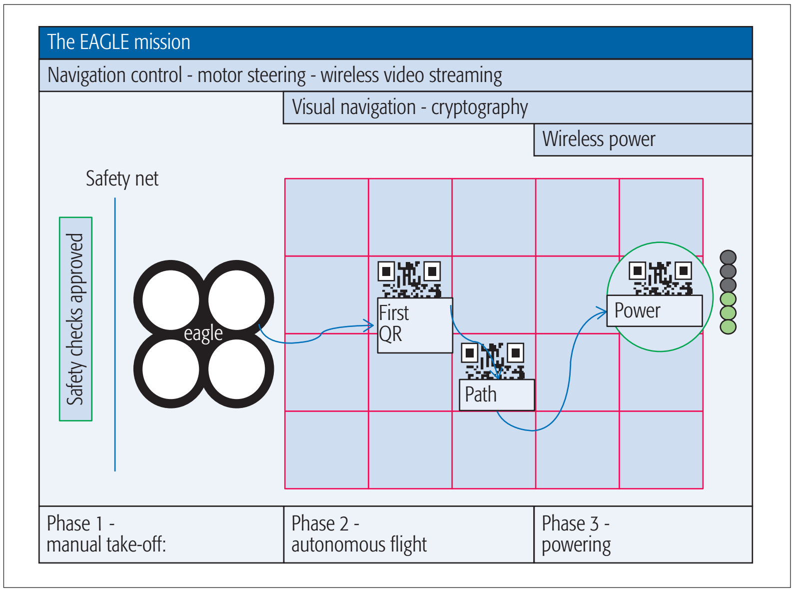


Figure 1. The EAGLE mission featuring the three described phases, where an increasing number of modules need to be operative to successfully perform each task.

the navigation controller should also be operative in this phase in order to ensure a stable flight. In the second phase the drone switches to autopilot mode and flies over the grid describing a predefined path. This path is obtained from the QR codes, which contain encrypted information about the next coordinate to which the drone should fly. For this phase the cryptography, autopilot, and vision processing modules need to be functional in order to detect the lines of the grid and decode the QR information. Finally, in the third phase the drone obtains the position of a landing platform surrounded by a copper coil that is connected to a power receiving circuit, able to power a light emitting diode (LED) meter. The drone then needs to autonomously land on the platform and inductively transfer power from its coil to the one underneath the platform. Based on the performance of the landing maneuver and the wireless power transfer module, the LED meter will display a higher or lower level, and finally the mission is accomplished.

The Wireless Video Link Module

The wireless video link module is the focus of this article, and covers about one-sixth of the course. During the first semester two to three students per team work on a partial VHDL implementation of a DVB-S2 standard-compliant transmitter. This code will be deployed on the ZyBo board installed in the drone. The video stream is transmitted to an off-the-shelf DVB-S2 receiver in order to obtain the signal being recorded by the camera.

The ZyBo development board was chosen for this project due to its low cost and small learning curve. This board has several I/O ports (USB, HDMI, Ethernet, VGA, audio, and GPIO), two ARM Cortex-A9 CPUs, and one FPGA. The camera functions like an IP surveillance capture device: H264 encoded packets are sent via Ethernet to the ZyBo at a rate of 1.5 Mb/s and encapsulated into a unidirectional lightweight

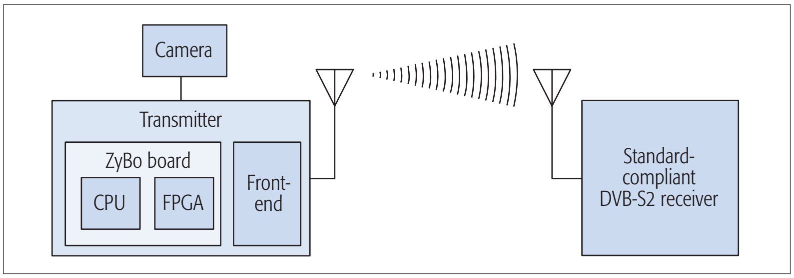


Figure 2. Architecture of the wireless video link, consisting of a ZyBo board with added camera and custom front-end for the drone and an off-the-shelf DVB-S2 receiver for the ground station. The FPGA on the ZyBo board runs the custom DVB-S2 implementation.

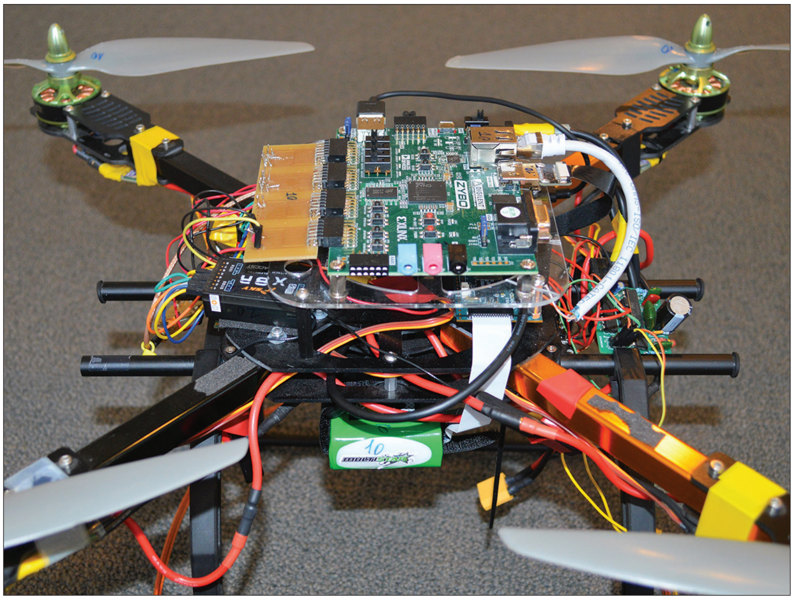


Figure 3. EAGLE drone with the ZyBo board on top of it. The camera is attached below, facing the ground in order to recognize the QR codes.

encapsulation (ULE) stream [9] which has an acceptable overhead [10]. This MPEG transport stream is then handed to the DVB-S2 transmitter core in the FPGA fabric. The IP core is able to process approximately one bit of the transport stream data per clock cycle. Since it is capable of running at up to 100 MHz, much higher throughputs are possible. The IP core supports all standard code modulations: quadrature phase shift keying (QPSK), 8-PSK, 16-amplitude and phase shift keying (APSK), and 32-APSK. All code rates are supported: 1/4 (short block only), 1/3, 2/5, 1/2, 3/5, 2/3, 3/4, 4/5, 5/6, 8/9, and 9/10. In the case of the EAGLE project, the output signal has a symbol rate of 866 ksamples/s. A root-raised cosine (RRC) filter with a roll-off factor of $\alpha = 0.35$ is used. This digital baseband stream is finally delivered to a custom designed front-end that transmits it in the 5.8 GHz industrial, scientific, and medical (ISM) band. For the EAGLE course, the constant coding and modulation (CCM) mode is used along with the regular single transport stream broadcasting service configuration. This architecture is depicted in Fig. 2.

Other low-cost boards for the Xilinx Zynq-7000 system on chip (SoC) can be used instead of the ZyBo, such as the ZedBoard [11]. The camera selected for this project is a regular Raspberry Pi camera [12], while the receiver chosen is the TBS5980 QBOX CI DVB-S2 TV tuner [[13], which is connected to a PC. An EAGLE drone showing this hardware is depicted in Fig. 3.

The wireless video link task is divided into two different phases: design and implementation. In the initial design phase, students need to become familiar with the DVB-S2 standard and understand the functionality of the transmitter sub-blocks. They receive a copy of the full DVB-S2 standard, as well as a description of a partial VHDL transmitter implementation. This is a version of a fully functional custom implementation of the DVB-S2 standard for the ZyBo FPGA,[1] with some blocks omitted that they will have to implement in the second phase. This approach reduces the complexity of the task and ensures that students are not overwhelmed by the standard and the complexity of VHDL coding, while still leaving space for a sense of achievement and experimenting with the results of their work. The software and drivers for controlling and configuring the transmitter are also given to them, and run on the CPU. The software can dynamically modify the code rate and modulation parameters of the DVB-S2 transmitter and calibrate the front-end, while the drivers manage the communication with the camera.

We emphasize that this approach is tailored to the technical background of third-year Bachelor students. They have already covered all mathematical background and been introduced to topics such as digital signal processing, programming, and circuit design. However, they are not familiar yet with error coding, modulation theory, propagation issues, or VHDL. If the EAGLE course was designed for the Master level, much less or no content would need to be provided.

Evaluation Criteria and Resources Available to Students

In the EAGLE course every group needs to autonomously elaborate a plan for the design of the six modules and justify the amount of resources (team members) dedicated to each one of them. No information is given to them in this regard, but as the wireless video link module needs to already be operative in the first phase of the EAGLE mission, it is expected that all groups prioritize this task from the beginning, dedicating two to three students for its implementation.

Regarding grading, every group is evaluated based on both technical (meeting a predefined set of deadlines) and non-technical aspects (managing skills, out-of-the-box thinking, work ethics), and several demo sessions are expected during the year. An online platform stores all material given to the students for the design of each module. Inside this platform every group also has a blog in which they have to explain their progress and fulfill a peer review assessment, from which the individual grade is obtained in combination with the group grade. Knowledge transfer among team members is done through their regular meetings and the online platform, where several question and answer (Q&A) sessions are additionally planned with the teaching staff after the weekly sessions.

There are six technical experts involved in the course, advising students with respect to each of the six technical modules. In addition, each team

[1] The open source implementation of the DVB-S2 transmitter developed for this project can be found at http://opendvb.org. Any code related to parts students need to implement is available only upon request.

of students has a team of three coaches for the non-technical aspects. The number of coaches per team may seem high, but usually only one takes an active role in the group, while the remaining two are left as backups in case of absence. Students work on the project every Monday afternoon throughout the year, and are supposed to organize themselves and ask relevant questions of the technical experts. The first author of this article is the technical expert for the wireless video link.

An In-Depth Overview of the Wireless Video Link Module

This section offers a more detailed description of the wireless video link module and elaborates on the results obtained in the current academic year. The DVB-S2 standard is briefly introduced, and the breadth of topics relevant to engineering curricula covered is discussed. We describe the design and implementation phases of the wireless video link module. Finally, we give an overview of the progress of the students and present their feedback on the learning process.

The DVB-S2 Standard and Its Educational Applications

The complete specification of the DVB-S2 standard can be found in [3], while we give a short overview of the main sub-blocks and discuss their application to engineering curricula.

The block diagram of the DVB-S2 transmitter can be found in Fig. 4. It consists of several sub-modules that can be indexed into four main categories:

1. Baseband (BB) processing blocks. These blocks format the incoming video data, and build the baseband frame by inserting the BB headers and adding the cyclic redundancy check (CRC). Finally, the whole stream is scrambled.
2. Forward error correction (FEC) blocks. These comprise Bose Chaudhuri (BCH) and low-density parity check (LDPC) error correction coding, as well as an interleaver.
3. Constellation mapping and physical layer (PL) framing blocks. They build the actual quadrature symbol stream.
4. Modulation and filtering block. In this block the RRC filter is applied. The data is formatted in a way the front-end can understand.

The input interface inside the BB processing blocks can be used for explaining the relevance of headers and control information. It shows how they allow the receiver to know how to decode the data stream. Moreover, it can be taken as an example of how communications systems are layered. The data stream from one technology, the camera output, is fragmented and packetized into upper layer frames for another technology.

The CRC encoder serves the students as an introduction to error detection codes, while the scrambler can explain the need for bias-removal techniques.

The FEC blocks extend the knowledge introduced by the CRC encoder. They can also be used to further strengthen the mathematical background of the students as they are based in polynomial division and several mathematical properties that allow one to detect if the received bitstream is corrupted by the channel.

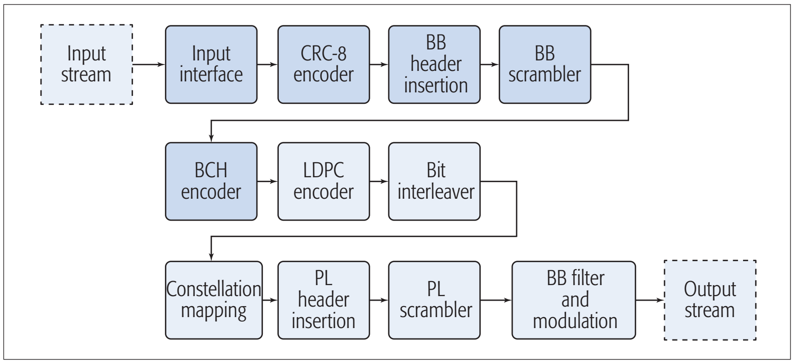


Figure 4. DVB-S2 transmitter block diagram. Shaded blocks are to be implemented by students during the project.

The constellation mapping blocks introduce the students to several modulation schemes, and their performance and robustness to errors.

The use of the final RRC filtering block in the educational process can be twofold: first, it can reinforce the digital signal processing (DSP) background of the students on digital filter design as well as offer a textbook illustration of the Nyquist pulse shaping criterion and the concept of matched filtering; second, it allows discussion about the spectral efficiency of the chosen modulation scheme and more high-level topics such as spectrum regulation, where such a filter is used to meet spectrum mask constraints.

As introduced before, students receive a partial implementation of the transmitter, starting from the LDPC encoder onward. This is done to specifically introduce or reinforce the topics covered by the BB processing and BCH blocks, which are taught in the remaining subjects studied during the academic year.

Design and Implementation of the Wireless Video Link

Design Phase: When the complete picture of the transmitter is clear, students need to design the BB processing and FEC blocks.

For the design of the BB processing an input interface, CRC-8 encoder, frame builder (BB header insertion), and scrambler need to be designed. In the ZyBo architecture the incoming MPEG-2 video stream is delivered in byte form, while the transmitter must process one bit per clock cycle. Therefore, the input interface must serialize this data stream and handle the synchronization of the transmitter with the camera to ask for a new byte of data only when the previous one has been processed and the remaining blocks are ready. Regarding the frame builder, the students need to choose the appropriate values of the header according to the application and build the BB frames, ensuring that the incoming video stream is fragmented to fit within the frame length. For the CRC-8 encoder, a linear feedback shift register (LFSR) must be designed in order to compute the checksum. This design can be reused for the scrambler, teaching the students the different applications of LFSRs.

For the design of the BCH encoding block, the standard describes the mathematical operation using polynomial division. Students need to

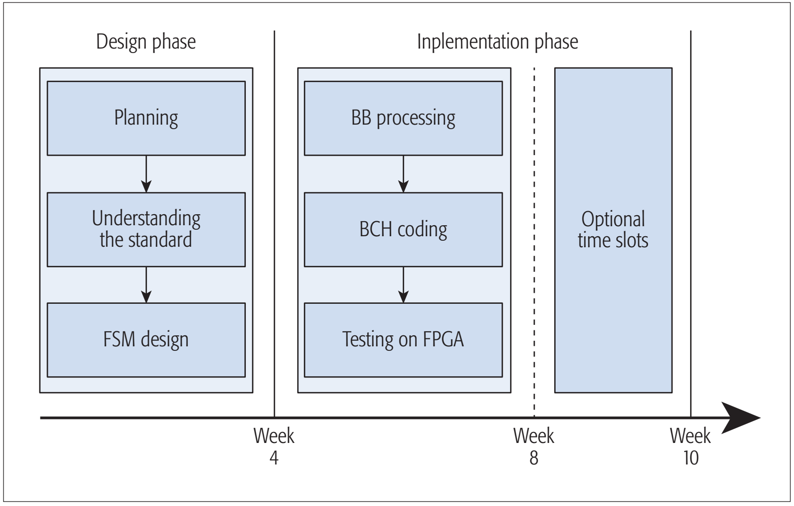


Figure 5. Expected schedule and milestones for the design and implementation phases of the wireless video link module. Extra time slots are reserved to optionally extend the functionalities of the transmitter, or to ensure that the baseline transmitter is working.

understand this operation and how the result is obtained, but implementing it this way would not be feasible in the FPGA. They need to again relate this operation with LFSR and propose a simple design for computing the BCH parity bits.

Implementation Phase: The implementation phase covers the actual coding of the selected sub-blocks, based on the proposed design. As testing and debugging a VHDL implementation with the off-the-shelf receiver results in a complex and binary process (it correctly receives and decodes the signal if there is no error in the implementation and completely fails to do so if there is), several standard compliance testbenches are provided to help locate faults in sub-blocks. This reinforces their critical thinking and teaches them the process of modular design, implementation, and debugging. The tests use custom test vectors containing the bitstreams before and after each block, so the code can be tested in pieces. Students are provided with tests for the BB processing with and without scrambler, the BCH encoder, the full FEC module, and finally the full transmitter, in which handshaking issues are tested. In order to test the handshaking, a `bit_ready` and `bit_valid` input and output port for each sub-block are defined. Each of the blocks coded by the students needs to process the signal only when the incoming bit is valid, while always waiting to deliver the output until the next block is ready to receive it. We generate random pulse sequences in these signals to ensure the proper synchronization of the whole transmitter. In this way, the students become aware of not only the theoretical concepts but also the practical challenges of a real-life system.

Once all tests are passed, the VHDL project is compiled and flashed into the FPGA for performance testing. In this last phase students need to also become familiar with the software for controlling the transmitter, and will finally be able to see the streaming of the camera.

Milestones and Deliverables: The wireless video link is expected to be designed and implemented during 10 weekly worksessions of 5 hours. There are three milestones in the global planning, which correspond to the three mandatory demo sessions planned for the first semester. The expected milestones are depicted in Fig. 5.

The design phase takes place during the first four weeks, and its main deliverable is a complete pseudocode design of the BB processing and BCH blocks using a finite state machine (FSM) approach. The complete DVB-S2 standard as well as material covering FSM design and VHDL coding is given to the students in this phase.

The implementation phase is divided into two sub-phases: another four weeks are planned for the VHDL coding, testing, and flashing into the FPGA, with two optional sessions at the end. These two sessions may be used for implementing adaptive coding and modulation (ACM) on the transmitter, characterizing its performance vs. signal-to-noise ratio (SNR) through a measurement campaign or simply finishing the project in case it is not done by then.

Student Progress and Feedback

Figure 6 shows the progress of the students during the first semester of the current academic year. The main conclusion that can be obtained is that the course represents a very challenging but achievable task for the students. Their performances are well spread, with all groups except one having a working VHDL code after the 10 sessions. However, the optional tasks (ACM and characterization of performance vs. SNR) were neglected by all groups.

Additionally, after the final week students are asked to submit a report providing feedback on the course. They have to reflect on the main difficulties found during the project and express their level of satisfaction with the results obtained and the learning process. A comparison with other project-based courses studied before as well as suggestions to improve the project for coming academic years are also asked.

Regarding the main difficulties, understanding the standard was the common one for all groups. However, while perceived as difficult, the standard did improve the overall satisfaction of the students with the learning process. The vast majority of them reported that, although challenging, they preferred using a real application as the driver for learning VHDL. This is contrary to other project-based learning courses they have done. Moreover, as the standard covers more topics than they have already studied, they appreciated being introduced to them. As a general suggestion, some initial sessions explaining the standard in more detail would have reduced the complexity of the project.

Summary of Educational Aspects

First, through the study of a real standard that is naturally decomposed in layers and functional blocks per layer, students learn how the divide and conquer approach is done in real technologies.

Second, students get their first hands-on experience in VHDL coding. After understanding the basics of VHDL, they learn to implement the individual blocks and to connect them considering the relevant interdependencies.

Of particular interest is the fact that students learn not only the relevant functionality for DVB-S2, but also how to implement this efficiently in an FPGA. Here, they learn to understand the trade-off between functionality and implementation, and see how a standard is actually a representation of this trade-off.

Next to technical skills, students also improve their soft skills. Project management, planning, and communication also play a crucial role in EAGLE in order to meet the target deadlines. Students first have to decide how to divide the team of ten people over the different modules, how to meet regularly to discuss progress, and how to interact with the technical experts. It involves having discussions and making decisions with other teams about the functional decomposition of the tasks and the interfaces between the teams in order to guarantee smooth integration. It is very interesting to see that there is a strong correlation between technical progress and teamwork; that is, teams that have good planning, communication, and teamwork skills also outperform on the technical milestones.

Finally, students learn to write technical reports about the progress they made and the problems they encountered. This helps them to identify what is going wrong, narrow down the problem to the real bottleneck, and learn how to avoid it in the future.

Conclusions

In this course we used the DVB-S2 standard to teach students how to go from theory to practice and how to manage a complex engineering project.

There was some initial skepticism about the challenging nature of the project, and during the first two to four weeks students indeed felt overwhelmed. But it was remarkable to see that, after the initial reading of the standard, students conquered the most important insights, managed to get their VHDL code implemented, compiled, and tested, and delivered a working prototype. This disproves the idea that standards are too complex for Bachelor student education.

Several parts of the project were already given to the students. Depending on the curriculum and time allocated for the course, more blocks could be covered. If knowledge about PCB design or control of an analog front-end is needed, students could be asked to design a transmitter or focus more on the control of the front-end. If more focus is on baseband modulation theory, students could be asked to implement the extension for terrestrial broadcast, DVB-T2. It could also be interesting to compare the performance of both indoor and outdoor scenarios and understand the impact of multipath fading. If more focus is on the higher layers of the protocol stack, such as IP encapsulation or even the application layer, students can be asked to focus on the network stack. Finally, even application layer video compression could be optimized to meet the specific requirements of the EAGLE mission, and the students could be introduced to the concepts of cross-layer design.

As a conclusion and to emphasize the benefits of using a standard in a project-based learning course, being able to recognize a familiar platform (TV broadcasting) and learning the standard increases the motivation of the students to learn VHDL and code the transmitter, contrary to building a completely made up project that is not related to a real-world application.

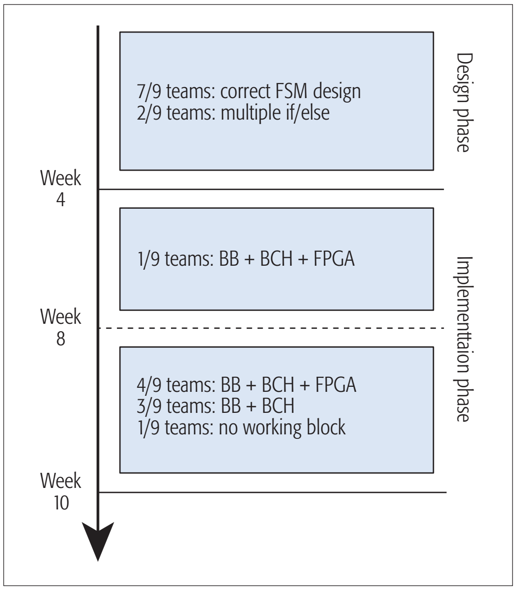


Figure 6. Student progress on the milestones. In the design phase all teams were able to make a plan and a correct VHDL design, although two teams did not follow an FSM approach and used multiple *if else* statements. In the implementation phase one team was able to finish in less than eight sessions, and four teams finished in ten sessions. Three teams had minor bugs in the FPGA implementation, while one team was not capable of having a working VHDL code.

To emphasize the benefits of using a standard in a project-based learning course, being able to recognize a familiar platform (TV broadcasting) and learning the standard increases the motivation of the students to learn VHDL and code the transmitter, contrary to building a completely made up project that is not related to a real-world application.

References

[1] A. M. Agogino *et al.*, "Making Connections to Engineering During the First Two Years," *Proc. 1992 IEEE Frontiers in Education Conf.*

[2] M. J. Pavelich, B. M. Old, and R. L. Miller, "Real-World Problem Solving in Freshman/Sophomore Engineering," J. Gainen and E. Willemsen, Eds., *Fostering Student Success in Quantitative Gateway Courses*, Jossey-Bass, 1995.

[3] "Digital Video Broadcasting (DVB); Second Generation Framing Structure, Channel Coding and Modulation Systems for Broadcasting, Interactive Services, News Gathering and Other Broadband Satellite Applications (DVB-S2)," Nov. 2014, https://www.dvb.org/standards/dvb-s2, accessed Feb. 2017.

[4] Xilinx Inc. Zybo Zynq-7000 ARM/FPGA SoC Trainer Board, https://www.xilinx.com/products/boards-and-kits/1-4azfte.html, accessed Feb. 2017.

[5] C. L. Dym *et al.*, "Engineering Design Thinking, Teaching, and Learning," *J. Engineering Education*, vol. 94, 2005, pp. 103–20.

[6] A. J. Dutson *et al.*, "A Review of Literature on Teaching Design Through Project-Oriented Capstone Courses," *J. Engineering Education*, vol. 76, no. 1, 1997, pp. 17–28.

[7] B. Bennett *et al.*, "DVB-S2 Benefits for Military Broadcast Systems," *MILCOM 2005*, Atlantic City, NJ, 2005, vol. 1, pp. 148–54.

[8] KU Leuven, Project and Design EAGLE Course, https://onderwijsaanbod.kuleuven.be/syllabi/n/H01Q6CN.htm (in Dutch), accessed Feb. 2017.

[9] "Unidirectional Lightweight Encapsulation (ULE) for Transmission of IP Datagrams over an MPEG-2 Transport Stream (TS)," IETF RFC 4326, https://tools.ietf.org/html/rfc4326, accessed on Febr. 2017.

[10] A. Mayer *et al.*, "Analytical and Experimental IP Encapsulation Efficiency Comparison of GSE, MPE, and ULE over DVB-S2," *2007 Int'l. Wksp. Satellite and Space Commun.*, Salzburg, Austria, 2007, pp. 114–18.

[11] Zedboard, Xilinx Zynq-7000 ARM/FPGA SoC Based Development Board, http://zedboard.org/product/zedboard, accessed Feb. 2017.

[12] Raspberry Pi Camera Module v2, https://www.raspberrypi.org/products/camera-module-v2/, accessed Feb. 2017.

[13] TBS5980 QBOX CI DVB-S2 TV tuner USB, external TV tuner box for laptop and PC, http://www.tbsdtv.com/products/tbs5980-dvb-s2-ci-tv-tuner-usb.html, accessed Feb. 2017.

Biographies

Yuri Murillo (yuri.murillo@esat.kuleuven.be) obtained his M.Sc. degree in telecommunication engineering from the Public University of Navarre (UPNA), Spain, in 2013. He wrote his Master's thesis at TU Delft, the Netherlands. From 2013 to 2015 he worked as an IT security consultant at GMV Secure-eSolutions and obtained a Master's degree in business administration from UAH, Madrid. He is currently pursuing a Ph.D. degree at KU Leuven, Belgium. His main research interests include optimization of dense wireless mesh networks and opportunistic networking protocols.

Bertold Van den Bergh (bertold.vandenbergh@esat.kuleuven.be) is a Ph.D. student in the Department of Electrical Engineering (ESAT), KU Leuven, Belgium. His research is focused on reliable communication systems for UAVs and other applications. Before starting his Ph.D. he obtained his Master's 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.

Jona Beysens (jona.beysens@esat.kuleuven.be) obtained his B.Sc. and M.Sc. degrees in electrical engineering from KU Leuven in 2014 and 2016, respectively. After the first year of his Master's, he did an internship at National Chiao Tung University, Taiwan, designing a lightweight preprocessing algorithm for a deep learning convolutional neural network. Currently, he is pursuing a Ph.D. degree at KU Leuven, focusing on visible light communication for dense indoor networks and MIMO processing.

Alexander Bertrand [M'08, SM'16] received his Ph.D. degree in engineering sciences (electrical engineering) from KU Leuven in 2011. He is currently an assistant professor with the Electrical Engineering Department, KU Leuven. His research focus is on signal processing algorithm design for (distributed) sensor arrays.

Wim Dehaene has been a professor at KU Leuven since 2002. His main research interests are ultra low-power DSP and memories. Besides that, he also heads high-speed digital circuit research, mainly in the domain of digital RF transmitters and digital AD converters. He is also very active in research toward novel STEM education programs for the Flemish secondary school system. He is a member of the ESSCIRC TPC and served on the ISSCC TPC for many years.

Panagiotis Patrinos has been an assistant professor at ESAT, KU Leuven since 2015. During fall/winter 2014 he held a visiting assistant professor position at Stanford University. His current research interests are in structured optimization and model predictive control with a broad range of application areas including smart grids, water networks, aerospace, and machine learning.

Tinne Tuytelaars is a professor at the PSI research division of KU Leuven's ESAT.Her research interests lie in computer vision, and in particular object recognition, image representations, and transfer learning.

Ruth V. Sabariego is an associate professor with EnergyVille, ESAT, KU Leuven since October 2013. In October 2000, she joined the Department of Electrical Engineering and Computer Science, University of Liège, Belgium, where she received her Ph.D. degree in applied sciences in 2004 and stayed as a post-doctoral researcher until September 2013.

Marian Verhelst [S'01, M'09, SM'14] became an assistant professor at the Micro-Electronics and Sensors (MICAS) laboratories of ESAT, KU Leuven in 2012. Her research focuses on self-adaptive circuits and systems, embedded machine learning, and low-power sensing and processing for the Internet of Things. She is a member of both the ISSCC and DATE TPCs, and of the DATE and ISSCC Executive Committees. She was an Associate Editor for *TCAS-II* and currently is for *JSSC*.

Patrick Wambacq [SM] received his Master's degree in electrical engineering in 1980, and his Ph.D. degree in electrical engineering in 1985, both from KULeuven. He started as a lecturer at KU Leuven in 1985, becoming a professor in 1989 and a full professor in 2004. Initially he was interested in computer vision and industrial applications thereof, but since 1998 has been the head of the speech group at KU Leuven ESAT focusing on speech recognition and its applications.

Sofie Pollin [SM] (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 the University of California, 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 ESAT, 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.

Teaching Communication Technologies and Standards for the Industrial IoT? Use 6TiSCH!

Thomas Watteyne, Pere Tuset-Peiró, Xavier Vilajosana, Sofie Pollin, and Bhaskar Krishnamachari

The authors discuss how the 6TiSCH stack can be incorporated into existing and new curricula to teach the next generation of electrical engineering and computer science professionals about designing and deploying such networks. They also give a comprehensive overview of the 6TiSCH stack and the tools that exist to support a course based on it.

ABSTRACT

The IETF 6TiSCH stack encompasses IEEE802.15.4 TSCH, IETF 6LoWPAN, RPL, and CoAP. It is one of the key standards-based technologies to enable industrial process monitoring and control, and unleash the Industrial Internet of Things (IIoT). The 6TiSCH stack is also a valuable asset for educational purposes, as it integrates an Internet-enabled IPv6-based upper stack with state-of-the-art low-power wireless mesh communication technologies. Teaching with 6TiSCH empowers students with a valuable set of competencies, including topics related to computer networking (medium access control operation, IPv6 networking), embedded systems (process scheduling, concurrency), and wireless communications (multipath propagation, interference effects), as well as application requirements for the IIoT. This article discusses how the 6TiSCH stack can be incorporated into existing and new curricula to teach the next generation of electrical engineering and computer science professionals about designing and deploying such networks. It also gives a comprehensive overview of the 6TiSCH stack and the tools that exist to support a course based on it.

INTRODUCTION

All signs indicate that the Industrial Internet of Things (IIoT) revolution is upon us. This revolution relies on ubiquitous low-cost sensor- and actuator-enabled devices with constrained computational capabilities that operate autonomously and are connected to the Internet through low-power wireless communication technologies. The deployment of such devices within the factory floor will bring together operation technologies (OT) and information technologies (IT), and enable industrial process monitoring and control applications. It is therefore of great importance to train the next generation of electrical engineering (EE) and computer science (CS) professionals on these low-power wireless communication technologies and standards that will be used to connect devices to the Internet in the industrial space.

We have observed a rising trend to use Wi-Fi (IEEE 802.11) and the IPv4 stack as the communication standards to teach IIoT-related courses. We acknowledge that there are many benefits to using Wi-Fi and IPv4 for this purpose, including the fact that Wi-Fi connectivity is ubiquitous, and IPv4 is still the de facto protocol stack to connect to the Internet. However, if the focus is on connecting devices to the Internet for industrial applications, this approach is flawed for several reasons. First, Wi-Fi is not low-power and does not provide the determinism and robustness levels required by industrial applications, thus limiting its applicability. Second, the IPv4 stack has been superseded by the IPv6 stack, which provides several advantages in terms of simplicity, scalability, efficiency, quality of service, and security. Third, chips implementing Wi-Fi and IPv4 are typically provided as a black box, thereby limiting the opportunities to learn how they operate at the lower levels.

In contrast, standards for low-power wireless communications specifically targeted at the IIoT are reaching technical maturity and being adopted by industries around the world. One particular set of standards is being developed by the Internet Engineering Task Force (IETF) 6TiSCH standardization Working Group. 6TiSCH builds on the IEEE802.15.4 time slotted channel hopping (TSCH) link layer without changing it, and combines it with an IPv6-ready "upper stack." The result is the best of both worlds: industrial-grade performance and ease of integration into the Internet. Despite the fact that the standardization process is not finished, there are "pre-6TiSCH" products already available on the market, such as Linear Technology's SmartMesh IP.

Compared to the Wi-Fi/IPv4 approach, there are several elements that make 6TiSCH an ideal match to teach how to connect devices to the Internet. First, it is based on truly low-power wireless mesh communication technology that has been validated in the industry for the past decade. Second, it is based on next-generation Internet standards currently being developed, which allows following the standardization process and learning from it. Third, it is a simple but complete stack, making it easy to understand the basic concepts but allowing us to go deep into protocols and the implementation details.

This article shows how 6TiSCH can serve as an accelerator for students to learn a fully standards-based communication stack and get familiar with key IIoT protocols.

TECHNICAL OVERVIEW OF 6TISCH

This section offers an overview of 6TiSCH, discussing the protocol stack, its relevance, and success stories.

Digital Object Identifier: 10.1109/MCOM.2017.1700013

Thomas Watteyne is with EVA; Pere Tuset-Peiró and Xavier Vilajosana are with the Universitat Oberta de Catalunya; Sofie Pollin is with KU Leuven; Bhaskar Krishnamachari is with the University of Southern California.

 0163-6804/17/$25.00 © 2017 IEEE

The 6TiSCH Protocol Stack

Figure 1 shows the 6TiSCH protocol stack. It is composed of standards developed by the IEEE[1] and the IETF.[2] Together, these standards achieve IPv6-based end-to-end connectivity while meeting the determinism, robustness, and low-power operation required for battery-operated devices targeted at industrial process monitoring and control applications [1].

This section provides a brief introduction to each layer in the 6TiSCH stack — from the bottom up — highlighting the educational potential of each.

IEEE 802.15.4 defines the physical and data link layers for low-power wireless communication between battery-operated devices. It offers an appropriate trade-off between transmit power (0–10 dBm), data rate (250 kb/s), and maximum payload size (127 bytes). The new IEEE 802.15.4 TSCH mode is targeted at demanding industrial applications. TSCH combines network-wide synchronization and channel hopping to achieve over 99.999 percent end-to-end reliability and over a decade of battery lifetime.

TSCH is a basic but complete link layer protocol to teach students the basics of multiple channel access, frequency diversity, scheduling, and coexistence. It trades off throughput with packet error rate and delay by scheduling more or less redundant transmissions. Students can derive equations and perform simulations to learn how to do the performance analysis of TSCH-enabled networks of various sizes.

The IETF 6top Protocol (6P) [2] is being standardized by the 6TiSCH Working Group. It defines a distributed scheduling protocol whereby neighbor nodes negotiate to add/remove one or multiple cells in the TSCH schedule. Each cell is a "communication opportunity" for the neighbor nodes to exchange a link layer frame. Adding more cells increases the bandwidth and lowers the end-to-end latency, while at the same time increasing the nodes' power consumption.

6P is a simple protocol that allows students to not only see how the TSCH schedule is managed "down to the wire," but also touch on notions such as schedule consistency, network stability, and network churn.

IETF 6LoWPAN [3] allows long IPv6 packets (up to 1280 bytes) to fit into short IEEE 802.15.4 frames (at most 127 bytes). It is composed of two main mechanisms. First, it defines rules for compacting the IPv6 header. This is done by:

1. Removing fields that are not needed
2. Removing fields that always have the same contents
3. Compressing the IPv6 addresses by inferring them from link layer addresses

The result is that the 40-byte IPv6 header gets compacted down to a couple of bytes in the most favorable case. A low-power border router (LBR) sits at the edge of the low-power wireless network and is responsible for doing the transparent IPv6→6LoWPAN and 6LoWPAN→IPv6 translation: a computer outside the low-power wireless network interacts with a low-power wireless directly with its IPv6 address. Second, it defines fragmentation rules so that multiple IEEE 802.15.4 frames can make up one IPv6 packet.

6LoWPAN allows students to really understand every single bit in the IPv6 header and feel the power of the Internet. That is, once the devices implement 6LoWPAN correctly and the LBR is running, just by injecting a correctly formatted packet into the Internet, the tiny low-power wireless device sitting on the student's desk can interact with a computer halfway around the globe.

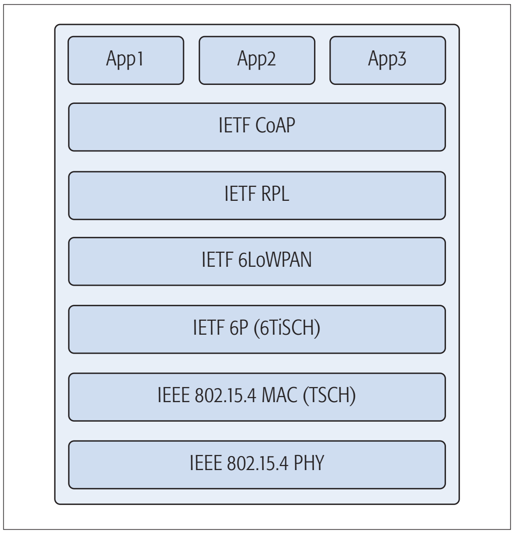


Figure 1. The IETF 6TiSCH protocol stack, composed of IEEE 802.15.4, IETF 6P, 6LoWPAN, RPL, and CoAP.

CoAP is not particularly complicated from a protocol point of view; it's header is 4 bytes long and it is "just" an application protocol. But by using this protocol, students can "put everything together" and really get that interacting with a 6TiSCH devices is just as easy as interacting with a web browser.

IETF RPL [4] is an intra-domain routing protocol for low-power wireless mesh networks. RPL organizes the network into a directed acyclic graph (DAG) rooted at a gateway device that connects to the Internet. As with other distance vector protocols, devices regularly advertise their distance to the root, allowing neighbors to compute their own.

Despite being "simple" (there are only two required packet formats), by playing with RPL students can understand the complexity of maintaining a coherent multihop routing structure in a network made up of "lossy" wireless links.

IETF Constrained Application Protocol (CoAP) [5] is the "HTTP for constrained devices." It turns every low-power wireless device into a web server and a browser, and allows web-like interactions. A constrained node can publish its sensor readings onto a server on the Internet, or a smartphone can operate smart blinds, all by using the communication paradigms of the Internet. Extensions to popular browsers and open-source libraries make adding CoAP support to an computer application easy.

CoAP is not particularly complicated from a protocol point of view; its header is 4 bytes long and it is "just" an application protocol. But by using this protocol, students can "put everything together" and really get that interacting with a 6TiSCH devicesis just as easy as interacting with a web browser. Moreover, thanks to standardization, there is no need to learn exotic protocols, as third party libraries and tools are readily available.

Success Stories and Relevance

The 6TiSCH stack did not come out of thin air; it is the result of a rigorous multi-year standardization effort. 6TiSCH is the latest generation of protocols exploiting TSCH technology, and inherits from the lessons learned from TSMP (2006, proprietary protocol), WirelessHART (2008), and ISA100.11a (2011).

[1] https://www.ieee.org/

[2] https://www.ietf.org/

There is a clear momentum behind the 6TiSCH stack. From an educational perspective this means that it makes sense to teach 6TiSCH to next generation EE and CS professionals, facilitated by a wide availability of tools and implementations. One challenge is, of course, that the standardization activity is ongoing, which means the technology might still evolve.

Since its formation in October 2013, 6TiSCH has followed the rigorous IETF standardization process. Three hundred ninety-eight people — with a mix of technology providers, end-user companies, and academics — follow the standardization activity of the group through its mailing list and periodic face-to-face and online meetings.

Several companies have commercialized or are preparing "pre-6TiSCH" products. One example is Linear Technolgy's SmartMesh IP product line. A vibrant community has grown around 6TiSCH, with open source implementations, open hardware platforms, open testbeds, "pre-6TiSCH" commercial off-the-shelf solutions, simulation platforms, and other tools.

There is a clear momentum behind the 6TiSCH stack. From an educational perspective this means that it makes sense to teach 6TiSCH to next generation EE and CS professionals, facilitated by a wide availability of tools and implementations. One challenge is, of course, that the standardization activity is ongoing, which means the technology might still evolve.

IIoT Course Objectives

The topic of "Industrial IoT" is an educational goldmine. A course on IIoT should be targeted at the advanced undergraduate or graduate level. It can target both EE and CS students, and can be taught to mixed groups. Ideally, prerequisites include classical computer networking, computer architecture, computer programming, and digital electronics. Depending on the type of course, some of these requirements can be lifted, at the cost of more "hand-holding" of students.

Probably the most specific factor of an IIoT course — and a key differentiator with other courses at that level — is its systems nature. That is, the course will allow the student to "put everything together" and build a minimal but fully functional device that connects the physical and digital worlds by interfacing physical sensors and actuators to a micro-controller, and interacts with other devices on the Internet. During the process, the student will be able to get his/her hands on the complete chain (hardware, software, standards) and put his/her prior knowledge to the test.

One important limitation of 6TiSCH is that standardization is still underway, so the technical specifications the students will be working on might not be stable.

Core Educational Focus

There are three core educational focus points to an IIoT course.

Core Focus 1: Medium Access Protocols. What makes 6TiSCH different from other low-power wireless technologies is the fact that it uses TSCH technology. Hence, a good first focus point are medium access control (MAC) protocols. In particular, the first goal is to teach the students about the main challenges of building a reliable low-power wireless network: multi-path fading and external interference. The instructor can then touch upon TSCH, the technique used to achieve both channel hopping (to bring reliability) and low-power operation.

Core Focus 2: Standardization Process. Based on core focus 1, the course can then focus on the standards that exploit channel hopping and synchronization. This discussion is a great way to introduce the standardization process: What are the main standardization bodies? Who contributes to them and for what reasons? It is a good idea to highlight what makes the IEEE and the IETF different, but stressing the point that they both contribute equally to building the Internet. Somewhat counterintuitively, standardization is usually well received by students who appreciate learning about what it takes to turn an idea into interworking devices. 6TiSCH, still being standardized, is a great opportunity to learn about the process as standardization happens.

Core Focus 3: End-to-End Systems. The goal of any IoT system, including the IIoT, is to connect low-power wireless devices to the Internet. Probably the most important focus in a course is to take a systems approach and *build* an end-to-end system. Students are usually very engaged when "seeing it all work together." It is therefore important to spend some time discussing the backend system (including turn-key solutions such as MQTT, Node-Red, Thingworx, and IBM Bluemix).

Optional Educational Focus

An additional three focus points can be added as options to the course, depending on the students' background and interests.

Optional Focus 1: Low-Level Programming (*for CS*). CS students usually enjoy getting to the bottom of a computer system, and the embedded nature of an IIoT system is a great example. The availability of open source implementations are a great opportunity to discover the inner details of embedded operating systems (concurrency, task switching, memory footprint) and low-level peripherals (timers, communication buses, radio transceivers).

Optional Focus 2: Hardware (*for EE*). IIoT devices are low-power wireless and battery operated in the vast majority of cases. They consist of a micro-controller and a radio connected to sensors and actuators. Hence, they present a unique opportunity to delve into the design of low-power circuits (including the trade-off between performance and power consumption when choosing a micro-controller and a radio chip) and the integration of sensors and actuators. Advanced groups can even go through the exercise of manufacturing and characterizing a printed circuit board.

Optional Focus 3: Security (*for CS and EE*). The cyber-physical nature of the IIoT, combined with the critical requirements of the industrial applications in which it is used, makes the importance of security clear to the students. With IIoT, it is important to stress that security does not come as an optional add-on, but must be built in from the start. As with all systems, security is built into every layer of the protocol stack, including link layer (Cisco Call Manager [CCM]), transport (Datagram Transport Layer Security [DTLS]), and application (object security). A security focus point should contain an overview of these different levels and some hands-on exposure to at least one of them.

Tools and Resources around 6TiSCH

This section serves as a listing of the 6TiSCH-related tools and resource available to a person teaching IIoT using 6TiSCH. We only list the tools and resources the authors of this article have either used for teaching, or have directly contributed to.

Open Source Academic Implementations

As detailed in [6], 6TiSCH is now supported by the three major open source low-power wireless stacks:

OpenWSN[3] [7] is an open source implementation specifically targeted at 6TiSCH. It was started at the University of California (UC) Berkeley in 2010, and has been considered the reference implementation for the 6TiSCH protocol stack, including in 6TiSCH interoperability events. Its use in research and industrial applications is increasing, and OpenWSN is now supported on the most popular platforms. There are several tutorials available. OpenWSN includes the OpenSim emulator, detailed below.

Contiki[4] is an open source operating system targeting low-power devices and aiming to provide the user with a complete toolbox for the design of IoT solutions. There are plenty of examples of how to use it online, and it is coupled elegantly with the Cooja simulator, which enables simulation of a protocol implementation before it is deployed in a testbed. Currently, Contiki supports TSCH.

RIOT[5] is an open source operating system for the IoT. It is developer friendly, as it allows easy programming in C/C++ and is supported by a wide range of platforms. Currently, RIOT supports TSCH through the OpenWSN project.

Open Hardware Platforms

In principle, all hardware platforms that include a micro-controller and an IEEE 802.15.4 radio transceiver are suitable for implementing the 6TiSCH stack, since the core requirement to support it is a timer with capture and compare registers. For example, the 6TiSCH stack has been successfully ported on the TelosB platform, which was developed at UC Berkeley back in 2004. Modern platforms based on the Cortex-M architecture are better suited to support the 6TiSCH stack given the fact that they are less constrained in terms of RAM and Flash memory.

A good example is the OpenMote platform[6] (Fig. 2), which was developed at UC Berkeley in 2013 as a replacement for the TelosB. The OpenMote platform is based on the Texas Instruments CC2538 system on chip (SoC). The main advantage of the SoC architecture compared to the dual-chip approach is that it removes the timing limitations associated with the micro-controller to radio transceiver communications based on a bus. This is particularly important for the 6TiSCH time-sensitive implementation.

Open Testbeds

Testbeds are an important tool for teaching since they enable large-scale experimentation while removing the need for each student to buy hardware [8].

The **IoT-Lab**[7] is a 2728-node testbed deployed over 7 locations in France and Germany. Three main hardware platforms are supported in the IoT-LAB to conduct experiments, representative of the variety of IoT hardware. IoT-LAB offers a web-based reservation system, a set of tools that enable application development and real-time debugging, and tools to measure current consumption.

TutorNet[8] is a 100-node low-power wireless embedded IoT testbed that has been operational for nearly 10 years. It has been used for research and teaching of many WSN protocols, including multi-channel scheduling and routing protocols relevant to 6TiSCH. Originally consisting of TMote Sky nodes, it is currently being upgraded to the OpenMote platform to support new operating systems, such as OpenWSN and RIOT, and new standards like 6TiSCH. The testbed allows researchers to deploy code on any subset of nodes, run the experiments, and collect the data. Currently, visual tools to program the nodes are being developed, and the whole software package is being documented to allow for replication at other institutions.

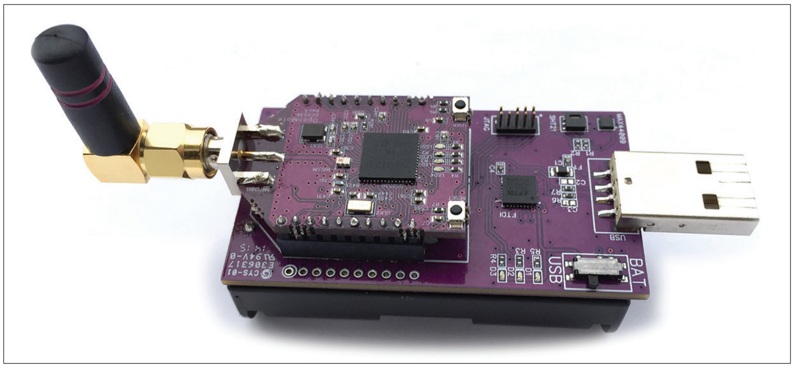


Figure 2. The OpenMote platform with the OpenUSB carrier board. The OpenUSB contains various sensors (temperature, relative humidity, light, 3-axis acceleration) and allows a user to reprogram it over USB.

"Pre-6TiSCH" Commercial Off-the-Shelf Solutions

The 6TiSCH standardization process is not finished yet, so commercial 6TiSCH products do not exist. That being said, several "pre-6TiSCH" commercial solutions exist, which combine the industrial performance of TSCH with an IPv6-enabled stack in a way very similar to the approach taken by 6TiSCH.

One example is Linear Technology's SmartMesh IP product line, which has been commercially available since 2011. The specificity of this solution is that Linear Technology manufactures its own SoC. The result is a best-in-class solution that achieves over 99.999 percent end-to-end reliability and over a decade of battery lifetime. A SmartMesh IP starter kit[9] (Fig. 3) is available that contains all the pieces necessary to build and benchmark a network. This kit has been used in numerous courses and educational materials.

Simulation Platforms

Several tools have been created to simulate the performance of a 6TiSCH network.

NS-3 is the most widely adopted tool for network simulation. It provides a large database of protocols, tools, and propagation models to emulate almost any kind of network. There are several ongoing efforts to add support for IEEE802.15.4 TSCH and 6TiSCH.[10]

The **6TiSCH simulator**[11] is a Python-based high-level simulator maintained by the 6TiSCH Working Group. It allows students to benchmark the performance of a network implementing the full 6TiSCH stack.

OpenSim is the emulator that is part of the OpenWSN project. That is, the OpenWSN firmware can be compiled as a Python extension module, and run within OpenVisualizer, the OpenWSN software running on the gateway computer. In this setup, the same firmware runs, and allows debugging and performance estimation without requiring real hardware.

[3] https://www.openwsn.org/

[4] http://www.contiki-os.org/

[5] https://riot-os.org/

[6] http://www.openmote.com/

[7] https://www.iot-lab.info/

[8] http://anrg.usc.edu/www/tutornet/

[9] http://www.linear.com/solutions/3106

[10] https://github.com/EIT-ICT-RICH/ns-3-dev-TSCH

[11] https://bitbucket.org/6tisch/simulator/

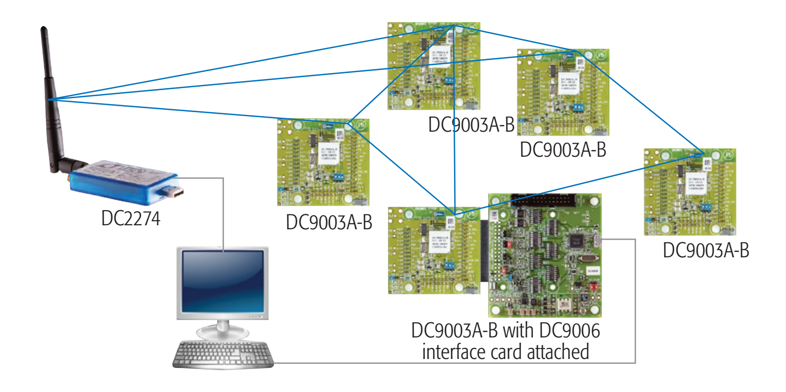


Figure 3. Linear Technology's SmartMesh IP starter kit allows students to experiment with the market leading low-power wireless IIoT solution.

Cooja is a network simulator linked to the Contiki project. It is developed in Java and is built around the concept of the COOJA mote. The Contiki native code, written in C, is compiled as a shared library for the MSP430 16-bit architecture using the GCC tool chain for the MSP430 platform. Since Contiki supports the 6TiSCH stack using a platform-agnostic approach, it is possible to simulate the operation of a node and a whole 6TiSCH network in Cooja.

Miscellaneous Tools

The channel-hopping nature of a 6TiSCH networks makes it challenging to capture over-the-air traffic, as typical sniffers can only listen to a single channel at a time. Beamlogic's sniffer[12] listens to all 16 frequencies in the 2.4 GHz band simultaneously, thus allowing the capture of all packets on air. The Argus project https://github.com/openwsn-berkeley/argus builds on it to allow co-located students to share one sniffer through the Internet.

Recent versions of Wireshark[13] come with 6TiSCH dissectors, including IEEE 802.15.4-2015 and the 6TiSCH 6top protocol. These dissectors are developed and maintained by the OpenWSN project team.[14] There are countless CoAP libraries written in virtually all programming languages (including Python[15]), which allow adding CoAP support to computer applications. The Copper extension to the Firefox browser[16] is still the easiest to use.

Available Experimental Results

Over the last couple of years dozens of papers have been published presenting results on the reliability [9], energy consumption [10], and synchronization accuracy [11] of 6TiSCH, complemented by application results exploiting and evaluating the 6TiSCH technology [12, 13]. These results bring fundamental insights into the performance of 6TiSCH.

Teaching the IIoT with 6TiSCH

The availability of numerous related resources and tools makes it easy to use 6TiSCH as a basis for an IIoT course. The instructor's role is to select the pieces he/she wants the students to focus on and to assemble them in a logical order. With such a diversity of educational focus points and tools to choose from, we believe it would be wrong to recommend a single course outline. Rather, we highlight the two main questions an instructor should ask him/herself when preparing a 6TiSCH-based course. For instructors in a hurry, we list several completely packaged and ready-to-teach 6TiSCH materials.

Add to an Existing Course or a New Course?

As an instructor, a first question to ask is whether the IIoT material will be added to an existing course or be the basis of a full new course.

The first approach is adding 6TiSCH to an existing course. Courses suitable for such an addition include introductory courses on computer networks, wireless networks, distributed systems, and embedded systems. Materials pertinent to 6TiSCH could be taught as a single module to illustrate concepts ranging from protocol layering in computer networks to interference mitigation in wireless networks. Alternatively, they may be interleaved throughout the course, for instance, by exposing students sequentially to different layers of the 6TiSCH stack, as they cover different layers of the protocol stack in a bottom-up (physical to application layer) or top-down (application to physical layer) manner.

The second approach is creating a full course based on 6TiSCH. The main challenge for such a course is that it should be broad and multidisciplinary, going from hardware all the way up to the application layer. When relying on the 6TiSCH standard, it is possible to cover all required protocols and technologies in one course, as 6TiSCH is supported by a broad range of tools used in networking design and test, from analysis to implementation. Within a full 6TiSCH course, we recommend a four-fold methodology that allows coverage of both the theoretical and hands-on knowledge:
- Deployment
- Simulation
- Testbeds
- Analysis

Depending on the choice of the instructor, it is possible to first start with hands-on experiments and then gradually move to simulation and analysis for understanding and explaining the experimental results. Alternatively, a course can follow the traditional design flow, starting from analysis, to design and simulation, and finally to deployment.

Open Source or Commercial?

One important question to ask ourselves as instructors is whether to use an open source implementation or an off-the-shelf commercial solution.

Understanding the distinction is crucial, and it is important for students to also understand it. Regardless of the enthusiasm of the communities behind them, open source implementations are never at the level of commercial solutions. The latter tend to "just work," while the former are often riddled with little bugs and undocumented subtleties, even if they might appear similar on paper.

As educators ourselves, our rule of thumb has been simple. For a course that focuses on building a real system which just works, we use a commercial solution for the low-power wireless portion. For a course that focuses on the internals of the wireless system, an open source implementation is acceptable, as long as the students are reasonably comfortable with the tools involved.

[12] http://www.beamlogic.com/

[13] https://www.wireshark.org/

[14] https://github.com/openwsn-berkeley/dissectors

[15] https://github.com/openwsn-berkeley/coap

[16] https://addons.mozilla.org/en-US/firefox/addon/copper-270430/

Building an IIoT system is complex, as it consists of a chain of elements (sensors, operating system, protocol stack, backend system, etc.) where each of them can fail. In our experience, using a commercial solution has allowed students to "go further" and build more polished solutions, at the cost of hiding some of the details of the low-power wireless network.

Teaching Track Record and Teaching Resources

The authors of this article have been using 6TiSCH as a vehicle for teaching the IIoT from half-day tutorials at conferences to a multi-week advanced undergraduate teaching courses. As a result, a lot of material has been prepared, which is readily available for instructors to use.

For a course using the open source approach, the OpenWSN project has been the foundation for half a dozen short courses. All the material, including presented slides, virtual machines, and pictures, are maintained on the project website.

For a course using the commercial approach, the DustCloud community[17] has been maintaining the "Dust Academy," a series of 32 labs students can follow using a SmartMesh IP starter kit. This material is ready to be presented, and instructors can get access to instructor-specific material, including the answers to the questions.

Conclusion

The 6TiSCH stack is an educational goldmine. It is a combination of protocols for the IIoT, resulting from a rigorous standardization process in the most prominent standardization development organizations. The 6TiSCH stack is widely seen as the key enabler for the IIoT, as it combines the performance of proven industrial low-power wireless communication technologies with the ease of use of an IPv6-enabled stack. A vibrant community has grown around it, making a plethora of tools and resources available to instructors, from open source implementations to "pre-6TiSCH" commercial products.

This article serves as a guide for an instructor to teach an IIoT course using 6TiSCH. Rather than recommending a single approach, it describes the educational focus points of an IIoT course, provides an exhaustive list of the tools and resources available, and discusses the key options an instructor has to choose when preparing the course.

Acknowledgments

This work is partially supported by the EU H2020 F-Interop and H2020 ARMOUR projects, Inria through the DIVERSITY associate team, the Spanish Ministry of Economy and the FEDER regional development fund through the SINERGIA project (TEC2015-71303-R), and the Center for Cyber-Physical Systems and the Internet of Things at the University of Southern California.

References

[1] M. R. Palattella *et al.*, "Standardized Protocol Stack for the Internet of (Important) Things," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 3, Mar. 2013, pp. 1389–1406.

[2] Q. Wang and X. Vilajosana, "6top Protocol (6P)," 31 Oct. 2016, IETF, draft-ietf-6tisch-6top-protocol-03, work in progress.

[3] J. Hui and P. Thubert, "Compression Format for IPv6 Datagrams over IEEE 802.15.4-Based Networks,"Sept. 2011, IETF RFC 6282.

[4] T. Winter *et al.*, "RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks," Mar. 2012. IETF RFC 6550.

[5] Z. Shelby, K. Hartke, and C. Bormann, "The Constrained Application Protocol (CoAP)," June 2014, IETF RFC 7252.

[6] T. Watteyne *et al.*, "Industrial Wireless IP-Based Cyber-Physical Systems," *Proc. IEEE*, vol. 104, no. 5, May 2016, pp. 1025–38.

[7] T. Watteyne *et al.*, OpenWSN: A Standards-Based Low-Power Wireless Development Environment. *Trans. Emerging Telecommun. Technologies*, vol. 23, no. 5, 2012, pp. 480–93.

[8] A. Gluhak *et al.*, "A Survey on Facilities for Experimental Internet of Things Research," *IEEE Commun. Mag.*, vol. 49, no. 11, Nov. 2011, pp. 58–67.

[9] T. Watteyne, A. Mehta, and K. Pister, "Reliability Through Frequency Diversity: Why Channel Hopping Makes Sense," *Int'l. Conf. Modeling, Analysis and Simulation of Wireless and Mobile Systems*, Tenerife, Canary Islands, Spain, 26–30, Oct. 2009, pp. 116–23.

[10] X. Vilajosana *et al.*, A Realistic Energy Consumption Model for TSCH Networks," *IEEE Sensors J.*, vol. 14, no. 2, 2014, pp. 482–89.

[11] D. Stanislowski *et al.*, "Adaptive Synchronization in IEEE802.15.4e Networks," *IEEE Trans. Industrial Informatics*, vol. 10, no. 1, 2014, pp. 795–802.

[12] T. Watteyne *et al.*, "PEACH: Predicting Frost Events in Peach Orchards Using IoT Technology," *EAI Endorsed Trans. Internet of Things*, vol. 16, no. 5, Dec. 2016.

[13] Z. Zhang *et al.*, "Long-Term Monitoring of the Sierra Nevada Snowpack Using Wireless Sensor Networks," *IEEE Internet of Things J.*, Dec. 2016.

Biographies

Thomas Watteyne (thomas.watteyne@inria.fr) is a researcher at Inria, Paris, and a senior networking design engineer in the Dust Networks product group at Linear Technology in Silicon Valley. He co-chairs the IETF 6TiSCH Working Group. He did his postdoctoral work in Prof. Kristofer Pister's lab at UC Berkeley. He founded and co-leads Berkeley's OpenWSN project. He holds a Ph.D. in computer science (2008), an M.Sc. in networking (2005), and an M.Eng. in telecommunications (2005) from INSA Lyon, France.

Pere Tuset-Peiró (peretuset@uoc.edu) is an assistant professor at the Computer Science, Multimedia and Telecommunications Department of Universitat Oberta de Catalunya (UOC), Barcelona, Spain. He received his B.Sc. and M.Sc. in telecommunications engineering from Universitat Politècnica de Catalunya (UPC) in 2007 and 2011, respectively, and his Ph.D. in network and information technologies from UOC in 2015. Currently, he has published over 20 journal and conference publications, and holds 7 international patents related to embedded systems, wireless communications, and signal processing.

Xavier Vilajosana (xvilajosana@uoc.edu) is a principal investigator of the Wireless Networks Research Lab at UOC. He is also a co-founder of Worldsensing and OpenMote Technologies. He is an active member of the IETF 6TiSCH WG, where he has authored different standard proposals. He also holds 30 patents. He has been a visiting professor at Prof. Pister's UC Berkeley lab. He co-leads Berkeley's OpenWSN project. He has been a senior researcher at the HP R&D labs (2014–2016) and a visiting researcher at the France Telecom R&D Labs Paris (2008). He holds a Ph.D. (2009) from UOC, and M.Sc. and M.Eng. degrees (2004) from UPC.

Sofie Pollin [SM] (sofie.pollin@esat.kuleuven.be) obtained her Ph.D. degree at KU Leuven, Belgium, 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. Her research centers around networked systems which require networks that are ever more dense, heterogeneous, battery powered, and spectrum constrained. She is a BAEF and Marie Curie fellow.

Bhaskar Krishnamachari [M02, SM14] (bkrishna@usc.edu) received his B.E. degree in electrical engineering at The Cooper Union, New York, New York, in 1998, and his M.S. and Ph.D. degrees from Cornell University, Ithaca, New York, in 1999 and 2002, respectively. He is currently a professor and Ming Hsieh Faculty Fellow in the Department of Electrical Engineering at the Viterbi School of Engineering, University of Southern California, Los Angeles. He is also the director of the USC Viterbi School of Engineering Center on Cyber-Physical Systems and the Internet of Things. His primary research interest is in the design and analysis of algorithms, protocols, and applications for next-generation wireless networks.

The 6TiSCH stack is widely seen as the key enabler for the IIoT, as it combines the performance of proven industrial low-power wireless communication technologies with the ease of use of an IPv6-enabled stack. A vibrant community has built around it, making a plethora of tools and resources available to the instructors.

[17] http://www.dustcloud.org/

Hands-On Education about Standardization: Is That What Industry Expects?

Damjan Katusic, Pavle Skocir, Mario Kusek, Gordan Jezic, Carlo Ratti, and Iva Bojic

The university curricula rarely include standards analysis and rarely encourage students to create solutions that are based on the standards. This fact poses a hindrance in the adoption of new technologies due to the fact that recent graduates as young professionals will most certainly have difficulties in using the new standards they may get in touch with on their first job.

ABSTRACT

Technical standards support compatibility and interoperability among numerous entities existing in the telecommunications market. However, the university curricula rarely include standards analysis and rarely encourage students to create solutions that are based on the standards. This fact poses a hindrance in the adoption of new technologies due to the fact that recent graduates as young professionals will most certainly have difficulties in using new standards they may get in touch with on their first job. In our approach, we acknowledge the need for students to work with standards before they graduate. Research activities, connected with industrial projects in which application of new standardized technologies is in focus, are combined with student projects, enabling students to get their own hands-on experience in analyzing and applying telecommunications standards. As the result, students gain valuable experience for their future careers, an academic setting is used as a *safe* environment for testing and possibly even enhancing new standards, and the industry attains deeper knowledge about standards and technologies that could be used in their real-world deployments.

INTRODUCTION

The telecommunications sector is guided by rapid innovation, bringing new technological marvels at an unprecedented rate. Every year we are witnessing the deployment of not just upgraded versions of existing, but also completely new technologies. In the area of telecommunications, this roughly refers to new networking equipment building up network infrastructure, new protocols enabling their operation and communication, and new devices serving as platforms for innovative services. However, widespread use of technology, old or new, cannot be achieved without at least partial standardization.

Technical standards enable compatibility and interoperability between numerous devices in the telecommunications market. They allow firms to focus on specific individual components within the telecommunications system, which they do not need to develop themselves entirely, as their products are now compatible with other firms' products. Consumers gain more certainty in products they are offered on the market, and more choice because standards increase competition between firms. Development of devices connecting to the existing network infrastructure, offering services, and communicating with the help of various protocols would be next to impossible without standardization efforts building common ground for particular technologies to be used.

Competing incompatible technologies can bring a specific marketplace to a halt, or even irretrievably hinder it. Obviously, a completely new technology cannot be standardized before it is invented, but as soon as it gains some importance on the market, some form of standardization is inevitable. However, if defined and used rigidly, standards also have a tendency to restrict new innovation. Therefore, it is important for telecommunications professionals to understand the importance of standards in technology, and if necessary how to read, implement, or even define them in practice.

Education should prepare students for their future careers. The telecommunications sector, as are science, technology, engineering, mathematics, and computer science (STEM-C) fields in general, is rapidly changing, which brings many challenges to the educational process [1, 2]. Higher education institutions need to continuously adapt to society's changing demands and new scientific achievements. Telecommunications education is not only focused on learning and mastering science and the technical skill set of future engineers, but also basics of economics, management, and quality assurance.

Technical standards, apart from the very latest still in development, are subject to moderate change and, in a very limited scope, are already integrated into faculties' curricula. However, more important than bits and pieces of a particular standard's content is the notion of their importance, as well as the ability of future professionals to read them and, if necessary, implement correctly in practice. This segment of professional development is largely absent from courses students can attend in telecommunications or similar study programs.

Consequently, when one day on their first job, given the task to carefully analyze a particular standard describing, for example, the newest version of a communication protocol or a network device, which is not yet available as a polished and tested software module ready for integration into their firm's product, most recent

Digital Object Identifier:
10.1109/MCOM.2017.1601134

Damjan Katusic, Pavle Skocir, Mario Kusek, and Gordan Jezic are with the University of Zagreb; Carlo Ratti is with Massachusetts Institute of Technology; Iva Bojic is with Massachusetts Institute of Technology and SMART Centre.

graduates would not know what to do and how to start the given task. This is a likely outcome of the current educational approach due to the fact that a particular student can graduate with an M.Sc. in a telecommunications related field without once being engaged in a practical exercise regarding technical standards. This could be potentially avoided if the students, in addition to learning about standards by listening to lectures, which are part of their regular academic curriculum, were presented with opportunities to get hands-on experience implementing standards by working on team-based student projects.

The Latest Trends in Telecommunications Standards

We are witnesses of how the complexity and the speed of technological development constantly increases, and different markets, especially those in the field of the telecommunications, show evidence of a higher variety of products and solutions in a more frequent manner. The need for technological standardization is growing as successful standards allow a large number of market players to interoperate and innovate. Consequently, this helps commercial firms to expand, technology to progress faster, and end users to enjoy a wider choice range, richer functionality, and lower costs.

Standardization efforts are of vital importance because they can make a significant difference in the success of an innovation by creating a shared framework for innovation and establishing the rules of the game [3]. Moreover, it has been proven that the structures of convergence in technological development and standardization show a moderate positive correlation [4]. Finally, by providing a minimum level of quality, the reduction of variety, and supporting interoperability or compatibility between different parts, standards also show economic benefits [5].

In the past only standardization organizations played a major role in the telecommunications industry to guarantee interoperability between products and to allow different voices to be heard when it comes to expressing their opinion on the future directions the industry should take. However, the standardization landscape has changed over the past 20 years, and now in addition to *formal standard*s (i.e., de jure standards) released by standardization organizations, there are also de facto standards, which can be *proprietary* standards developed by a single firm or *informal standards* as developed in standards consortia.

Standards consortia are private industry alliances who share the same interest in developing technologies for standardization, and unlike formal standardization organizations, they are more flexible and able to set standards right on time [6]. Although standards consortia lack formal accreditation, their standards can still be widely accepted and of great importance, or even follow up certain formal standards [7]. Finally, the activities of formal and informal standardization bodies are complementary rather than substitutive, meaning that most technical issues are addressed by both formal standardization organizations and standards consortia [8].

The Current State in Standards Education and the Road Ahead

On their first job, recent graduates can come in contact with standardization and standards management in a variety of ways. On one hand, firms that are strategically active in standardization expect entry-level staff to have extensive knowledge in this area, and on the other hand firms that are not involved in standards and standardization have much lower expectations, as they themselves are not yet aware of how standardization can help them in achieving their business goals [9]. However, even in the latter case, when recent graduates are not expected to have a certain level of knowledge on standards and standardization, their knowledge can, for example, enable them to set up a new system within a firm.

The need for standardization education as part of their academic curriculum has been previously identified for students across different disciplines — technical sciences, business science, science of public administration, economics, and law [10]. The gap between manifest and latent needs for standardization education can be bridged by a strong national policy, which may be a part of a regional policy, and by cooperation between government, industry, national standards bodies, academia, and other educational institutions [11]. There are lessons to be learned from some Asian countries — in particular from Indonesia, where in 2010 23 universities cooperated with their national standards body, and South Korea, where standardization education is implemented in particular universities and secondary schools [12].

When it comes to academia, one of the problems is that only a limited number of university professors pay attention to standardization in their education and research activities. The second one is that even in cases when there is an elective course focused on standardization, it is often the case that it may not appeal to students. We propose that the solution for this problem may be including it in other courses or as a compulsory part of the curriculum at three different levels.

The first level would include learning only basic information about standardization. Students are, for example, presented with links to standard documents as part of their regular lectures. In this way students become aware of the fact that most of the materials they use when learning about telecommunications technologies are actually based on standards.

The second level would feature specialized lecture(s) for more detailed introduction into standardization issues. Students in these lectures find out more details about not only standards, but also standardization organizations and their workflows (i.e., specific work groups responsible for defining standards and issuing reports).

Finally, the third level would consist of a practical task (as part of a homework or laboratory assignment) of reading the actual technical standards with practical implementation as the final goal. Thus, students become familiar with these documents, read them for the first time in their professional careers, and then proceed to practical implementation in defined programming lan-

When it comes to academia, one of the problems is that only a limited number of university professors pay attention to standardization in their education and research activities. The second one is that even in cases when there is an elective course focused on standardization, it is often the case that it may not appeal to students.

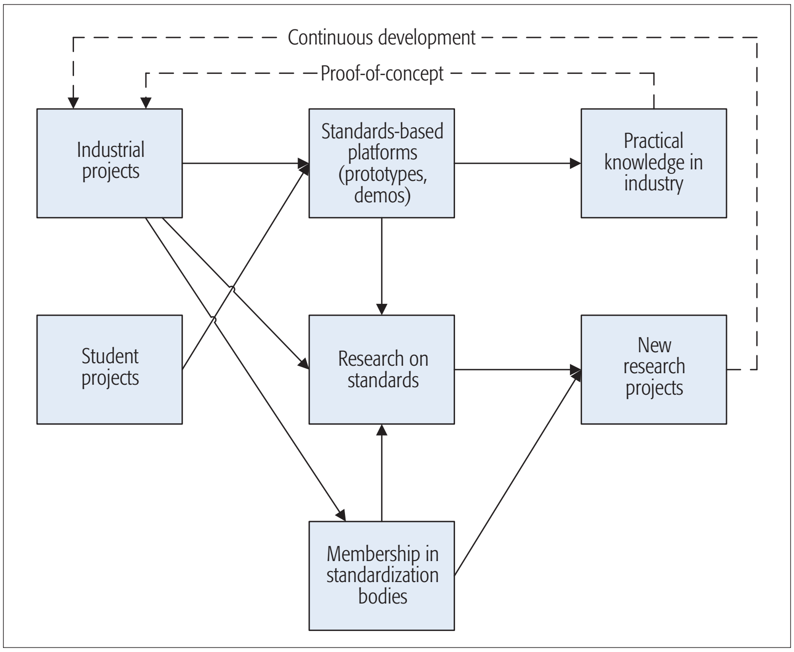


Figure 1. The proposed framework for enabling students to get hands-on education on standardization.

guage (if applicable) or write an essay outlining the most important characteristics of the technology specified in the chosen standard.

However, in addition to professors' willingness to include the topic in their courses, the problem is the curricula already being full, resulting in competition of a standardization topic with other more established ones. In order to overcome that, we propose a framework presented in Fig. 1. Specifically, at many universities, the teaching staff is involved in different types of activities. Primarily, there are educational activities, mainly in the form of lectures, but also often in the form of designing and overseeing team-based student projects. These are often without a strictly defined syllabus, as the content of student projects is specified in agreement with their mentors.

Apart from educational activities, the staff is also usually involved in research activities and projects done in collaboration with the industry. We thus propose to combine student teams with research teams composed of research assistants employed at the university to develop standards-based platforms, as seen in Fig. 1. As a result, student teams gain valuable insight into new technologies based on standards, and by working together, research and student teams can take on larger amounts of work. Students can then afterward use the knowledge gained from deploying standards-based platforms within a *safe* university environment as practical knowledge in the industry. Moreover, to gain insights into the process of defining standards and choosing the areas for which standards should be defined, it is strongly encouraged that the university staff signs up for membership in standardization bodies with significant impact on the area of their interest. The proposed process is iterative, and the results of ongoing industrial, student, and research projects serve as input for new industrial, student, and research projects.

Good Practices of Standards Education at the University of Zagreb

Using the methodology proposed in the previous section, we examined Telecommunications and Informatics profiles that are offered at the University of Zagreb Faculty of Electrical Engineering and Computing (UNIZG-FER) in Croatia. Education at UNIZG-FER is divided into three levels: Bachelor, Master, and postgraduate. Courses are grouped into categories: Mandatory, Theoretical, Specialization, Elective, Mathematics and Science, Humanistic or Social, and Skills courses. Our analysis of telecommunications studies only involved the first four categories, as shown in Table 1. Students must enroll in all mandatory courses, but only a finite number of courses from other categories, as recommended by maximum European Credit Transfer System (ECTS) score for each category. In practice, this would roughly map into 30 ECTS points per student per semester.

Learning About Standardization in Different Courses

We have analyzed the published curricula on the official websites of all the courses taught as part of Telecommunications and Informatics profiles in the fifth and sixth semesters of the Bachelor's study program, and all four semesters of the Master's study program. The results may slightly vary year after year as some changes to homework's or laboratory assignments' content is expected. We have also interviewed several colleagues from the Department of Telecommunications and Department of Wireless Communications. Table 1 summarizes our conclusions. Several courses in the curriculum included instruction on standards, especially in those dealing with telecommunications and networking, although not explicitly stating such material in the courses' syllabus.

In the first level of standards education, materials explaining preeminent communication protocols, for example, did not present published technical standards, but abbreviated overviews instead. Notably, links to full texts of the technical standards were provided in lectures' presentations. However, our concern is that limited academic exposure to standards inadequately prepares students to use the protocols professionally. In the case of Bachelor studies, there are 50 percent of courses with the basic standards education in the category of mandatory courses, 25 percent of specialization courses, and 33 percent of elective courses. At the Master's level, the situation is the following: 17 percent of courses in the category of mandatory courses, 63 percent of theoretical, 42 percent of specialization, and 31 percent of elective courses offer basic standardization education.

Several courses in second-level standards education commit at least one lecture explicitly to the overview of technical standards. In the case of Bachelor studies, there are two courses with standards lectures as defined in syllabus in the category of mandatory courses, none in the category of specialization courses, and one elective course. A similar situation is at the Master's degree, with none of the courses in the category of mandatory courses, but 25 percent of theoretical, 21 percent of specialization, and 19 percent of elective courses.

	Bachelor degree				Master degree					
	5. sem	6. sem	Total	Stand.	1. sem	2. sem	3. sem	4.sem	Total	Stan.
Mandatory courses	5	3	8	4/1/2*	2	2	1	1	6	1/0/3*
Theoretical courses	–	–	–	–	3	5	–	–	8	5/2/1
Specialization courses	2	2	4	1/1/0	4	11	9	–	24	10/5/3
Elective courses	–	12	12	4/2/1	–	–	16	–	16	5/3/1

Table 1. Analysis of standards education at UNIZG-FER.

Several years ago, we started to enhance standards education using the proposed framework by incorporating hands-on experience within five available seminar and project team-oriented courses. These courses are without a strictly defined syllabus, as the content of students' projects is specified in agreement with students' mentors.

Third-level course homework assignments required students to read, analyze, and implement a technical standard as a programming solution or author a report on the subject. Seminars and group-project-based courses, of which there are two Bachelor and three Master, required students to gain a working competency of standardization tasks. In the Bachelor's study program, there are only two mandatory, no specialization, and one elective course with such a high level of standards education. The situation is, as expected, a bit better in the Master's study program with three of the courses in the category of mandatory courses, one theoretical, three specialization, and one elective course requiring students to analyze and practically implement a technical standard. Importantly, students cannot attend all those courses because of limitations in ECTS points in each of the course categories. Moreover, standardization education in group-project-based courses is limited to a particular situation and depends principally on professors' willingness to include such topics. These courses are marked with asterisks in Table 1. Unfortunately, the curriculum allows students to graduate with an M.Sc. in information and communication technologies without once being engaged in a practical exercise regarding technical standards.

Hands-on Experience in Standardization Education

Several years ago, we started to enhance standards education using the proposed framework by incorporating hands-on experience within five available seminar and project-team-oriented courses. These courses (i.e., software design project and B.Sc. thesis at the Bachelor's degree, and seminar, project, and graduation thesis at the Master's degree) are without a strictly defined syllabus, as the content of student projects is specified in agreement with students' mentors. Apart from our educational duties, we also work on projects with our industrial partners, such as Ericsson Nikola Tesla (ETK) and the Croatian regulatory authority for network industries (HAKOM). These projects bring opportunities to explore the latest technologies and use them to develop our prototype solutions to the problems on which we focus.

The institution's students benefit from hands-on experience with state-of-the-art telecommunications technology by partnering with the industry. The aforementioned project-based courses with flexible curricula provide us with the best platform to achieve this goal. Students work on small-scale problems with the latest technologies, expanding our competencies in the field. This tactic proved useful to UNIZG-FER in applying for a research and innovation project symbIoTe, and in defining the system architecture. In the next sections, we describe three use cases in which teams of students worked on industry focused projects.

Case Study 1: Machine-to-Machine Platform with Semantic Support: The research focus of 2012–2013 was development of a standard-based machine-to-machine (M2M) system enabling communication between different types of devices [13, 14]. Our goal was the student-led development of a platform, based on European Telecommunications Standards Institute (ETSI) technical specifications, that supported semantic description of sensors, both physical and virtual ones. The performance of semantic repositories was meant to be analyzed.

The architecture of the system is shown in Fig. 2. It consists of virtual sensors (1) that simulate changes in the value of the sensor, and send it to the eHealth Semantic App started on the Cocoon Gateway (2), an implementation of an ETSI proposed specification for M2M communication. Physical sensors (3) are based on the Libelium Waspmote platform, and communicate with the gateway using the IEEE 802.15.4 standard. The semantic repository (4) is deployed on the Sesame Semantic Server, and stores data in an OWLIM-Lite semantic repository. Information from sensors is stored in resource description framework (RDF) form. Cloud storage (5) is implemented by using OpenNebula and Eucalyptus solutions. Web application executed in cloud was able to show information about a particular sensor.

During the project, students had to become familiar with the actual ETSI standard in order to connect physical and virtual sensors with the Cocoon Gateway, implemented according to the same standard. Additionally, students gained experience in using semantic ontologies to simplify configuration of M2M applications and represent information from heterogeneous data sources. Furthermore, students presented the prototype in a demo session at a scientific conference, shown in Fig. 3. Our industrial partner gained insights about the usability of the Cocoon Gateway: whether it is easy to configure, what its functionalities are, and so on.

Case Study 2: 6LowPAN on Waspmote Mote Runners: One student group studied how different types of low-power communication protocols affected network self-organization. The flexibility for self-organization was enabled by the Low-Energy Adaptive Clustering Hierarchy (LEACH) protocol, commonly used to distribute energy consumption of sensors in a network, allocating more power to specific nodes designated as cluster heads.

Apart from our educational duties, we also work on projects with our industrial partners, such as Ericsson Nikola Tesla (ETK) and the Croatian regulatory authority for network industries (HAKOM). These projects bring opportunities to explore the latest technologies and use them to develop our prototype solutions to the problems on which we focus.

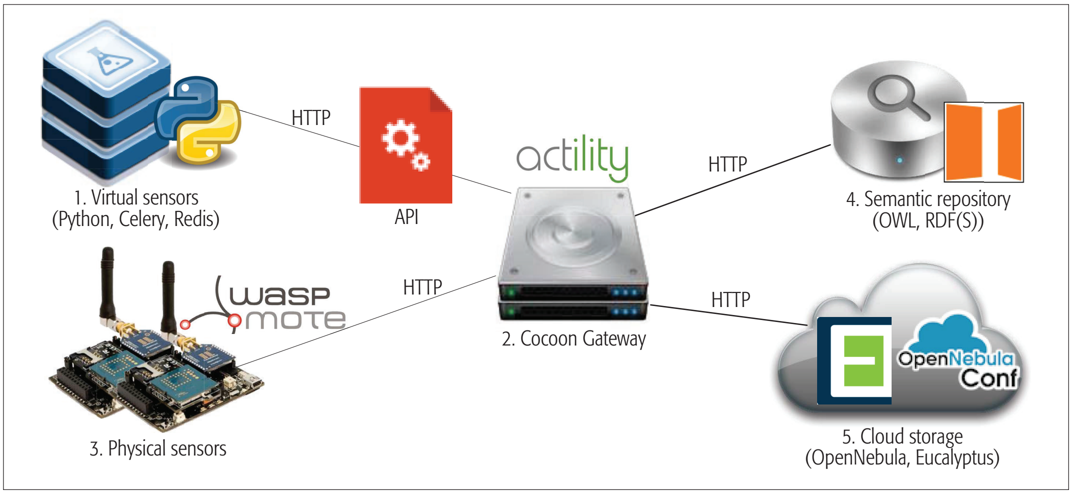


Figure 2. ETSI-based architecture of developed M2M system.

Figure 3. Demo at a scientific conference.

The goal of the research was to compare several low-power communication protocols by analyzing each node's energy consumption in a network organized into clusters using the LEACH protocol. The communication protocols surveyed included IPv6 over Low power Wireless Personal Area Networks (6LoWPAN), Bluetooth Low Energy (BLE), XBee, and Low Power Wide Area Network (LoRaWAN). The platform we used for the 6LoWPAN demonstration setup was a result of the partnership of two firms, one responsible for the hardware, and the other that developed the software. Unfortunately, the software had several bugs, and the company responsible for it dropped the development and support. Since there was no response to our direct inquiries aimed at solving the problem, there was no other option but to preclude the 6LoWPAN protocol from inclusion in the comparison.

This experience taught us and students participating in the project a very valuable lesson: very often problems may occur with new, unproven technologies. Sometimes the reason may be that there is no official support, or as in our case, the development and support for this particular item was abandoned. For our industrial partner, our experience was also very valuable, exposing that a particular product was not ready for market, allowing for improvements before deployment to real environments.

Case Study 3: User-Centric Solution for Controlling M2M Devices: Another topic for a student group was to analyze control protocols for M2M devices, enabling starting and stopping of tasks on M2M devices through mobile or web applications. On-demand tasking achieves data transmission and other functionalities at the expressed intent of the user. When paired with environmental sensors, such as motion sensors in a room, the proposed infrastructure conserves energy, initiating tasks based on need, for example, turning on the lights when someone enters the room.

The proposed architecture of the system, based on Message Queue Telemetry Transport (MQTT) protocol, is shown in Fig. 4. A request from the device management application begins communication between web server, gateway, and M2M devices. A similar solution (with the same network entities) was also implemented using Constrained Application Protocol (CoAP).

This study's principal communication protocols deployed to control M2M environments were MQTT and CoAP. The usage of both protocols will certainly grow in the nearer future, as M2M systems will become ubiquitous. This study equipped students with the technical knowledge and hands-on experience with both protocols they will almost certainly use in their future professional careers. As for our industrial partners, they gained insights into a prototype demonstrating what can be done by using protocols currently in the spotlight.

Discussion and Conclusion

Nowadays the benefits of standardization in the telecommunications industry are well known. This has resulted in not only formal standardization organizations, but also informal industry-driven standards consortia producing widely adopted and important standard solutions. Consequently, this increased our understanding about the importance of education on telecommunications standardization, which can be *formal* or *post-formal*, where the former is divided into four sub-categories: primary education, secondary education, and higher education composed of undergraduate and graduate education [15].

A very extensive analysis conducted on 16 general promotion, 10 primary/secondary education, 27 higher education, and 65 post-formal education programs from 21 countries of a regional

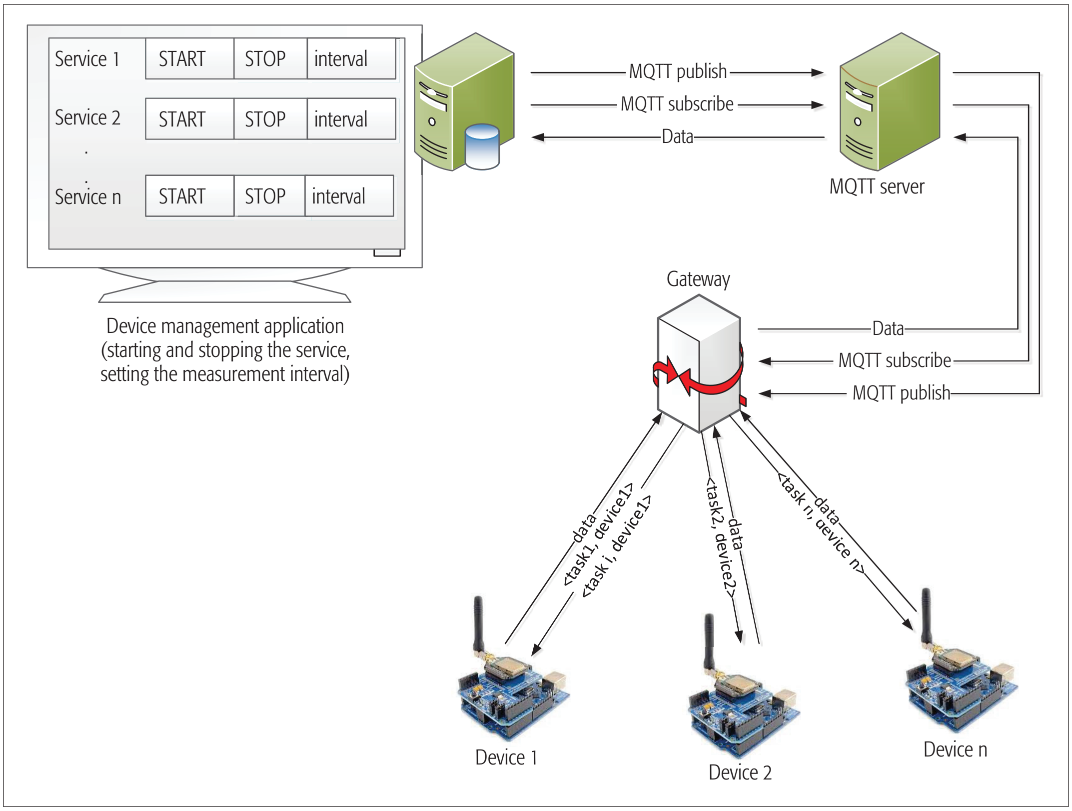


Figure 4. MQTT-based architecture of the developed M2M system.

Industry settings are much less flexible and the industry can benefit by learning from mistakes made in the academic environments, rather than spend much more money to make the same mistakes in the industry environments.

variety showed that education on standards mostly consisted of lectures and group contest activities [15]. However, when it comes to academic settings, adding anything to already jam-packed curricula is dismissed as virtually impossible by many faculty professors. Faced with those challenges, the solution we propose is to allow students to "get their hands dirty" by implementing certain standards during their project-based learning.

The lessons learned during the last five years of implementing our solution at UNIZG-FER are:
- An academic setting is a "safe" environment in which students can fail and most certainly will once when involved in working with new technologies as early adopters.
- This failure will allow them to be more confident in the future when faced with new problems.
- Industry settings are much less flexible, and the industry can benefit by learning from mistakes made in the academic environments, rather than spending much more money making the same mistakes in the industry environments.

One can say that the outcomes of an internship in a company are similar to those of our approach, which is really something that we want to achieve. However, there are a few substantial differences between the two of them:
- Proposed student projects are focused on new standards-based technologies, preferably as early adopters.
- Not all students have an opportunity to get an internship, whereas all students have to work on their student projects, making our approach more inclusive.
- In the proposed setting, students are supervised by university professors and work side by side with Ph.D. students and postdoctoral researchers, enabling them to learn basic skills of conducting research.
- By including students in work on projects that are of interest for a particular industry, we build stronger ties between academia and industry, enabling a transfer of knowledge.

Acknowledgment

The research was supported by Croatian Science Foundation under the projects 8065 (Human-centric Communications in Smart Networks) and 8813 (Managing Trust and Coordinating Interactions in Smart Networks of People, Machines and Organizations), and by the European Union's Horizon 2020 research and innovation programme under grant agreement No 688156 (project symbIoTe). We further thank the MIT SMART Program, Center for Complex Engineering Systems (CCES) at KACST and MIT, the National Science Foundation, the MIT Portugal Program, the AT&T Foundation, Audi Volkswagen, BBVA, the Coca Cola Company, Ericsson, Expo 2015, Ferrovial, GE and all the members of the MIT Senseable City Lab Consortium for supporting the research. Finally, the research was also partially supported by the National Research Foundation, Prime Minister's Office, Singapore, under its CREATE programme, Singapore-MIT Alliance for Research and Technology (SMART) Future Urban Mobility (FM) IRG. The authors would like to thank Vedan Podobnik, Miljenko Mikuc, Dragan Jevtic, Tomislav Kos, Aleksandar Antonic, Matija Sulc, and Hrvoje Maracic for

When it comes to the academic settings, adding anything to already jam-packed curriculum is dismissed as virtually impossible by many faculty professors. Faced with those challenges, the solution we propose is to allow students to get their "hands dirty" by implementing certain standards during their project-based learning.

fruitful discussions, encouragement, and help in preparing the manuscript, and Jessica Snyder for proofreading the final version of the manuscript.

References

[1] I. Bojic and J. F. Arratia, "Teaching K-12 Students STEM-C Related Topics through Playing and Conducting Research," *Proc. IEEE Frontiers in Education Conf.*, 2015, pp. 1–8.

[2] I. Bojic, T. Jagust, and A. Sovic, "Selected Examples of Cooperation between Universities and Schools in STEM Education," *Proc. IEEE Integrated STEM Education Conf.*, 2015, pp. 189–94.

[3] D. Datla, S. F. Bush, and S. Hasan, "Innovation and Standardization: Oxymoron or Pleonasm?," *IEEE JSAC*, vol. 33, no. 4, 2015, pp. 703–10.

[4] S. Gauch and K. Blind, "Technological Convergence and the Absorptive Capacity of Standardisation," *Technological Forecasting and Social Change*, vol. 91, 2015, pp. 236–49.

[5] L. Heuser, Z. Nochta, and N. Trunk, "ICT Shaping the World: A Scientific View," 2008.

[6] C. Cargill, "Intellectual Property Rights and Standards Setting Organizations: An Overview of Failed Evolution," report issued by the U.S. Department of Justice and the Federal Trade Commission, 2002.

[7] K. Blind, S. Gauch, and R. Hawkins, "How Stakeholders View the Impacts of International ICT Standards," *Telecommun. Policy*, vol. 34, no. 3, 2010, pp. 162–74.

[8] K. Blind and S. Gauch, "Trends in ICT Standards: The Relationship between European Standardisation Bodies and Standards Consortia," *Telecommun. Policy*, vol. 32, no. 7, 2008, pp. 503–13.

[9] I. A. Hövel and I. M. Schacht, "Standardization in Education — A Building Block for Professional Careers," *Proc. Int'l. Cooperation on Education about Standardization Conf.*, 2013, pp. 1–6.

[10] H. J. De Vries and T. M. Egyedi, "Standardization Education: Developments and Progress," *New Applications in IT Standards: Developments and Progress*, 2002, pp. 1–26.

[11] H. J. de Vries and T. M. Egyedi, "Education about Standardization: Recent Findings," *Int'l. J. IT Standards & Standardization Research*, vol. 5, no. 2, 2007, p. 1.

[12] H. J. de Vries, "How to Implement Standardization Education in a Country," *Modern Trends Surrounding Information Technology Standards and Standardization within Organizations*, 2014, pp. 262—90.

[13] V. Galetic *et al.*, "Basic Principles of Machine-to-Machine Communication and Its Impact on Telecommunications Industry," *Proc. 34th Int'l. Convention on Info. and Commun. Technology, Electronics and Microelectronics*, 2011, pp. 380–85.

[14] D. Katusic *et al.*, "Market, Standardization, and Regulation Development in Machine-to-Machine Communications," *Proc. 20th Int'l. Conf. Software, Telecommunications and Computer Networks*, 2012, pp. 1—7.

[15] D. G. Choi and H. J. de Vries, "Standardization as Emerging Content in Technology Education at All Levels of Education," *Int'l. J. Technology and Design Education*, vol. 21, no. 1, 2011, pp. 111–35.

Biographies

DAMJAN KATUSIC [S'11] is a Ph.D. student at the University of Zagreb, Faculty of Electrical Engineering and Computing, Internet of Things Laboratory (IoTLab@FER). His research interest include machine-to-machine communication, Internet of Things (IoT) systems, and complex services in sensor networks. He currently works on an industry collaboration project with HAKOM and serves as a teaching assistant in several courses on both the undergraduate and graduate levels.

PAVLE SKOCIR [S'12] is a research assistant and Ph.D. student at the University of Zagreb, Faculty of Electrical Engineering and Computing, Internet of Things Laboratory (IoTLab@FER). His research interests include energy efficiency of Internet of Things (IoT) systems with battery-powered IoT devices and achieving IoT platform interoperability. He is a member of the IEEE Communications and Computer Societies, and serves as a Chair of the IEEE Computer Society Chapter within IEEE Student Branch Zagreb.

MARIO KUSEK [M'97] is an associate professor at the University of Zagreb, Faculty of Electrical Engineering and Computing. His main research interests include distributed systems, software agents, and the Internet of Things. He has participated in scientific projects financed by Croatian Ministries, Croatian Science Foundation, EU COST actions, H2020 (EC), and others. He is Chair of the IEEE ComSoc Croatia Chapter and Conference Coordinator in the Croatia Section, and of the European Telecommunications Standards Institute (ETSI).

GORDAN JEZIC [M'97] is a professor at the University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Telecommunications. His research interests include parallel and distributed systems, M2M and IoT systems, communication networks and protocols, mobile software agents, and multi-agent systems. He is a member of KES International, IEEE FIPA, and the Technical Committee of IEEE SMC on Computational Collective Intelligence.

CARLO RATTI is a professor of the Practice of Urban Technologies at the Massachusetts Institute of Technology, where he directs the Senseable City Lab. He is also a founding partner of the international design and innovation office Carlo Ratti Associati. He graduated in engineering from the Politecnico di Torino and the École Nationale des Ponts et Chaussées in Paris, and later earned his M.Phil. and Ph.D. at the University of Cambridge, United Kingdom. He completed his Ph.D. at the Massachusetts Institute of Technology as a Fulbright Senior Scholar. He holds several patents and has co-authored over 250 publications, including the recent book *The City of Tomorrow* (Yale University Press, June 2016, with Matthew Claudel). In addition to scientific writings, he is a regular contributor to popular media including Project Syndicate, the *New York Times*, *Washington Post*, *Financial Times*, and *Scientific American*. His work has been exhibited worldwide at venues such as the Venice Biennale, the Design Museum Barcelona, the Science Museum in London, MAXXI in Rome, and MoMA in New York City.

IVA BOJIC [M'08] is a senior postdoctoral researcher at Singapore-MIT Alliance for Research and Technology, Singapore. She received her M.S. and Ph.D. degrees in computer science from the University of Zagreb Faculty of Electrical Engineering and Computing in 2009 and 2013, respectively. She also holds a second M.S. in mathematics from the University of Zagreb Faculty of Science. From 2014 to 2015 she was a Fulbright Scholar at Massachusetts Institute of Technology and from 2009 to 2014 she was with the University of Zagreb. Her doctoral thesis was awarded the Silver Plaque Josip Loncar ,and the Dean's award for outstanding doctoral dissertation and particularly successful scientific research. In 2012 she received the Google Anita Borg award from Google for strength of academic performance, leadership experience, and demonstrated passion for computer science. The same year she was also awarded the Richard E. Merwin award by IEEE for demonstrating outstanding involvement in IEEE and excellence in academic achievement. In the past, she served as an IEEE Student Ambassador for IEEE Region 8, a Chair of the IEEE Computer Society Student Chapter of Zagreb, Croatia, an IEEE Student representative, and a Chair of IEEE Educational Activities for Croatia Section.

Teaching Telecommunication Standards: Bridging the Gap between Theory and Practice

Antoni Gelonch-Bosch, Vuk Marojevic, and Ismael Gomez

Abstract

Telecommunication standards have become a reliable mechanism to strengthen collaboration between industry and research institutions to accelerate the evolution of communications systems. Standards are needed to enable cooperation while promoting competition. Within the framework of a standard, the companies involved in the standardization process contribute and agree on appropriate technical specifications to ensure diversity and compatibility, and facilitate worldwide commercial deployment and evolution. Those parts of the system that can create competitive advantages are intentionally left open in the specifications. Such specifications are extensive, complex, and minimalistic. This makes telecommunication standards education a difficult endeavor, but it is much demanded by industry and governments to spur economic growth. This article describes a methodology for teaching wireless communications standards. We define our methodology around six learning stages that assimilate the standardization process and identify key learning objectives for each. Enabled by software-defined radio technology, we describe a practical learning environment that facilitates developing many of the needed technical and soft skills without the inherent difficulty and cost associated with radio frequency components and regulation. Using only open source software and commercial of-the-shelf computers, this environment is portable and can easily be recreated at other educational institutions and adapted to their educational needs and constraints. We discuss our and our students' experiences when employing the proposed methodology to 4G LTE standard education at Barcelona Tech.

Introduction

Standards are fundamental for the development of products in many technical areas. Standardization tackles real problems and defines the requirements of a technological ecosystem where a diverse set of players can effectively pursue their business objectives. Any company developing a method, process, service, or device in compliance with a standard needs to pass the homologation process, which consists of a series of tests that are defined in the standard. Thus, this regulated interaction is the cornerstone that holds the ecosystem and allows interactions (compatibility, interoperability) among the stakeholders (manufacturers, service providers, etc.).

Standards have become a catalyst for technological innovation in numerous areas of science and technology because of the way standards are defined, leaving room for innovation and market differentiation [1]. Standards become a tool to coordinate efforts of various stakeholders while preserving competition. Involved companies can take benefits of economies of scale, build or strengthen collaborations, and participate according to their business model and capability.

The potential of standards to spur economy and impact society is apparent more than ever in the increasingly globalized world. Standards developed by the Third Generation Partnership Project (3GPP), a consortium of several standard setting organizations (SSOs) that standardizes cellular communications, led to an estimated global revenue of more than US$3.3 trillion in benefits and more than 11 million jobs in 2014 [2]. A billion human users enjoy wireless communications services today, and multiple billions of machines will be connected very soon.

Problem Formulation

Cellular communications are evolving toward the fifth generation (5G). Five revisions of the 4G Long Term Evolution (LTE) have been released to date, while in parallel, IEEE and other standardization bodies have evolved their WLAN or IEEE 802.xx series products, with a different mobile system profile. Many jobs in the wireless communications industry require telecommunication standards education. Implementing or evolving a complex standard such as LTE is challenging for anyone, but can be overwhelming for fresh graduates. The 3GPP specifications are written in an unusual language, are often intricate, and refer to other documents, requiring a steep learning curve. The technical reasons for specifying one parameter or technique over another are difficult to understand and often have historical, political, or economical foundations. Moreover, typical parameter values that can be useful for implementing an algorithm are extremely difficult to find in the specifications. Despite the minimalistic and formal description, standards have been developed with implementation in mind.

The authors describe a methodology for teaching wireless communications standards. They define their methodology around six learning stages that assimilate the standardization process and identify key learning objectives for each. Enabled by software-defined radio technology, they describe a practical learning environment that facilitates developing many of the needed technical and soft skills without the inherent difficulty and cost associated with radio frequency components and regulation.

Antoni Gelonch-Bosch is with Universitat Politècnica de Catalunya; Vuk Marojevic is with Virginia Tech; Ismael Gomez is with Software Radio Systems.

Digital Object Identifier: 10.1109/MCOM.2017.1601231

0163-6804/17/$25.00 © 2017 IEEE

Teaching wireless communications standards is a challenging objective. Important efforts are therefore being made by the IEEE Standards Association and others to show the importance of standards and the role that standardization plays for the industry and society.

Recent graduates are highly motivated and have strong theoretical background in many aspects of telecommunication systems, and may have basic familiarity with modern standards. The skills that are needed to implement a standard-compliant communications system are rare to find. Even after completing a Ph.D. in electrical engineering, graduates often lack implementation skills such as advanced programming or understanding the limitations and constraints of real systems. At the university, a student learns how to solve a particular problem, analyze the available solutions, and develop alternative approaches. However, until actually implementing an algorithm and facing the practical challenges in terms of complexity and performance, the student does not fully understand the true differences and practical implications of selecting one algorithm for a standard over another. Therefore, standard-specific implementation, compliance, and performance assessment should be components of the electrical engineering curriculum.

The Proposed Approach and Related Work

Teaching wireless communications standards is a challenging objective. Important efforts are therefore being made by the IEEE Standards Association (http://standards.ieee.org/about/stdsedu/index.html) and others to show the importance of standards and the role that standardization plays for the industry and society. Through the IEEE Standards Education program, IEEE creates and distributes a variety of educational material and actively promotes the integration of standards into academic programs. They understand that standards are a tool that allows transitioning from theoretical, simulation, and experimental results to real-world implementations. Standards combine fundamental concepts with system implementation and address conformance, interoperability, operation, and management tasks. We argue that the reasons behind the technical choices, and their impact on resources and performance vs. flexibility trade-offs are important components of telecommunication standards education. Moreover, project management, teamwork, and the development of realistic expectations and practical solutions to imminent problems are skills that are demanded by the industry in addition to the domain-specific technical background. We therefore propose a methodology that allows developing such skills.

The combination of lecture-centered educational methodologies [3] with laboratory-centered approaches [4, 5] has been adopted in the engineering curriculum with special emphasis when the conceive, design, implement, and operate (CDIO) methodology appeared in the last decade. CDIO defines a structured methodology to translate the expected education outcomes to the curriculum [6–8]. While lecture-centered education is considered one of the most effective learning methods [3], it is often criticized for not helping students to transform their knowledge into skills. Laboratory work enhances student skills and helps to consolidate the acquired knowledge. Other cognitive techniques that help to address the development of the much needed skills include the *scaffolding* approach, where students receive some support from the instructor, who incrementally reduces this support when no longer needed, the *collaborative learning* approach, where the collaborative process gives students the possibility of sharing thoughts and approaching a valid solution [9], or the *student-centered learning* approach, which provides support that attends to specific student needs [4].

Considering the nature of standards and responding to industry needs, implementation-orientated active learning methods, such as project-based learning (PBL, http://www.bie.org), provide a student-centered learning environment that is appropriate for the purpose. Learning by doing has been a major engineering education breakthrough, inspired by how humans learn, how they develop expertise, and what mechanisms they activate when thinking at a higher level [10]. PBL also has its drawbacks. Students typically experience difficulties in initiating their project and do not reach the necessary depth when they lack sufficient background knowledge [11]. An interesting proposal to overcome this issue is the spiral step-by-step method [12], where information is grouped into stages and provided sequentially so that students can better focus and develop the necessary background with sufficient depth.

We propose a PBL methodology for teaching telecommunication standards. We applied it for teaching 3GPP LTE beyond the basics by making use of free, open source software-defined radio (SDR) development tools. The presented methodology was applied to the Wireless Communications master course taught at the Castelldefels School of Telecommunications and Aerospace Engineering (EETAC) of Barcelona Tech. Along with the methodology and case study, we describe our experiences and observations while teaching this course. The methodology and SDR framework are portable and allow adaptation to different learning environments and learning objectives.

Enabling Technologies

A wireless communications standard defines the physical and logical components of the system, the processes, and performance requirements. The functionalities are split into basic functions that are formally presented in the specifications only once, following the established document organization and indexing. These functions comprise algorithms, often expressed as one or more mathematical operations or one or more tables, and interact with other functions through well defined interfaces to provide the desired functionality.

SDR technology and the availability of open source software libraries for several digital signal processing (DSP) functions allow implementing complete radio systems in a few laboratory sessions. Software libraries exist for implementing wireless communications standards, such as openBTS implementing the Global System for Mobile Communications (GSM), and Amarisoft, OpenAirInterface, and srsLTE for LTE. These software libraries help to experiment with these systems at low cost.

SDR technology intrinsically supports hands-on learning, facilitating system implementation and practical analysis. We therefore advocate

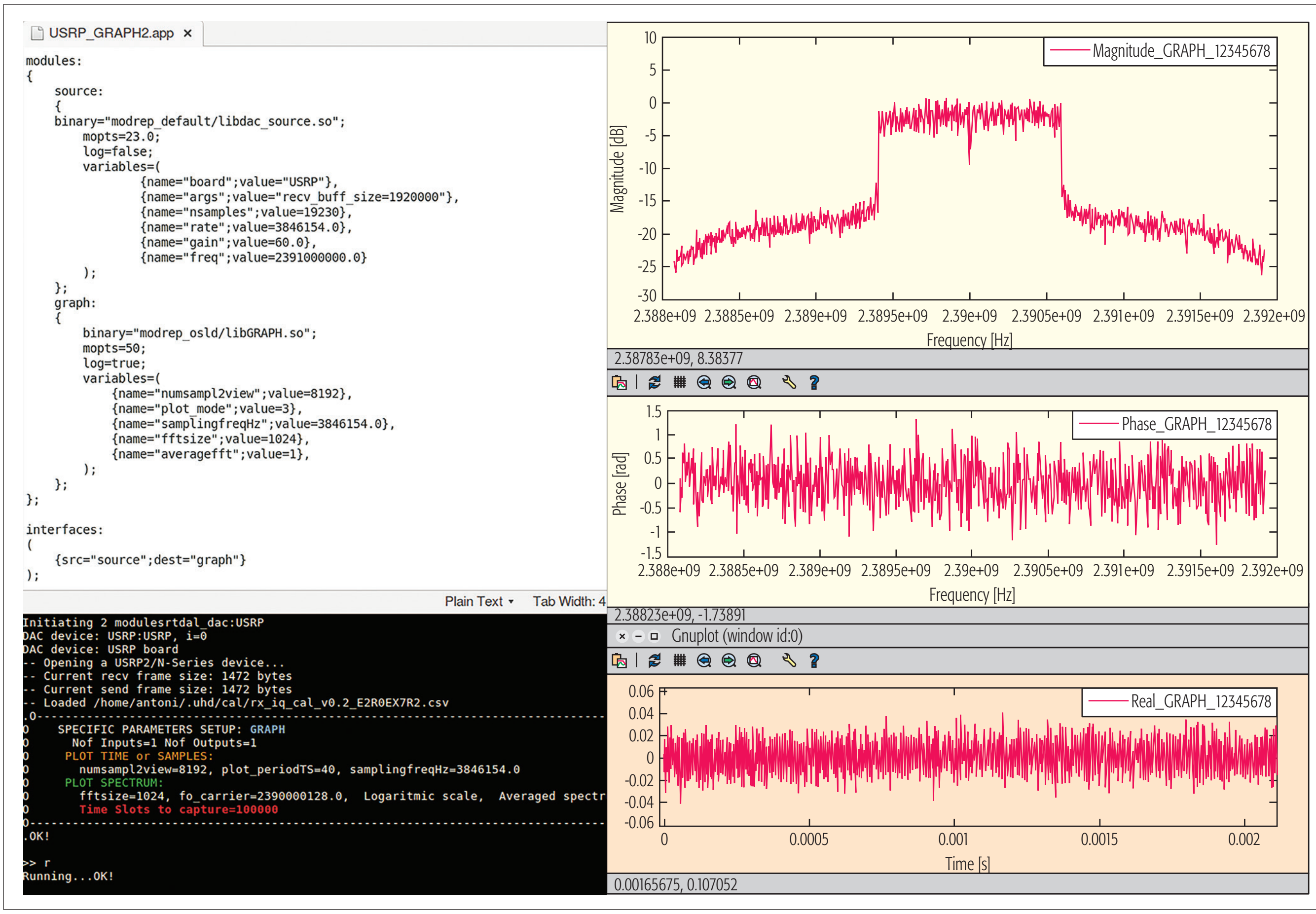


Figure 1. ALOE working environment. The upper left window shows an .app file that defines the modules involved, their configuration parameters, and the interfaces between them. The lower left window provides execution control and system status information. The graphs show a 1.4 MHz LTE spectrum and signal. The LTE signal was generated and captured using a pair of universal software radio peripherals (www.ettus.com).

for using SDR tools to implement, validate, and evaluate the performance of a wireless communications standard. SDR development and implementation frameworks, such as the software communications architecture (SCA), primarily used in military radios [13]; GNU Radio, primarily used in research and education (http://gnuradio.org); and the application layer and operating environment (ALOE), also used in research and education (http://flexnets.upc.edu), have certain features in common with the specifications of wireless communications standards. SDR frameworks use modular programming and support the concatenation of modules and access to external equipment through common interfaces.

ALOE is an open source SDR framework that is specifically designed for the implementation of modern radio systems [14]. It takes advantage of the regular data flow of DSP chains and provides a limited set of customizable services. The framework abstracts and virtualizes heterogeneous multiprocessor platforms, provides a packet-oriented network with first-in first-out (FIFO)-based interfaces between processors, and coordinates the real-time execution of the entire system. ALOE dynamically monitors the computing cost for every processing module and allows observing other critical system parameters in real time as well. Figure 1 provides a screenshot of the working environment of ALOE for a specific experiment. We can see the description of the processing chain in an .app file, a terminal for control of execution, and some graphs for visualizing signals in different formats. SDR boards or other peripherals can be interfaced through specific modules that use the vendors' application programming interfaces (APIs). Switching from a simulated channel to over-the-air transmission or reception then involves modifying the .app's sink or source module.

Teaching Telecommunication Standards: Methodology and Case Study

Instead of reverse-engineering the standard, building the standard out of fundamental building blocks, or functions, aligns with the human learning process. Many basic functions are introduced in prior undergraduate and graduate classes. Here the student can focus on learning how to use these functions in concert and combine them into larger functionalities to achieve the desired system behavior. The assembly of function and the analysis of how these functions work together and how they affect the subsystem or system performance allows invaluable insights to be gained into the specifications and reinforces the practice-oriented learning process.

We provide students with a preliminary system implementation and define assignments that lead

We provide students with a preliminary system implementation and define assignments that lead to gradually building and testing part of a standardized communications system. The students used ALOE to implement, validate, and evaluate the performance of the adopted solution with respect to its complexity.

to gradually building and testing part of a standardized communications system. The students used ALOE to implement, validate, and evaluate the performance of the adopted solution with respect to its complexity. Students obtain grades from measurable system performance results. The motivation for students to make the system work helps them to acquire a solid background of the technical details of the standard.

Building a prototype that follows an industry standard gives significance to the implementation profile of standards and covers many of the expected skills. Having successfully completed this course, the student will be able to:

- Read a telecommunication standard and find the desired information
- Design and implement a telecommunication system that is standard-compliant
- Discuss the pros and cons of alternative technical solutions
- Discuss possibly evolutionary paths for the standard being analyzed

The class is divided into groups of five or six students. Inspired by the scaffolding and spiral step-by-step educational methodologies, after providing a high-level overview of the standard under study in the first quarter of the class, we narrow down the focus. More precisely, the students implement part of the system and test its proper functionality, performance, and standard compliance based on previously defined metrics (about two quarters of the class period). Finally, the students discuss the technical decisions that were made during standardization and identify alternative solutions or improvements (last quarter).

The development and testing, being the main part of the class, are continuously monitored by the professor during weekly sessions (two to three hours), where students describe their progress and the troubles encountered, followed by discussions about the solutions adopted and the progress along the roadmap.

The proposed methodology balances the teaching material and assignments to fit the schedule and accommodate the specific learning objectives that the instructor considers of highest relevance. We propose six learning stages to guide the students through their projects, grouped into the modeling (I), development (II), and evaluation and review (III) phases. These are summarized in Table 1 and discussed in continuation. We provide a brief description of the methodology and exemplify it using a real case study.

The results presented in our case study are extracted from the documentation delivered by the project teams. This article discusses the project "Study of the computing cost of the LTE PHY," carried out by multiple student groups in 2012–2014. Starting from a baseline implementation, the project objective was to implement the missing pieces of the physical downlink shared channel (PDSCH), the data channel of LTE, and analyzing the impact of adaptive modulation and coding and the decoder on the system performance, but also on the computing demand.

LTE defines about 30 modulation and coding schemes (MCSs) and employs turbo coding and decoding. The LTE base station, or eNodeB, assigns the mobile terminal, or user equipment (UE), the highest possible MCS according to the channel conditions reported by the UE. Changes in the MCS are reported to the UE receiver as part of the control signaling. The UE decodes the control messages first and accordingly modifies the operational parameters of the receiver processing chain. We provided a simplified LTE PHY processing chain through ALOE [14], which features the eNodeB transmitter and UE receiver and a simulated channel. The students download and install ALOE on their computers and do not need any additional hardware.

General Overview of the Standard

Methodology: The student needs to get familiar with the standard and the standardization mechanics. We therefore provide:

- A high-level description of the standard, from a general description to some details, describing theoretical concepts and employed technologies, identifying relevant working parameters and expected behaviors
- An overview of the standard specifications and the relationship among the main and auxiliary documents

According to our working approach and temporal restrictions, we suggest providing tutorials in no more than two or three lecture periods. These tutorials should also cover the SDR framework or tools that the students will use in continuation of the course.

Case Study: The instructor provides LTE tutorials that cover the following topics:

- Overall LTE architecture description and functional split
- Radio protocol architecture: a description of functionalities of user plane and control plane signaling
- Fundamental resources, timing, multiuser access, and scheduling
- LTE PHY: Logical and physical channels and mapping to physical resources, synchronization process, retransmission protocol, and so forth
- System performance metrics
- Conformance test and RF regulation
- Organization of LTE specifications with focus on the physical layer (PHY)

The LTE tutorial includes a description of how LTE specifications are organized with emphasis on how the Technical Specifications Group Radio Access Networks (TSG RAN) and their working groups (WGs) specify the LTE air interface. A flavor of the information provided to the students is shown in Fig. 2.

Observing student progress over the years, we have found that the tutorials should be defined around a handful of key themes and involve the students. A technique that has worked well is having the students summarize each session based on a template and specific questions that emphasize the key take-home messages. This way the students obtain a general overview of the LTE standard, how the specifications are organized, and how to search for details.

Following the LTE tutorials, we introduced ALOE and the tools for building an LTE system. The ALOE tutorial describes the ALOE architecture, tools, and services, and makes reference to the ALOE website (http://flexnets.upc.edu), where the entire ALOE code base resides and can be downloaded, installed, and modified for free.

	Learning stages	Learning objectives (To be able to…)	Tasks [Instructor] [Students]
I	1. Overview of the standard	**Good understanding of the standard** • Identify, at a high-level, the critical components of the standard, the relation among key components of the standard as well as some of the important options and tradeoffs • Discuss how and where to search for specific information	Tutorials • Standard technology and concepts description • Standardization mechanics and specification documents organization • SDR framework to be used in the project
	2. Abstract modeling	**Design the system** • Assemble a model of the main processing chain of the standard-specific transmitter and receiver • Discuss the processing tradeoffs and how they impact key performance parameters, such as synchronization, throughput, latency, and spectral efficiency	• Propose the project and define the specific assignments and milestones • Develop a complete model of the system • Document
II	3. Narrow the focus	**Define tests and figures of merit (FOMs)** • Identify the key FOM for a system of interest • Design performance and conformance tests based on the FOMs while taking into account the practical circumstances and limitation	• Define conformance tests and FOMs • Define performance test and FOMs • Document
	4. Development and testing	**Implement and test** • Implement the design from available building blocks • Test the implementation in terms of functionality, compliance with the standard specifications (conformance) and performance	• Provide a baseline implementation • Develop prototype to perform conformance and performance test • Support to validate results • In case of failure, propose and perform corrective measures
III	5. Review	**Review the product and process** • Identify where failures happened and discuss short-term remediation techniques as well as long-term solutions • Analyze and design possible system evolution	• Discuss what went right and what went wrong • Document
	6. Publicity and evaluation	**Demonstrate the product and process** • Demonstrate how objectives have been met and what process has been followed in obtaining the results • Defend the work and discuss alternative approaches • Evaluate the system and the team and individual team member performances	• Demonstration • Poster • Document and software library • Students provide a self-evaluation of the team and individual team members • Instructors evaluate group and individual performances

Table 1. Proposed learning stages.

Abstract Modeling: Modeling the Processing Chain

Methodology: For wireless communications standards, the PHY is a key component of the system and is therefore a candidate for more detailed analysis. By abstracting the PHY, other parts of the standard can be analyzed instead.

According to the project specification, defined by the professor, the student teams are tasked to develop a system model. This model should identify not only the functionalities (boxes, modules) and their interconnections, but also the working parameters as well as an estimation of complexity. This stage is part of Phase I, where the students develop a model based on the standards overview and available tools.

Case Study: Student teams develop a connected graph that illustrates the LTE PHY. One realization is shown in Fig. 3 and illustrates the simplified LTE PHY processing chain of the downlink transmitter and receiver. The colored blocks represent processing functions or processing chains and are specific to the LTE standard (http://www.3gpp.org/). For example, the resource demapping module, RESDEMAP, extracts the control and data symbols and demultiplexes them to be processed by different processing chains. The tables identify the amount of data flowing through the interfaces between the modules for two MCS instances.

Such high-level modeling along with the analysis of relations among modules and functionalities and the impact of some of the important parameters provide a good perspective for addressing the partial implementation and analysis of the LTE system.

Narrowing the Focus

Methodology: The extension of modern wireless standard specifications and the limited course duration require further narrowing down the focus of the project to specific aspects of the standard. The focus could, for example, be on breakthrough technological concepts that distinguish this standard from its predecessors or emerging concepts incorporated as part of the evolution of a standard. According to the specific project goals, students need to identify those parts of the standard's specifications that require deeper analysis.

Conformance tests are an important part of telecommunication standards. Performance tests help to understand the system behavior and to identify key figures of merit (FOMs). The focus could therefore be to identify standard-specific tests by the students under the close supervision of the instructor:

- *Define Conformance Tests* to check the suitability of the proposed implementation and fulfill the project specifications based on those defined in the standard.
- *Define Performance Tests* to address the impact of the employed technologies on the overall system performance.

Conformance tests are an important part of telecommunication standards. Performance tests help to understand the system behavior and to identify key figures of merit (FOM). The focus could therefore be to identify standard-specific tests by the students under the close supervision of the instructor.

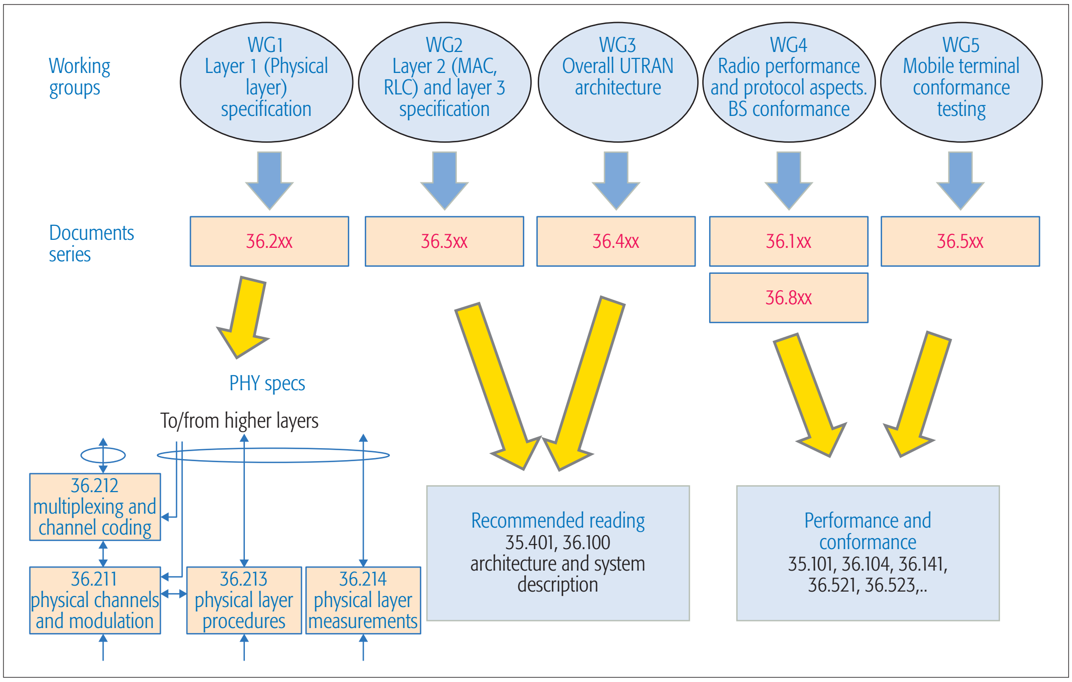
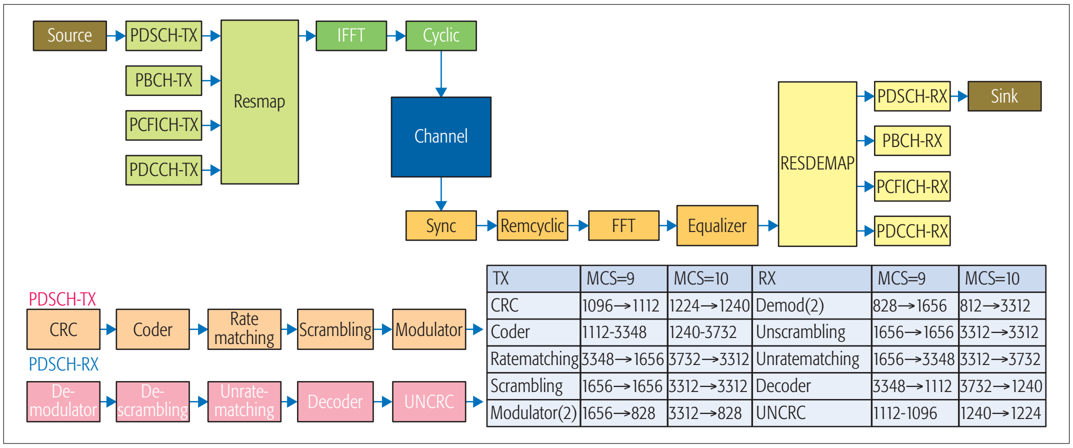


Figure 2. Working groups and documents specifying the LTE PHY.



TX	MCS=9	MCS=10	RX	MCS=9	MCS=10
CRC	1096→1112	1224→1240	Demod(2)	828→1656	812→3312
Coder	1112-3348	1240-3732	Unscrambling	1656→1656	3312→3312
Ratematching	3348→1656	3732→3312	Unratematching	1656→3348	3312→3732
Scrambling	1656→1656	3312→3312	Decoder	3348→1112	3732→1240
Modulator(2)	1656→828	3312→828	UNCRC	1112-1096	1240→1224

Figure 3. Modeling the LTE PHY processing chain.

As a result of both activities, the team develops a set of FOMs to quantitatively characterize the performance of the system w.r.t. the project requirements.

Case Study: The objective of the chosen project was to analyze the impact of the MSC on the LTE system performance. By measuring the computing cost, the system performance can be plotted vs. computing overhead to emphasize the growing importance of computing in modern wireless systems. The student team working on the project defined the following tests to validate the system and analyze its behavior:

- *Conformance tests*: The first test validates the behavior in terms of bit error rate (BER) of the downlink processing chain when using the three LTE modulation formats that map 2, 4, or 6 bits to modulation symbols and the simulated additive white Gaussian noise (AWGN) channel. The second test checks that the block error rate (BLER) is always below 0.1 or 10 percent, according to the specifications.
- *Performance tests*: The performance tests measure the BLER and the computing complexity (processing time overhead) for a selected set of MCS values and different signal-to-noise ratios (SNRs) in a simulated channel.

The student team understood how to validate the processing chain according LTE standard specifications for later analyzing system performance.

Development and Testing

Methodology: This is the core part of the proposed methodology, where the students actually implement part of the standard they have previously examined (1) and designed (2), and analyze their implementation based on the FOMs (3). The availability of a partial system implementation facilitates this phase and narrows it down to fit the course schedule. Students build the processing chain for their project from the standard specifications using the provided tools and baseline implementation. The system components and the subsystems are continuously validated for correct functionality using test vectors and known output statistics.

The second testing phase evaluates the system or subsystem for conformance, based on the FOM defined in the previous stage. The results obtained from the conformance tests are validated by the instructor. In case of failure, an analysis of the implication on the overall subsystem performance follows. The team then makes a decision whether to continue or solve the problem.

Once the conformance tests are satisfactory, the third testing phase can be initiated. The performance tests are performed and the results analyzed by the team in two sub-stages:

1. Analyze the system or subsystem performance w.r.t. the expected performance and discuss the differences, if any.
2. Devise corrective strategies if system performance does not match the expected results or discuss alternative solutions to improve the performance further.

In some cases, both conformance and performance test may require the use of simulated channels (e.g., simulated fading channels), whereas in other cases controlled over-the-air transmission and reception would be more appropriate.

Case Study: After having defined the FOMs in the previous stage, the students develop the partial system using a baseline implementation and perform the conformance and performance tests. The following figures and discussion, extracted from the project team's documentation, provides insights about the quality of the work as an indicator of success of the proposed methodology.

MCS and System Performance: Figure 4 plots BLER over MCS for different SNRs with the objective to check the compliance of available implementation with LTE standard specs. The project team realized that the demodulation and decoding process requires a certain SNR to achieve the 0.1 BLER target (3GPP LTE specs), which varies according to the chosen code rate or MCS.

The students learned how to use the turbo decoder and its relevance for error correction. Whenever the receiver implementation did not fulfill the 3GPP LTE specifications, the number of iterations was increased, from 1 decoding iteration (solid lines) to 5 (round points) in this case (Fig. 4).

MCS and Computing Cost: A second FOM of the learning process is capitalized in Fig. 5.

Figure 5 plots the user throughput and computing cost for 1.4 MHz LTE and different MCS values. LTE uses three modulation schemes, 4-, 16-, and 64-quadrature amplitude modulation, mapping 2, 4, and 6 bits to one modulation symbol. The computing cost was a measure of the time spent to execute the processing chain using an ASUS X200CA Netbook PC (Intel Core i3-3217U) and Ubuntu 12.04.3 LTS operating system. Figure 5a results from analyzing the relationship between the transport block size — the number of bits transmitted in one transmission time interval — the number of resource elements, and the throughput. The nearly linear relationship between computing cost and MCS matches the expected behavior. According to the 3GPP specifications, an LTE UE can send one of 16 channel quality indicators (CQIs) to inform the eNodeB about the highest MCS that it can decode with a BLER not exceeding 10 percent. Students experienced that more than one MCS can provide the required performance, but each has a different computing cost. Figure 5b shows the computing cost of the main processing blocks in the PDSCH processing chain of LTE. The students analyzed these figures to learn which blocks are critical and need careful (optimized) implementation.

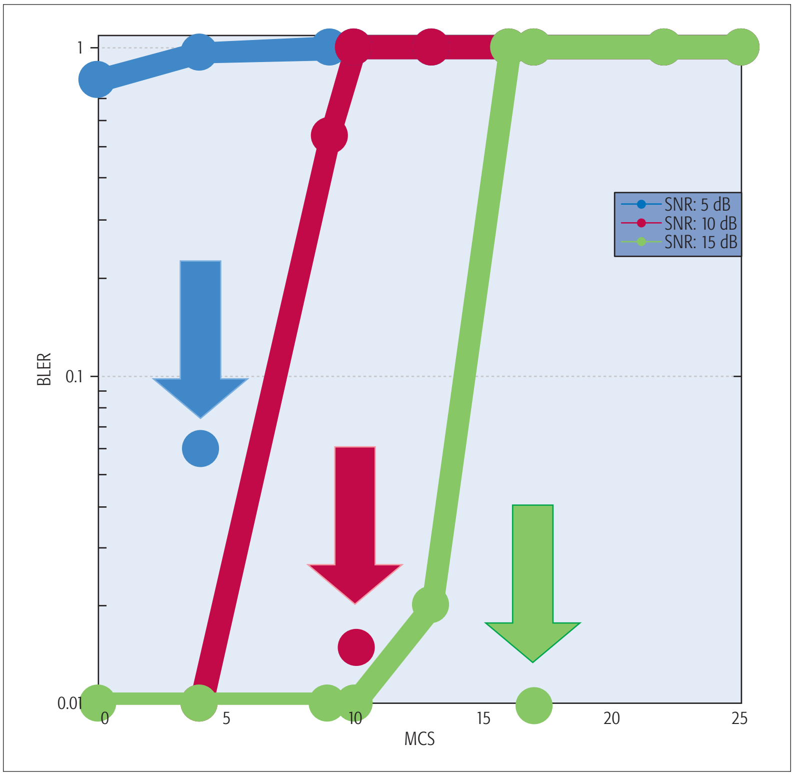


Figure 4. BLER vs. MCS for different SNR values. The solid lines indicate the achieved BLER with one turbo decoder iteration, whereas the round points correspond to the BLER achieved with five turbo decoder iterations.

Revision

Methodology: After successful completion of the tests, the students discuss what went right and what went wrong. In the case of unsatisfactory results, an analysis is conducted to identify the cause. This could need a new design (2), new FOMs (3), a review of the project procedure and goals, or even a revision of the standard [15].

Case Study: The use of FOMs based on implementation objectives helps to clarify to a student team their current status. Starting from the provided baseline implementation, students revise their system implementation continuously while progressing step by step, Corrective measures are taken when misalignments with specifications are detected. Regarding this case study, students experienced how the number of iterations of a turbo decoder impact the BLER but also the computing cost, and discussed about solutions from different points of view.

Along the years, the feedback provided by students has revealed that implementing a wireless standard requires advanced skills and more time. The following list summarizes the student feedback, which helped improve the tools and our methodology over the years:

- An optimized implementation of LTE that meets the timing requirements and FOMs requires experience with code optimization.

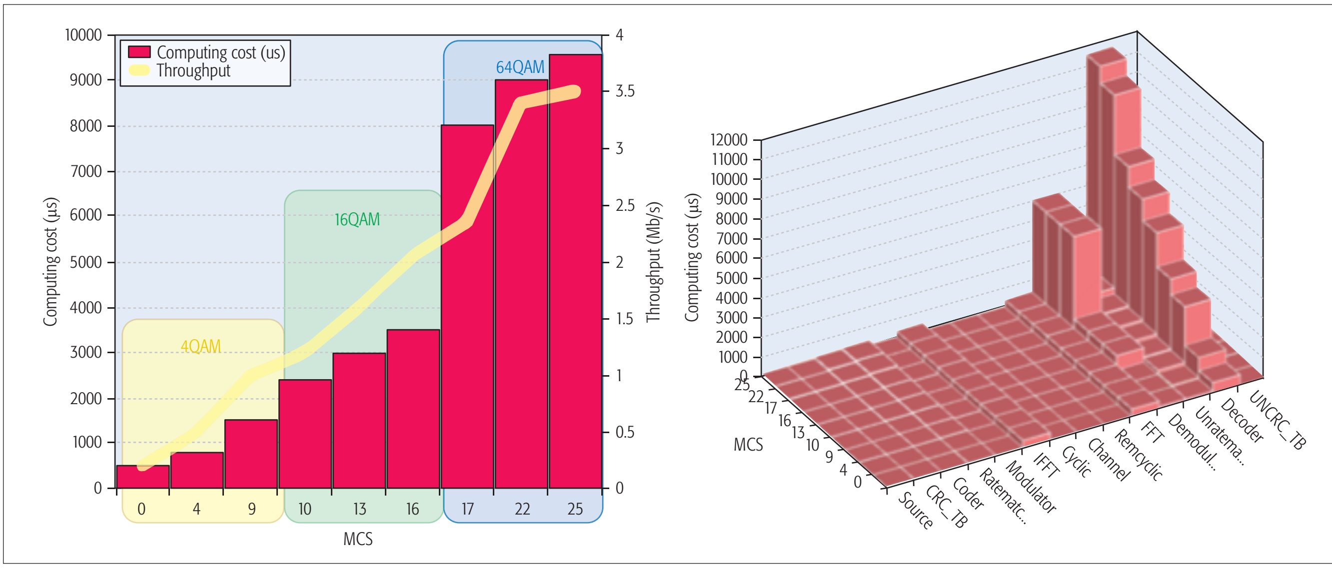


Figure 5. Computing cost: a) global computing cost and throughput vs. M.Sc. for one decoding iteration;d b) computing cost vs. MCS of the PDSCH LTE processing chain modules for five decoding iterations.

- A more detailed documentation of the provided baseline LTE implementation is desired for familiarization with the code.
- Incorporate means to identify potential bottlenecks in the project development early.
- Unbalanced or uncommitted teams need careful guidance.

Publicity and Evaluation

Methodology: The student evaluation is defined at three levels. The first one is based on the delivered documentation that describes the work done, the decisions made, and the system performance accomplished. A second level is done through a public presentation and demonstration of the work done to the entire class in a session open to other students and faculty. A third evaluation level is provided by each team member. They, better than anyone else, know the level of commitment and responsibility of each participant in the project team. This approach aims to enhance the cooperation skills of future engineers.

Case Study: Student teams provide a comprehensive document summarizing the standard pieces they have analyzed in more detail, the phases of the project, the realized tests and accomplished results, conclusions, and suggestions for improvement. At the end of each semester, the student teams present their accomplishments with demos, videos, or posters in a demo/poster session. All class instructors and students assist in this session, ask questions, and make suggestions. The evaluation is in part based on how well a group presents its work w.r.t. the class learning objectives and the specific project objectives.

Conclusions

Testing a prototype or product for performance or standard compliance is a valuable experience for electrical engineering students looking forward to contributing to current and next generation standards. The telecommunication industry is constantly looking for graduates with strong theoretical background as well as hands-on experience. Developing prototypes is always a huge endeavor, and dealing with concurrent processes of complex real-time systems is challenging for students. It is difficult to teach these skills as part of the engineering curriculum.

This article presents a PBL methodology and case study for teaching telecommunication standards. We identify three learning phases — modeling, development, and evaluation and review — subdivided into a total of six learning stages, and describe our methodology in terms of activities of students and instructors to meet specific learning objectives. Since standards are developed with implementation in mind, using the specifications to build a (simplified) product provides the best way of gaining a solid understanding of the standard. We suggest using SDR technology and the ALOE framework, which provides an effective working environment and baseline implementation for the project development in a confined class period. The staged PBL approach allows identifying the necessary skills, transmitting these to the students, and providing an effective learning environment for acquiring them.

We introduced the methodology into the electrical engineering curriculum at Barcelona Tech several years ago. Our students have had different interests and prior experiences. Some were motivated and acted as group leaders. Those students got the most out of the class. Other students delivered good work, but too narrow and specific. A balance is needed to gain broad knowledge without abstracting too many details. SDR technology and open source software frameworks, such as ALOE, provide an ideal framework for experiencing telecommunication standards and learning how to read, implement, and analyze the standards specifications in as much detail as considered adequate by the instructor.

Acknowledgments

This work has been partially supported by the Spanish Government, Ministerio de Ciencia e Innovación, through award number TEC2014-58341-C4-3-R.

References

[1] Standards Australia, "Economics Benefits of Standardization," research paper; http://www.standards.org.au, accessed Feb. 2017.

[2] J. Bezerra, *et al.* "The Mobile Revolution: How Mobile Technologies Drive a Trillion-Dollar Impact," Boston Consulting Group, Jan. 2015;, https://www.bcgperspectives.com, accessed Feb. 2017.

[3] O. A. B. Hassan, 2011, "Learning Theories and Assessment Methodologies — An Engineering Educational Perspective," *Euro. J. Engineering Education*, vol. 36, no. 4, doi: 10.1080/03043797.2011.591486, pp. 327–39.

[4] R. Hamid *et al.* ,2013, Competition as an Innovative Student-Centered Learning Method for Open-Ended Laboratory Work, *Procedia — Social and Behavioral Sciences*, 102, 2012, pp. 148–52, doi: 10.1016/j.sbspro.2013.10.726

[5] J.-S. Chenard, Z. Zilic and M. Prokic, "A Laboratory Setup and Teaching Methodology for Wireless and Mobile Embedded Systems," *IEEE Trans. Educ.*, vol. 51, no. 3, Aug. 2008, pp. 378–84.

[6] J. W. Thomas, "A Review of Research on Project-Based Learning," San Rafael, CA: Autodesk Foundation, 2000; http://www.bie.org/index.php/site/RE/pbl_research/29, accessed Feb. 2017.

[7] K. Edström and A. Kolmos, "PBL and CDIO: Complementary Models for Engineering Education Development," *European J. Engineering Education*, vol. 39, vol. 5, 2014, doi: 10.1080/03043797.2014.895703, pp. 539–55.

[8] E. Crawley *et al.*, 2014, *Rethinking Engineering Education: The CDIO Approach*, 2nd ed., doi: 10.1007/978-3-319-05561-9.

[9] S. L. Wismath and D. Orr, "Collaborative Learning in Problem Solving: A Case Study in Metacognitive Learning," *Canadian J. Scholarship of Teaching and Learning*, vol. 6, no. 3, article 10, 2015.

[10] S. Boss, "Project-Based Learning: A Short History"; https://www.edutopia.org/project-based-learning-history, Sept. 2011, accessed Feb. 2017.

[11] D. T. Rover *et al.*, "Reflections on Teaching and Learning in an Advanced Undergraduate Course in Embedded Systems," *IEEE Trans. Educ.*, vol. 51, no. 3, Aug. 2008, pp. 400–12, doi: 10.1109/TE.2008.921792.

[12] L. Jing *et al.*, "A Spiral Step-by-Step Educational Method for Cultivating Competent Embedded System Engineers to Meet Industry Demands," *IEEE Trans. Educ.*, vol. 54, no. 3, Aug. 2011, pp. 356–65, doi: 10.1109/TE.2010.2058576.

[13] C. Belisle *et al.*, "The Software Communications Architecture: Two Decades of Software Radio Technology Innovation," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 31–37, doi: 10.1109/MCOM.2015.7263343.

[14] I. Gomez, V. Marojevic, and A. Gelonch, "ALOE: An Open Source SDR Execution Environment with Cognitive Computing Resource Management Capabilities," *IEEE Commun. Mag.*, vol. 49, no. 9, Sept. 2011, pp. 76–83, doi: 10.1109/MCOM.2011.6011737.

[15] M. Labib *et al.*, "Enhancing the Robustness of LTE Systems: Analysis and Evolution of the Cell Selection Process," *IEEE Commun. Mag.*, vol. 55, no. 2, Feb. 2017, pp. 208–15.

Biographies

Antoni Gelonch-Bosch (antoni@tsc.upc.edu) has been an associate professor at the Department of Signal Theory and Communications of Barcelona-Tech University since 1997. His research interest has moved along the years from the development of suitable hardware platforms for implementing wireless systems, attending real-time processing constraints, to application of the software radio concept and development of appropriate frameworks to accelerate its development and deployment. Currently, he is focused on addressing the implementation, virtualization, and resource management issues of cloud RAN systems.

Vuk Marojevic (maroje@vt.edu) received his M.S. from the University of Hannover, Germany, and his Ph.D. from the Universidad Politècnica de Catalunya, both in electrical engineering. He joined Wireless@Virginia Tech in 2013. His research interests are in software-defined radio, spectrum sharing, 4G/5G cellular technology, wireless testbeds and testing, resource management, wireless security, and engineering education with application to public safety and mission-critical networks, the Internet of Things, and unmanned aircraft systems.

Ismael Gomez (ismael.gomez@softwareradiosystems.com) has over 10 years of experience in development and leadership of software defined radio using heterogeneous platforms. He is currently the lead developer of all LTE projects at Software Radio Systems (SRS). Prior to founding SRS, he held research positions in Trinity College, Dublin, Ireland, and Universitat Politècnica de Catalunya, Barcelona, where he has worked on the theoretical aspects of SDR processing, virtualization, and resource sharing in cloud RAN and massive distributed MIMO.

ITU Spectrum Management Training Program: A Comprehensive Modular Framework for Formalized Professional Education

Arturas Medeisis

The author describes the concept and composition of the ITU Spectrum Management Training Program that was developed by the International Telecommunication Union and is currently being implemented under the auspices of the ITU Academy. SMTP was designed as a flexible ECTS-based modular framework for professional training that allows various degrees of immersion as well as different specialization paths toward a formal professional certification in the field of radio spectrum management.

Digital Object Identifier:
10.1109/MCOM.2017.1601232

ABSTRACT

This article describes the concept and composition of the ITU Spectrum Management Training Program (SMTP) that was developed by the International Telecommunication Union and is currently being implemented under the auspices of the ITU Academy. SMTP was designed as a flexible ECTS-based modular framework for professional training that allows various degrees of immersion as well as different specialization paths toward a formal professional certification in the field of radio spectrum management. SMTP offers a solid and holistic system of professional education in both the theoretical and practical aspects of spectrum management as an important standardization activity supporting development of wireless communications. The ITU envisages SMTP eventually becoming a global "gold standard" for professional education in the highly specialized field of spectrum management. An essential differentiator for SMTP, compared to the limited offerings currently available for spectrum management training, is the formalized assessment of learning outcomes in terms of obtained professional skills and qualifications. Upon successful completion of SMTP, participants will be awarded a professional qualifications certificate by the ITU Academy or, possibly, a postgraduate professional education diploma of Master of Sciences from a partner university.

INTRODUCTION

Spectrum management (SM) is a highly specialized and vital activity in the field of telecommunication standardization, operational planning, and management. However, it rarely gets any serious degree of attention in the formal educational programs available at engineering colleges and universities. This may be explained by its niche role and by the fact that the target employment opportunities for professionals trained in the SM field would be very limited — primarily as the staff of the national regulatory authorities in charge of telecommunications. Secondary employment options for spectrum managers are few and far between, such as within the planning departments of mobile operators and similar large wireless telecommunication users.

The relative importance of this profession may be appreciated in light of the evolution and proliferation of wireless telecommunication services, which all depend on the timely availability and favorable access rules applicable to the radio spectrum. Therefore, the judicious apportionment of radio spectrum and supervision of its efficient use remain ever relevant and critical. At the same time, liberalization of markets means that the traditional technical aspects of efficient SM need to be increasingly aligned with economic and legal considerations. As a result of these trends, there is a pressing need for SM professionals that possess highly specialized technical skills as well as a solid understanding of many legal and economic issues involved. Tools of the trade are constantly expanding and now encompass anything from traditional electromagnetic compatibility formulas to the theory and practice of real-time spectrum auctions.

These professional competences contrast strongly with the generalist knowledge and limited practical skill set of a typical university or college graduate. Even those who have majored in telecommunication engineering often have a scant understanding of real-world radio spectrum use, not to mention the tools and specifics of how it is managed. Graduates entering the SM field would clearly benefit from additional specialized professional training in the subject; however, there are very few opportunities for comprehensive SM training in existence today, and those that are available only scratch the surface in terms of depth of the required knowledge.

Having identified this educational gap, the ITU Academy [1], a human capacity-building initiative of the International Telecommunication Union (ITU), has established a comprehensive formal Spectrum Management Training Program (SMTP), which is described in this article. It is hoped that the presented case study of setting up a specialized professional training framework could provide a useful example for other organizations and institutions that face similar training gaps in their respective fields.

The article begins with a review of the landscape of SM training options that are currently available and identifying the gap between those options vis-à-vis the professional skill set requirements. It then goes on to present the proposed concept of the SMTP framework composed of

The author is with the International Telecommunication Union.

a number of obligatory and elective modules, underpinned by a standardized system of academic credits. In the final sections the current status of SMTP operations is reviewed and followed by conclusions.

The Current SM Professional Training Landscape

On-the-Job Training

Our desk-based research has not identified a single formal higher education program dedicated to the training of SM professionals. This finding is echoed by anecdotal accounts from the international family of SM practitioners participating in various ITU and regional fora, who all seem to be introduced to the subtleties of the SM profession by means of on-the-job training.

On-the-job training of new SM recruits most often takes the form of an "apprenticeship" under the guidance of experienced colleagues. This approach might work reasonably well in large organizations with many employees and a wealth of accumulated personal and institutional experience. But it takes a long time to reach the desired level of professional performance and too often may leave significant gaps of professional understanding. Such a method is also prone to suffering from negative effects of new recruits acceding unquestioningly to the past institutional work culture and knowledge, and thus propagating any existing shortcoming or biases.

In young or small organizations, such as regulatory agencies in many developing countries, opportunities for acquiring the necessary qualifications by following experienced colleagues may be limited or nonexistent.

Internationally Available SM Training Options

Faced with the above limitations of on-the-job training, many new recruits and their managers seek opportunities for external professional SM training, but these are few and limited in their scope. The first port of call is often the ITU, which offers a number of sources for self-study through a large repository of handbooks, reports, and recommendations published by the ITU's Radiocommunication Sector (ITU-R). For decades now the ITU-R has also offered a number of regular SM seminars and workshops [2] with the aim of disseminating knowledge of best practices, policies, and national and international standardization processes in the field of SM. These events, especially the flagship biannual ITU World Radiocommunication Seminar, provide a good overview of state-of-the-art SM including updates on availability and use of various specialist software tools. But due to the ad hoc nature of these events, with a short duration and typically consisting of one-way plenary presentations, they do not provide a real alternative to formal and holistic professional training. An additional severe limitation of attending ITU seminars, or self-study by perusing ITU-R publications, is that they lack a formal assessment of the professional skills and qualifications achieved.

Another veritable non-profit institution that provides training in the fields of SM and more generally the ICT policy and standardization, is the U.S. Telecommunications Training Institute (USTTI) [3]. Each year USTTI offers a number of SM-related courses aimed primarily at young professionals from developing countries working in information, communications, and technology (ICT) industries and regulatory agencies. These courses are usually institutionally sponsored and tutored by staff from NTIA, FCC, or major U.S. companies engaged in wireless communications. The delivery format is face-to-face classroom training with duration from a few days to two weeks, held in the United States, usually at the facilities of the sponsor institution. The USTTI's courses offer professionally and culturally enriching experiences, but in terms of educational scope, outreach, and ultimate training impact, they are broadly similar to the ITU seminars.

Commercial SM Courses

Some SM training is offered on a commercial basis by consulting companies active in the SM field. This typically consists of an instructor-led classroom course delivered over a few days. Such courses are often aimed at beginners and offer a broad overview of various SM topics from the essentials (e.g., fundamentals of radio wave propagation and key spectrum engineering methods) to more advanced ones (market allocation methods and modern SM policy trends). Tuition is provided by one or more experienced consultants.

These courses generally offer newcomers a quick "crash course," based on a carefully selected package of topics. Learning from internationally recognized SM experts ensures access to some of the best professional know-how in the field. However, the condensed nature of these short courses does not allow for truly in-depth immersion in the subject matter, and the resulting outcome is inevitably patchy in terms of the professional skills and knowledge attained by the participants. Hence, these courses cannot be compared to formal comprehensive training that would cover the entire range of SM topics and lead to achieving measurable and lasting professional skills. Moreover, attending such commercial courses can be very costly.

The Educational Gap

When reviewing the available educational options in the field of SM, it is evident that there is a significant educational gap between what is offered today by various institutions and what is required from practicing SM professionals. As illustrated in Fig. 1, this gap exists in both the breadth of professional knowledge as well as the level of the practical skills required.

A further shortcoming of the current educational landscape is the absence of any real option for formal certification of the professional SM credentials achieved, be it through formal or on-the-job training. One existing certification option that comes close to the field of SM is the IEEE's Wireless Communications Engineering Technologies (IEEE WCET®) [4] professional certificate. WCET defines certain benchmarks for measuring professional qualifications in the field of wireless engineering and uses tests to determine whether or not a practicing professional has mastered those qualifications. Although IEEE WCET itself was not found to be directly applicable in the SM field, it provides a useful reference benchmark for estab-

Our desk-based research has not identified a single formal higher education program dedicated to the training of SM professionals. This finding is echoed by anecdotal accounts from the international family of SM practitioners participating in various ITU and regional fora, who all seem to be introduced to the subtleties of the SM profession by means of on-the-job training.

The flexible modular credit-based international programs were used as reference examples to design the SMTP structure. It is based on a number of obligatory modules, as well as some elective modules, with associated number of ECTS credits.

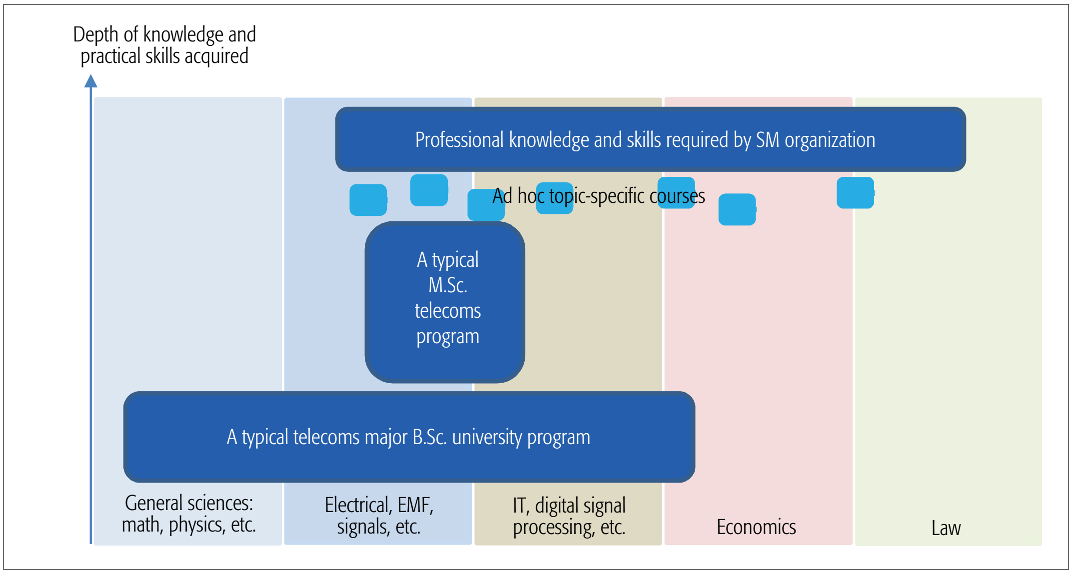


Figure 1. The educational gap between formal higher education and SM professional requirements.

lishing a comparable professional certification as the ultimate target of a purpose-built SM training framework.

Designing a Modular SMTP Framework

A Credit-Based Flexible Educational Framework

Over the past few years the ITU Academy has investigated the feasibility of establishing a dedicated SM training course at a post-graduate level to address the educational gap for SM professionals. It was decided that the course should cover the entire range of subjects involved in modern SM, be well structured and academically rigorous, and lead to a recognized professional certificate or diploma. Originally it was conceived as a "Spectrum Management Master Program" leading to an M.Sc. equivalent diploma. However, it became apparent that focusing solely on a formally accredited Master of Sciences program might become too much of a burden to set up while also becoming an unnecessarily rigid solution as it would have to be tied to one specific format and particular host institution(s) of higher education. Instead, the concept idea transformed to embrace a flexible framework of professional education consisting of a number of well defined and logically linked training modules, which can be used to build flexible training paths, as either part of the ITU Academy organized training or, eventually, part of formal M.Sc. programs at partner universities around the world.

After extensive analysis of the existing SM training options and reference sources vis-à-vis modern educational trends and practices, the SMTP concept was built based on several key cornerstones as illustrated in Fig. 2.

Taken together, these conceptual cornerstones both define the essence of the SMTP as a modern specialized educational program and provide answers to the previously identified shortcomings of existing training options, such as ensuring formal evaluation and providing solid and recognizable professional credentials.

The "academic credits" system used in many university programs today is an important enabler of modular education programs, permitting the trainees to tailor their study to any specific learning objectives at their desired study pace. After studying various options for SMTP, it was decided to use the well established principles of the European Credit Transfer and Accumulation System (ECTS) [5], which is an internationally recognized framework for evaluating the levels of university learning or other types of formal or informal continuing education. A key feature of the ECTS framework is that it is learner-oriented, where the credits correspond to the average time it takes a student to achieve certain learning objectives. Specifically, one ECTS credit corresponds to 25–30 hours of learning, and it is assumed that students taking a full-time formal educational course should be able to accumulate 60 ECTS credits per academic year. Then completing a first-level degree (B.Sc.) requires collection of 180–240 ECTS credits, while completion of a second-level degree (M.Sc.) requires further 90–120 ECTS credits.

The European academic landscape also offers some good examples of joint international teaching programs by university and industry consortia known as *Erasmus Mundus Master Courses*. Two examples in subject areas closest to SM include the Master of Science in Research on Information and Communication Technologies (MERIT) [6] and the Erasmus Mundus Joint Master in Economics and Management of Network Industries (EMIN) programs [7]. Although such programs do not change the essence or scope of university courses, they offer students more engaging and diversified studies, by combining modules taken at various participating universities and industry partners.

The flexible modular credit-based international programs were used as reference examples to design the SMTP structure. It is based on a number of obligatory modules, as well as some elective modules, with an associated number of ECTS credits. Similar to Erasmus programs, delivering SMTP through some (changing) forms and combinations of teaching consortia would provide two important benefits: the possibility to offer training locally in various regions of the world, and the possibility to promote incorporation of all or some of SMTP modules into formal M.Sc. programs of partner technical universities. This credit-based modular structure could provide a solid but versatile backbone for gradually building a formal SM training system to fill the identified educational gap, in terms of both the range of subjects covered and its global outreach.

Different Study Levels

It was also evident that the SMTP might be useful for both recent graduates beginning their SM careers and also more established professionals in the field who have previous practical experience, but may wish to round up their on-the-job training with an internationally recognized professional certificate. Therefore, the SMTP framework was designed with two entry levels in mind:

- A basic level, for people with little or no prior work experience in SM
- An advanced level, for those who either completed the basic level or have some equivalent practical experience in the field

The time required to complete the program would differ according to the entry level appropriate to the student, but in both cases, the studies would lead to the same professional certificate or diploma.

It is important to note that the possibility of joining SMTP at the advanced level should be conditional on meeting some formal requirements such as minimum number of years of practical experience and passing a formal entrance examination to assess previously obtained skills and competences. This would ensure compliance with the principles of ECTS, which allows recognition of outcomes of informal learning such as those obtained through practical work experience.

Modular Composition of SMTP

The initial composition of SMTP consisting of eight study modules is shown in Fig. 3. Each module is necessarily concluded by a standardized formal assessment tailored to the subject and content of the given module.

There are two types of modules: obligatory modules (OMs) and elective modules (EMs). The OMs deal with the core subjects of SM that should be understood by any practicing professional. At the basic level, they progress from detailed introduction of the legal basis of SM such as radio regulations and various other mechanisms and concepts constituting the entire regulatory framework of SM (OM1), to consideration of technical methods and tools of SM (OM2), and conclude with review of the role of SM as part of development of wireless telecommunications technologies (OM3). At the advanced level, it was considered to put more emphasis on specialization, so this level contains only two generic OMs: one reviewing the economic aspects and currently trending market-based tools of SM (OM4) and the second one dealing with impact analysis of SM regulations and policies (OM5)

The collection of EM modules, as their name suggests, allow trainees to shape their studies so that they could more closely correspond to the specific area of their professional interest within the SM field. It is also intended that the collection of offered EMs will be further expanded over time

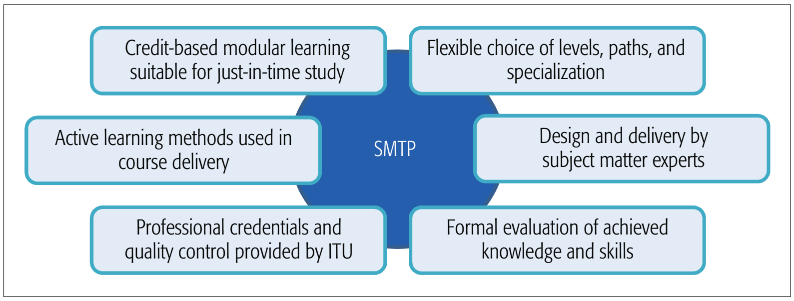


Figure 2. SMTP's key conceptual cornerstones.

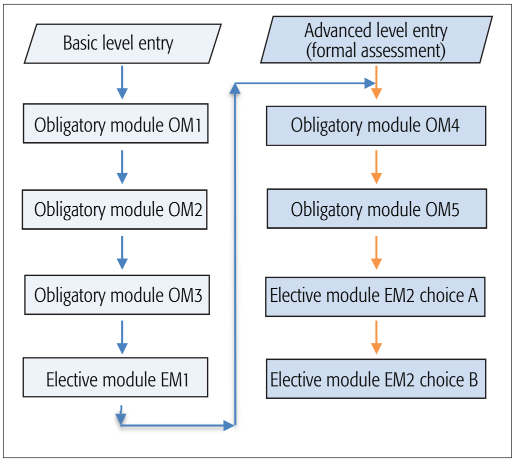


Figure 3. SMTP modular structure.

by the ITU Academy or partner training institutions to match the development of new wireless services and the appearance of novel SM toolsets, or to address some regionally or institutionally important topics. At the advanced level, where the taught skill sets are necessarily higher and thus imply a degree of specialization, the selection of EMs is streamlined by placing them into distinct specialization tracks: technical aspects vs. legal and market aspects. The initial collection of various SMTP modules is described in Table 1.

Delivery Options

The modular SMTP structure can easily be adapted to suit changing requirements and offer flexible delivery options including through cooperation of several institutions, for example, taking the following forms:

- A professional training course organized and delivered by the ITU Academy or its associated regional Centers of Excellence in the field of ICT. Upon completion of the training and passing all assessment tests, successful trainees would receive a professional SM qualification certificate granted in the name of the ITU Academy.
- An international course offered by one or a consortium of participating universities in partnership with the ITU and leading to a formal M.Sc. degree diploma, possibly along with supplementary SM proficiency certificate issued by the ITU Academy.

The SMTP could ultimately be used as a platform for entirely distance-based learning. In this case the ITU Academy would make available all training materials online, offer student counselling

Module #/track	Module subject
OM1	Legal Basis and Regulatory Framework of Spectrum Management
OM2	Spectrum Engineering Fundamentals
OM3	Wireless Telecommunications Technologies
EM1-1: Free choice	Spectrum Monitoring
EM1-2: Free choice	Enforcement and Type Approval of Equipment
EM1-3: Free choice	SM for Satellite Systems
EM1-4: Free choice	SM for HF Systems, Science, Maritime and Amateur Services
EM1-5: Free choice	SM for Aeronautical and Radio Determination Services and Military Systems
EM1-6: Free choice	Computer-aided Spectrum Management
OM4	Economic and Market Tools of Spectrum Management
OM5	Strategic Planning and Policies for Wireless Innovation
EM2-1: Legal track	Advanced Spectrum Authorization Regimes
EM2-2: Legal track	Socio-Economic Impact of SM; Competition and Consumer Protection
EM2-3: Technical track	Terrestrial TV Broadcasting Planning and Digital Transition
EM2-4: Technical track	Opportunistic Spectrum Access and Cognitive Radio

Table 1. Initial collection of offered SMTP modules: obligatory and electives.

and training guidance, and assume responsibility for final examination of learning outcomes. With this option in mind, the ITU Academy has developed study materials, lecture notes, as well as formalized assessment tests for all SMTP modules.

More detailed information on the design and composition of SMTP including initial syllabuses of offered modules may be found in the ITU report on the subject [8].

Status of SMTP Implementation

The syllabuses, training content, and assessment tests for the initial set of SMTP modules shown in Table 1 were developed by a large group of high-level international SM subject experts selected by ITU for this task based on their professional reputation and specialization areas. One important point, and challenge, addressed during this phase was to ensure that syllabuses for all modules follow a standardized syllabus format and incorporate as much as possible of active learning methods, such as peer-to-peer learning and interactive exercises.

Furthermore, all the teaching material developed for SMTP was subjected to comprehensive peer review by other members of the SMTP subject experts team as well as a number of volunteer SM professionals from the national regulatory authorities and the ITU staff, to ensure the highest quality of the final product.

Since completing the preparation of all teaching materials, the ITU Academy had been organizing pilot deliveries using its online training platform [1]. The most comprehensive and formal pilot to date was to test a delivery of module OM1 as a standalone course, since it is identical in its scope to the typical commercial crash courses mentioned previously. The test group was limited to 40 participants with 36 places equally distributed between applicants from the six world regions according to ITU's grouping of its regional activities: Africa, Americas, Arab States, Asia-Pacific, Commonwealth of Independent States, and Europe. All these participants were recent recruits from the national regulatory authorities with little or no previous experience in the SM field. The remaining four participants were ITU staff members acting as qualified observers to monitor the quality of teaching activities and level of formal assessments. The pilot lasted four weeks with 100 percent of initial participants completing the course and 70 percent successfully passing all formal assessments to meet the final qualification requirements. A questionnaire distributed at the end of the course returned very positive feedback from both trainees and trainers. The pilot thus provided a successful vindication of the SMTP concept and illustrated high demand from around the world for such training. Furthermore, the achieved high course completion and qualification rates indicated that the course format and teaching materials well matched the professional level and motivation of the participants.

Meanwhile, the ITU Academy had also been actively seeking academic partners as a means to facilitate regional SMTP delivery options and gauge opportunities for integration of SMTP into formal university M.Sc. programs. This invitation is open.

Some of the most recent implementation examples include the pilot delivery of all SMTP modules by the African Advanced Level Telecommunications Institute (AFRALTI) [9], which will offer the complete course as a combination of online and face-to-face sessions. Another example of pilot SMTP delivery by a regional partner university may be found at the Telecommunications Engineering Department of the Czech Technical University [10]. ITU sees a growing interest in delivering SMTP from a number of other universities and professional training institutions around the globe.

Conclusions

National management and international harmonization of radio spectrum is an increasingly important activity in the field of telecommunication standardization, which had been suffering from the lack of formal comprehensive professional education of specialist staff. The situation was further exacerbated by the fact that the SM profession today requires from its practitioners a broad collection of skills that span from technical to legal and economic fields of knowledge. Such broad but highly specialized knowledge, required by a handful of professionals globally, could only be meaningfully delivered through a suitably designed specialized professional training course, such as the SMTP training framework described in this article. Developed within the training and professional development branch of ITU, it is aimed to become a global standard of professional education in the SM field.

The proposed SMTP framework has a comprehensive curriculum that spans various SM subjects and functions, allowing participants to build a solid foundation of professional knowledge as well as optional specialization. SMTP's modular structure based on the ECTS credit framework allows flexible variation of study paths and portability of the entire program or its parts to various delivery platforms or training institutions. SMTP also offers a valuable option of incorporating the specialized SM training course as an integral part of formal university programs for postgraduate professional education. It is therefore considered that SMTP might be a useful case study for designing similar specialized programs of professional education in other fields.

Acknowledgments

The author's personal contribution to SMTP was limited to creating the original SMTP concept and its modular structure as described in this article and more fully in the referenced ITU Report [8]. The overall development of SMTP and its implementation is a result of team efforts by the staff in the Telecommunication Development Bureau and the Radiocommunication Bureau of ITU. Special acknowledgments go to the team in the Human Capacity Building Division (HCB) within the Projects and Knowledge Management Department of the Telecommunication Development Bureau. The work was carried out under the overall direction of Cosmas Zavazava, Chief of Department, with a team comprising Mike Nxele and Yevgeniya Minkova.

The author would also like to thank his colleague Conor Conran for reviewing the draft manuscript and providing many valuable suggestions for its improvement.

References

[1] ITU Academy, online information and training portal; https://academy.itu.int/, accessed Feb. 7, 2017.
[2] ITU Radiocommunication Seminars and Workshops; http://www.itu.int/en/ITU-R/seminars/, accessed Feb. 7, 2017.
[3] USTTI online information portal; http://www.ustti.org, accessed Feb. 7, 2017.
[4] IEEE WCET Certification online information portal; http://www.ieee-wcet.org, accessed Feb.7, 2017.
[5] ECTS Users' Guide 2015. EC, online information portal; http://ec.europa.eu/education/ects/users-guide/, accessed Feb. 7, 2017.
[6] MERIT online information portal: http://www.meritmaster.org, accessed Feb. 7, 2017.
[7] EMIN online information portal; http://www.icai.comillas.edu/en/master/emin-en, accessed Feb. 7, 2017.
[8] SMTP, ITU Telecommunication Development Sector Report, Dec. 2014; https://www.itu.int/en/ITU-D/Spectrum-Broadcasting/Documents/Publications/SMTP_2014.pdf, accessed Feb. 7, 2017.
[9] SMTP course offered by the African Advanced Level Telecommunications Institute; http://www.afralti.org/smtp/, accessed Feb. 7, 2017.
[10] SMTP course offered by the Czech Technical University; https://comtel.fel.cvut.cz/en/itu/academy/smtp, accessed Feb. 7, 2017.

Biography

Arturas Medeisis [M'95, SM'14] (arturas.medeisis@itu.int) received his doctorate from Kaunas University of Technology, Lithuania. He has been active in the field of radio spectrum management for more than 20 years, including positions with the National Regulatory Authority of Lithuania and the European Radiocommunication Office in Copenhagen, Denmark. Between 2007 and 2014 he held a position of associate professor at the Vilnius Gediminas Technical University, Lithuania. Currently he works as senior expert and government adviser at ITU.

SMTP also offers a valuable option of incorporating the specialized SM training course as an integral part of formal university programs for postgraduate professional education. It is therefore considered that SMTP might be a useful case study for designing similar specialized programs of professional education in other fields.

GREEN COMMUNICATIONS AND COMPUTING NETWORKS


Jinsong Wu


John Thompson


Honggang Zhang


RangaRao Venkatesha Prasad


Song Guo

The 2016 United Nations Climate Change Conference held in Marrakesh, Morocco, from 7 to 18 November 2016 was the 22nd Conference of the Parties (COP22) with delegates from 196 countries under the United Nations Framework Convention on Climate Change (UNFCCC) [1], and the 12th Conference of the Parties serving as the Meeting of the Parties to the Kyoto Protocol (CMP 12), and the 1st Conference of the Parties serving as the Meeting of the Parties to the historic Paris Agreement established in COP21 [2] (CMA 1). In COP 22, the Parties investigated the key steps to implement the Paris Agreement and develop the rule book to support its objectives and long-term goals [1]. CMA 2 or COP23 will evaluate progress in developing the rule book in 2017 [1]. The parties in COP22 decided that the Adaptation Fund, created in 2001, will support the Agreement, and endorsed the Marrakech Action Proclamation, which reaffirmed their commitment to implementing the Paris Agreement [1]. The Marrakech Partnership for Global Climate Action was established in COP22 as a platform to facilitate the involvement of non-state actors in pre-2020 climate action. Besides the formal negotiations, the Climate Vulnerable Forum (CVF), a group of 48 developing countries, declared their plans in COP22 to achieve 100 percent renewable energy between 2030 and 2050. Canada, Germany, Mexico, and the United States demonstrated their strategies in decarbonizing their economies by 2050 [1]. Developed countries established a roadmap to 2020 on achieving the agreed goal of US$100 billion per annum in climate finance for developing countries. The new Nationally Determined Contributions (NDC) Partnership in COP22 gathered together 33 countries and nine international institutions to develop tools, share best practices, and facilitate support for countries to reach ambitious climate and development goals as soon as possible [1]. As a parallel effort, in 2016, IEEE approved the IEEE Environmental Engineering Initiative, a pan-IEEE effort to support general environmental sustainability [2–4] under the IEEE Technical Activities Board (TAB) across more than 25 IEEE Societies, Councils, and other organization units (OUs).

All those above efforts will inspire the *IEEE Communications Magazine* Series on Green Communications and Computing Networks to continuing promoting green efforts. The sixth, May 2017, issue of this Green Series includes relevant articles.

The article "A Cost-Efficient Communication Framework for Battery-Switch-Based Electric Vehicle Charging" provides a brief review of state-of-the-art EV charging management schemes, and develops a fully distributed charging management scheme with consideration of urban travel uncertainties based on a publish/subscribe communication framework.

In the article "On the Fundamental Energy Trade-offs of Geographical Load Balancing," the achievable energy tradeoffs are considered via defining a service efficiency parameter for geo-dispersed data centers.

The article "Multi-Method Data Delivery for Green Sensor Cloud" addresses the potential applications and recent work on sensor cloud (SC), observes two issues regarding green SC, and further proposes a multi-method data delivery (MMDD) scheme for SC users.

The article "Energy Efficiency Challenges of 5G Small Cell Networks" investigates the computation power based on the Landauer principle in massive MIMO-supported 5G small cell networks.

ACKNOWLEDGMENTS

We would like to acknowledge the great support from Osman S. Gebizlioglu, the current Editor-in-Chief of *IEEE Communications Magazine*, and Peggy Kang, the Managing Editor of the 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), as well as the IEEE Environmental Engineering Initiative (EEI) under the IEEE Technical Activities Board (TAB).

REFERENCES

[1] Marrakech Climate Change Conference, Nov. 2016; http://unfccc.int/meetings/marrakech_nov_2016/meeting/9567.php, accessed Feb. 10, 2017.
[2] J. Wu *et al.*, "Green Communications and Computing Networks," [Series Editorial]," *IEEE Commun. Mag.*, vol. 54, no. 5, May 2016.
[3] J. Wu *et al.*, "Green Communications and Computing Networks Series Editorial," *IEEE Commun. Mag.*, vol. 54, no. 11, Nov. 2016.
[4] J. Wu *et al.*, "Big Data Meet Green Challenges: Big Data Toward Green Applications," *IEEE Systems J.*, vol. 10, no. 3, Sept. 2016 (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 officially elected Vice-Chair of Technical Activities, IEEE Environmental Engineering Initiative, a pan-IEEE effort under the IEEE Technical Activities Board (TAB). He is the founder and Founding Chair of the IEEE Technical Committee on Green Communications and Computing (TCGCC). He is an Area Editor of *IEEE Transactions on Green Communications and Networking*. He was the lead Editor and co-author of the comprehensive book *Green Communications: Theoretical Fundamentals, Algorithms, and Applications* (CRC Press, 2012).

John Thompson [FIEEE] (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 U.K. 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 with 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 of ComSoc during 2011–2012. He was the Lead Guest Editor of *IEEE Communications Magazine* Feature Topics on Green Communications. He served as the General Co-Chair of 2010 IEEE GreenCom and the Co-Chair of IEEE Online GreenComm 2015. He is the book 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, Esqube Communication Solutions. In 2005, he joined TUDelft. He worked on personal networks, IoT, CPS, and energy harvesting networks. His work at TUDelft has resulted in 180+ publications. He is also 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.

A Cost-Efficient Communication Framework for Battery-Switch-Based Electric Vehicle Charging

Yue Cao, Shusen Yang, Geyong Min, Xing Zhang, Houbing Song, Omprakash Kaiwartya, and Nauman Aslam

The authors present a brief review on state-of-the-art EV charging management schemes. Next, by incorporating battery switch technology to enable fast charging service, a publish/subscribe communication framework is provisioned to support the EV charging service. After that, they develop a fully distributed charging management scheme with consideration of urban travel uncertainties.

Abstract

Charging management for EVs on the move has become an increasingly important research problem in smart cities. Major technical challenges include the selection of charging stations to guide charging plans, and the design of cost-efficient communication infrastructure between the power grid and EVs. In this article, we first present a brief review on state-of-the-art EV charging management schemes. Next, by incorporating battery switch technology to enable fast charging service, a publish/subscribe communication framework is provisioned to support the EV charging service. After that, we develop a fully distributed charging management scheme with consideration of urban travel uncertainties, for example, traffic congestion and drivers' preferences. This would benefit from low privacy sensitivity, as EVs' status information will not be released through management. Results demonstrate a guidance for the provisioning of a P/S communication framework to improve EV drivers' experience, for example, charging waiting time and total trip duration. Also, the benefit of a P/S communication framework is reflected in terms of the communication efficiency. Open research issues in this emerging area are also presented.

Introduction

As the emerging key urban infrastructures, the smart grid and intelligent transportation systems (ITS) have been playing increasingly important roles in modern cities. This enables electric vehicles (EVs), which are expected to be widely adopted in individual, commercial, and public vehicle fleets.

However, compared to traditional gasoline-powered vehicles, on-the-move EVs are more likely to run out of energy and need to be charged during their journeys. As a result, how to manage the charging processes of EVs to improve their drivers' comfort is a vital research issue for the success and long-term viability of EV companies.

Existing charging techniques still require a relatively long duration to complete battery charging (typically half to several hours [1]), leading to frequently overloaded charging stations (CSs) caused by their typically limited forecourt areas [2]. The time and effort spent on seeking available CSs over the city, and waiting in the service queue would cause an uncomfortable and anxious driving experience for EV drivers. A promising alternative approach to the traditional charging service, battery switch service [3], can replace a fully charged battery for an EV within several minutes by using industrial automation robots.

As EVs become more prevalent, their charging demands will significantly rise for CSs throughout smart cities. Therefore, there is a necessity to design the communication infrastructure with efficiency and sustainability in mind. We aim to answer the following four questions in this article:

- How do we provision an ITS enabled communication framework for EV charging management, via techniques including the roadside unit (RSU), GPS, and standardization of vehicle-to-infrastructure (V2I) communications (e.g., 802.11p [4])?
- How does the provisioning of a communication framework affect the actual driving experience, and how cost-efficient is it?
- Which CS is ranked as the best plan for an EV on the move thatneeds the battery switch service to perceive a comfortable driving experience (e.g., minimized charging waiting time and trip duration)?
- What are the influences of urban driving uncertainties (e.g., traffic conditions and drivers' trip preferences) and battery switch systems on the driving experience?

By facilitating the battery switch service, we provision an ITS-enabled publish/subscribe (P/S) [5] communication framework, where necessary information (availability to provide the battery switch service) is shared among different EVs and other ITS entities such as RSUs. We further propose a CS selection scheme driven by the drivers' trip preferences. Throughout the case study in the Helsinki city scenario, the influence of communication provisioning and urban driving uncertainties on driving experience are studied.

Review of EV Charging Management

Parking Mode

The majority of previous works have addressed this use case (concerning when/whether to charge EVs), where EVs have already been parking at homes/CSs. Details can be found in [6]; herein we briefly summarize these works as:

Yue Cao and Nauman Aslam are with Northumbria University; Shusen Yang is with Xi'an JiaoTong University; Geyong Min is with the University of Exeter; Xing Zhang is with Beijing University of Posts and Telecommunications; Houbing Song is with West Virginia University; Omprakash Kaiwartya is with Universiti Teknologi Malaysia. The corresponding author is Shusen Yang.

Digital Object Identifier: 10.1109/MCOM.2017.1600913

0163-6804/17/$25.00 © 2017 IEEE

- Schedule and control the charging/discharging of EVs [7] (depending on different charging technologies, e.g., normal and fast charging), with different durations such that power grid constraints are maintained. This benefits the power grid such that peaks and possible overloads of the electricity network may be avoided.
- Address pricing issue [8] in order to encourage EVs not to charge during periods of high demand. This is mainly related to the economy issue, as the charging price is normally higher in peak hours than in off-peak hours.
- Integrate renewable energy, mainly solar and wind into grid as complimentary solution, from which sustainable energy could be provided to support massive demands.

On-the-Move Mode

A few works have studied how to manage EV drivers' charging plans when they are on the move:
- Route EVs (with charging event [9]) to minimize energy loss and maximize energy harvested during a trip, such that the time spent to fully recharge EVs is minimized. This would consider EV speed, as part of the efficiency of EVs results from their ability to recover some energy during deceleration.
- Where to deploy CSs (providing either plug-in charging or battery switch service [3]) such that EVs can access CSs within their driving ranges. Also, the capabilities of CSs to handle peak demands are taken into account due to different numbers of EV arrivals at different times.
- Select the appropriate CS as charging plan (or refer to where to charge). For example, select the CS that is not highly congested [10], so as to experience a minimized charging wait time.

Provisioning of a P/S Communication Framework For Battery Switch Service

In this article, we focus on the latter use case (i.e., on-the-move mode) and aim to determine where (which CS) to charge at in real time. Since previous works have not brought the benefit of battery switch technology for enabling fast EV charging in on-the-move mode, we lead an interdisciplinary contribution by bridging that advanced charging technology and provisioning of cost-efficient communication to drive information exchange within an EV charging system.

Battery Switch Service

To promote the popularization of EVs, it is necessary to build the infrastructure for charging batteries. Traditional plug-in recharging is accomplished by plugging EVs into charging slots (set by CSs placed at different city locations). In contrast, at a CS providing the battery switch service [3], the automated switch platform switches the depleted battery from an EV with a fully charged battery it maintains. The depleted battery is placed and recharged so that it can be used by other EV drivers. This means that each CS is able to maintain a certain number of batteries for switching.

The battery switch service could be described as a mixture of a drive-through car wash, which normally switches an EV's battery in several minutes, without requiring the driver to get out of the EV. Note that, as the cost to the battery's lifespan may be taken into account, the fast charging still takes a toll that should be avoided when possible. The fact that depleted batteries are charged by CSs (normally via a lower power) certainly removes the burden on EV drivers to maintain batteries.

To promote the popularization of EVs, it is necessary to build the infrastructure for charging batteries. Traditional plug-in recharging is accomplished by plugging EVs into charging slots. In contrast, at the CS providing the battery switch service, the automated switch platform switches the depleted battery from an EV, with a fully charged battery it maintains.

Centralized vs. Distributed Charging Management

In general, on-the-move EV charging management can be executed in both centralized and distributed manners. In the centralized manner, the charging management is executed by a global controller (GC) or other third party who is interested in charging management. However, this causes a big privacy concern, because the EV status information (e.g., location, trip destination, and ID) needs to be reported to the GC. The distributed manner benefits from low privacy sensitivity, where the charging management is executed by an EV individually (using accessed condition information from CSs). Thus, the accuracy of information plays an important role in charging management. This is because the CS selection decisions made by EVs would be suboptimal due to imperfect information acquired.

The Publish/Subscribe Communication Paradigm

The publish/subscribe (P/S) [5] paradigm is a suitable communication paradigm for building applications in vehicular ad hoc networks (VANETs) with a highly dynamic and flexible nature, for example, delay-tolerant networking (DTN) [11]. The following three network entities are involved.

Roadside Unit: It is strategically deployed at a certain location, and behaves as a broker to bridge the information flow from CSs to EVs. Each RSU is able to aggregate all CSs' condition information and caches it in local storage.

Electric Vehicle: Each EV has a status of charge (SOC). The EV, as subscriber, actively sends a query to subscribe to the information relayed by RSUs. If the ratio between its current energy and maximum energy is below the value of SOC, the EV starts to select a CS for the battery switch based on its gathered information.

Charging Station: CSs need to be distributed in a different way, usually in special parking spots or near shopping malls. Each CS maintains a number of fully charged batteries to provide the battery switch service. As EVs arrive, the number of maintained (fully charged) batteries will decrease because of switching. These depleted batteries from EVs may have some residual electricity but have not been fully charged yet. Since each CS needs to charge depleted batteries, its maintained number of fully charged batteries will increase. It periodically publishes its condition information to legitimate RSUs.

In Fig. 1, each CS publishes its condition information (i.e., availability of batteries for switching) to EVs as subscribers of this information. Along with this, strategically deployed RSUs can support information dissemination through V2I communication. The provisioning of a P/S communication framework fits the distributed charging management well, where EVs could access CSs' condition information from opportunistically encountered

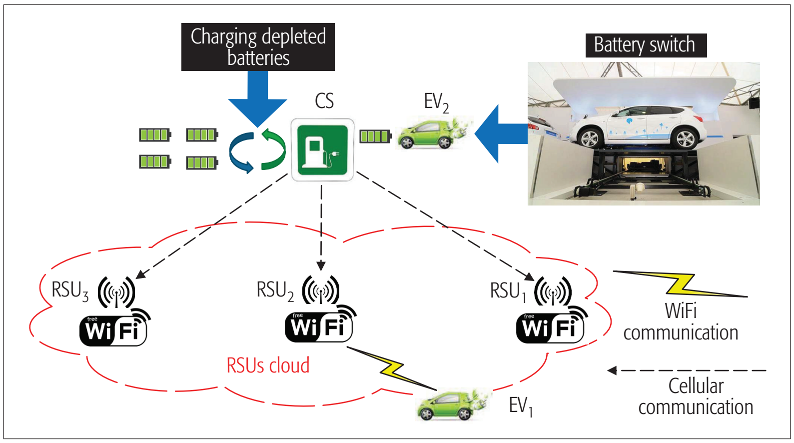


Figure 1. Big picture where EVs access periodically published CS information from RSUs, and further utilize collected information for CS selection.

RSUs (within the RSU cloud to share all CSs condition information), and perform their individual charging management when needed.

The Design of a P/S Communication Framework for Battery Switch Service

All CSs are geographically deployed, and their locations are pre-known by all EVs. Each CS is connected to all RSUs using a reliable channel such as authorized cellular network communication, and periodically publishes its condition information, for example, the available time for switch[1] (ATS). Given strategically deployed RSUs, there will not be an overlap between the radio coverage of adjacent RSUs. Such opportunistic communication between RSUs and EVs inevitably results in obsolete information accessed by EVs. The information exchange between CSs and EVs through RSUs is based on the P/S communication framework.

The communication in ITS enables information broadcasting to involved entities. In the case of EV charging application, the European Telecommunications Standards Institute (ETSI) TS 101 556-1 [12] standard is used. Its basic application is to notify EV drivers about the CSs' condition information, such that they are able to select a CS for battery switch. Further to [10], enabling the plug-in charging technology, we enhance its communication framework with additional effort to enable the battery switch with time sequences illustrated in Fig. 2.

Step 1: All CSs' information publications are synchronized. Each CS periodically publishes its ATS to RSUs, using the topic ATS_Update defined in Table 1. Strategically deployed RSUs could aggregate information from multiple CSs and cache it. Note that once new condition information has been received from CSs, RSUs will replace the obsolete one cached in the past.

Steps 2–3: Given an encounter opportunity with an RSU, the EV, being aware of updated service from that RSU, will send a subscription query[2] using the topic Aggregated_ATS_Access. This normally requires communication to br established from the EV to the RSU, via WiFi, enabling V2I communication.

Compared to [10], via a single topic for accessing CSs' queuing time, we bring two distinct topics illustrated in Table 1 and enable computation at the RSU side. Here, the basic idea is placing lightweight cloud-like facilities (e.g., RSUs) in the proximity of mobile users (e.g., EVs moving on the road) for authentication and reducing redundant information subscription. This is motivated by the recent trend of data services toward the edge of the cloud, resulting in novel architectures called fog computing [13].

Other Alternative Options

In Fig. 2, we also present two alternative options that can support information dissemination in the system.

Centralized Case (CC): It has been widely adopted by the majority of previous works. Throughout the cellular network communication, the GC globally monitors real-time CSs' ATS, and processes charging requests from on-the-move EVs.

Periodic Broadcasting (PB): This is a simple case where each CS periodically (with interval T) broadcasts its ATS to all EVs, also equivalent to the case where drivers use mobile phones to collect broadcasted CSs' ATS. The broadcasting is through the cellular network communication, and there is no RSU involved. Each EV can definitely access CSs' ATS within interval T, different from P/S affected by opportunistic encounters.

Discussion on Communication Cost

Possibility to Access Information in P/S: In [10], the P/S communication enabling the plug-in charging service (with CS queuing time published) was discussed. The fact that information is accessible depends on:

- Whether there is an encounter between the EV and RSU
- Whether an RSU (encountered by EV) has cached the information published from CSs

The analysis is based on a straight road model, where an EV (with constant/average moving speed) will pass through a number of N_{rsu} RSUs. The possibility $P_{p/s}$ of an EV to access information from at the least one of N_{rsu} RSUs, is given by

$$P_{p/s} \leq 1 - \prod_{i=1}^{N_{rsu}} \left\{ 1 - \left[\frac{(i-1)S + F + R}{V \cdot T} \right] \right\} \quad (1)$$

where N_{rsu} is the number of RSUs on the road, S is the distance between adjacent RSUs, and T is the publication interval (how often the information is published) of a CS. Also, V is the EV moving speed, while F is the distance from the starting point to the center of the first RSU.

In order to increase $P_{p/s}$, we obtain:

- Increased radio coverage R
- Increased number of RSUs N_{rsu}
- Reduced CS publication interval T (to increase CS publication frequency)

Communication Cost: Further to above, we herein denote N_{ev} as the number of EVs in the network. Then the communication costs of P/S, CC, and PB are given:

- In P/S, the cost for information access is given by

$$O\left(\frac{P_{p/s} \times N_{ev}}{T} \right),$$

since there are only ($P_{p/s} \times N_{ev}$) subscribers within each T interval.

[1] This information reflects the status of those batteries being charged. For example, given that a CS initially maintains 10 batteries, as time passes the status that 7 batteries are switchable while 3 batteries are being charged is published regarding the charging finish time (availability) of these 3 batteries being charged.

[2] It is worth noting that the subscription query received from EVs will be processed by RSUs, which only reply with the aggregated CSs' ATS associated with the updated time slot. This facilitates efficient radio resource utilization and alleviated interference to EVs.

- In CC, the cost at the GC side for handling EVs' charging requests is $O(N_{ev})$.
- In PB, each CS experiences a communication cost of

$$O\left(\frac{N_{ev}}{T}\right),$$

for broadcasting its ATS to all EVs.

Different from the CC communication framework, we can obtain scalability (i.e., the number of connections at CSs does not depend on the number of EVs), as the benefits of P/S-based communication against point-to-point communication.

On-the-Move EV Charging Management via the P/S Communication Framework

Urban Driving Uncertainties

Mobility Uncertainty: It refers to the situation in which there are several traffic jams occurring in a city. An EV within a certain range of a traffic jam has to slow down its speed, and it will accelerate its speed once leaving the range of that traffic jam. In particular, an EV may temporarily stop for a while if it is close to the central traffic. In such a case, an EV only resumes its movement once the closest traffic jam disappears. Due to the mobility uncertainty, the variation of EV moving speed will inevitably affect the arrival time at the CS and the electricity consumption for traveling toward that CS. Further to this, the mobility uncertainty also affects the traveling time taken from a CS to an EV's trip destination.

Trip Preference Uncertainty: It refers to the situation in which the EV drivers' direction of where to travel is uncertain. Here, EV drivers may have their daily routes or points of interest (POIs) to visit, for example, shopping malls or public parks for leisure. The trip preference uncertainty affects CS selection, where suboptimal charging during a journey may degrade a driver's comfort.

System Cycle of On-the-Move EV Charging Management

Figure 3 describes the system cycle of charging management:

Driving Phase: The EV is traveling toward its trip destination. This phase is affected by both the mobility uncertainty and trip preference uncertainty.

Charging Planning Phase: The EV reaching a threshold on its residual battery volume applies a policy to find a CS for the battery switch. Based on the locally recorded CSs' condition information, the EV runs a CS selection logic.

Battery Switch Phase: Upon arrival at the selected CS with parking place navigated [14], the EV's battery (electricity consumed) is switched with the one (fully charged) maintained at that CS. This happens if the selected CS already maintains a number of fully charged batteries for switching. Particularly, if the number of fully charged batteries at CSs is less than the number of EVs already parking therein, the EV charging scheduling (concerning when to charge) is based on the first come first serve (FCFS) order. This means that an EV with an earlier arrival time will be scheduled with a higher charging priority.

Battery Charging Phase: A number of batteries depleted from EVs will be charged by CSs in parallel. They will be switchable once being fully recharged.

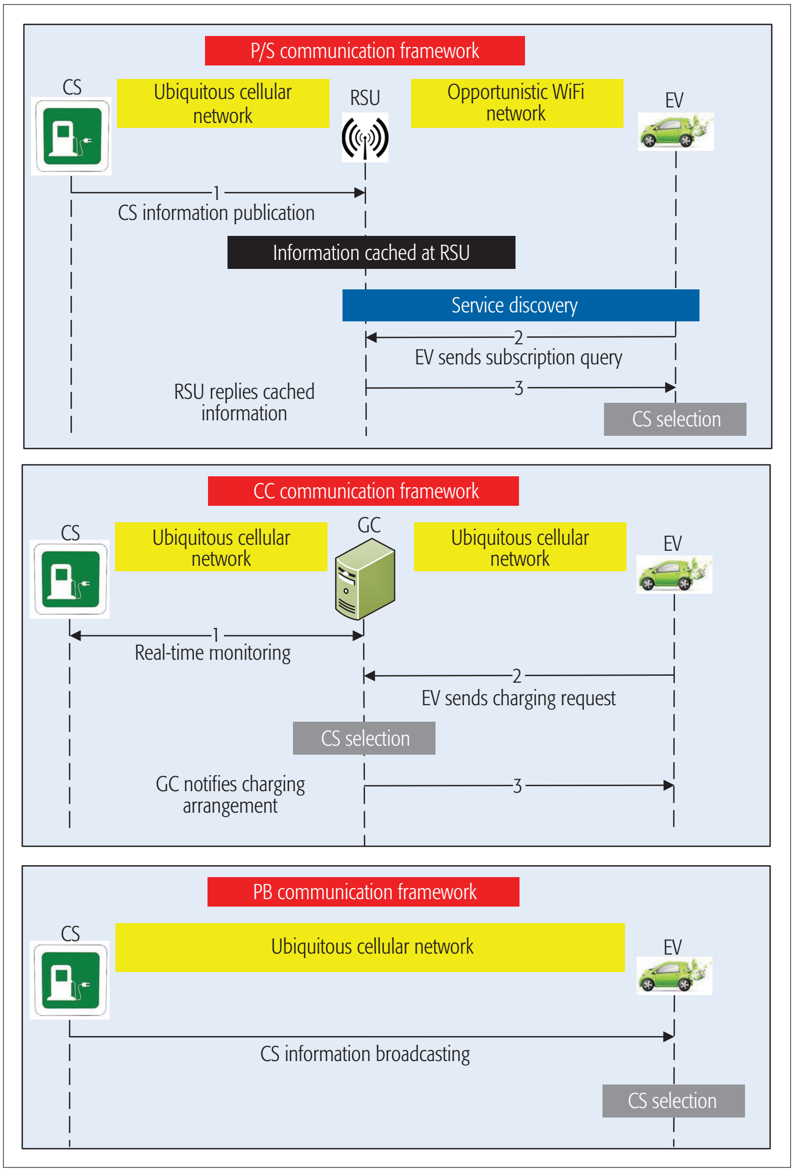


Figure 2. Time sequences of P/S, CC, and PB communication frameworks.

Note that the transition between *battery switch phase* and *battery charging phase* is bidirectional.

CS-Selection Logic

Note that the EV might have received aggregated CSs' condition information several times before it reaches the threshold for requesting the battery switch. The CS selection logic is to find the CS through which the EV will experience the shortest trip duration, including:
- Time to travel toward the selected CS
- Time to travel from the selected CS to the EV's trip destination
- Time to spend at the selected CS

The first two metrics can be computed directly; we introduce following steps to estimate the third metric.

Step 1: Run at the CS side, it first checks the maintained number of fully charged batteries. If there is availability for switching, the ATS is returned without containing those batteries' status. This means that the CS is currently able to provide the battery switch service.

Topic	Dissemination nature	Publisher(s)	Subscriber(s)	Payload
ATS_Update	One-to-many	CS	RSUs	<CS ID & CS's ATS, publication time slot>
Aggregated_ATS_Access	Many-to-many	RSUs	EVs	<Aggregated CS IDs CSs' ATS, publication pime slot>

Table 1. Topics defined in the P/S system.

Step 2: Run at the CS side, alternatively, if it is currently without switchable batteries, it then checks the number of batteries waiting for charging. Here, those batteries waiting for charging are sorted following the shortest time charge first (STCF) policy. This implies that the depleted battery with much remaining volume is charged with higher priority.

Step 3: Run at the CS side, only concerning those batteries already being charged, their charging finish time is included in a temporary list, called LIST, for computation purposes. Then the earliest time of LIST could be obtained, while all information in LIST is copied into ATS.

Step 4: Run at the CS side, the output of LIST from step 3 plus the charging time of each battery (waiting for charging) is calculated as the charging finish time of that battery, which is further included in ATS. This value is also replaced with the earliest value in LIST.

Step 5: Steps 2–4 are repeated until the number of resting batteries (waiting for charging) reaches 0. Then an updated ATS is returned.

Note that if the EV arrival time is later than any value in ATS, this means it would not experience any delay for battery switch upon arrival at a CS. Otherwise, the waiting time for battery switch is given by the earliest value in ATS-arrival time.

Case Study

We have built up an entire system for EV charging in the Opportunistic Network Environment [15]. In Fig. 4, the default scenario with 4500 × 3400 m^2 area is shown as the downtown area of Helsinki, Finland. 200 EVs with [30 ~ 50] km/h variable moving speed are initialized in the network. The configuration of EVs follows the charging specification (maximum electricity capacity, maximum traveling distance: 16.4 kWh, 140 km) of the Hyundai BlueOn EV (en.wikipedia.org/wiki/Hyundai BlueOn), and we set SOC ranging from 15 to 45 percent for all EVs.

Each CS maintains 30 fully recharged batteries at the beginning of the simulation, and can charge 20 depleted batteries from EVs in parallel using 10 kW power. The battery switch time is fixed as 5 min. There are totally five CSs provided with sufficient electric energy for battery charging. Also, 300 m radio coverage is applied for 7 RSUs and 200 EVs. The default update interval (publication frequency) of a CS is 120 s, and the simulation time is 43,200 s = 12 h.

The charging management scheme proposed above is evaluated under the P/S, CC, and PB communication frameworks discussed previously. We are mainly interested in:

- *Average charging waiting time* — the average period between the time an EV arrives at the selected CS and the time it finishes battery switching
- *Average trip duration* — the average time that an EV experiences for its trip through the battery switching service at an intermediate CS
- *Number of accesses* — number of times that EVs access CSs (through a cellular network or RSUs)

The Influence of Communication Framework Provisioning

Observation of Fig. 5 follows the analysis in Eq. 1, where an infrequent CS publication (300 s, 600 s) degrades the charging performance. This

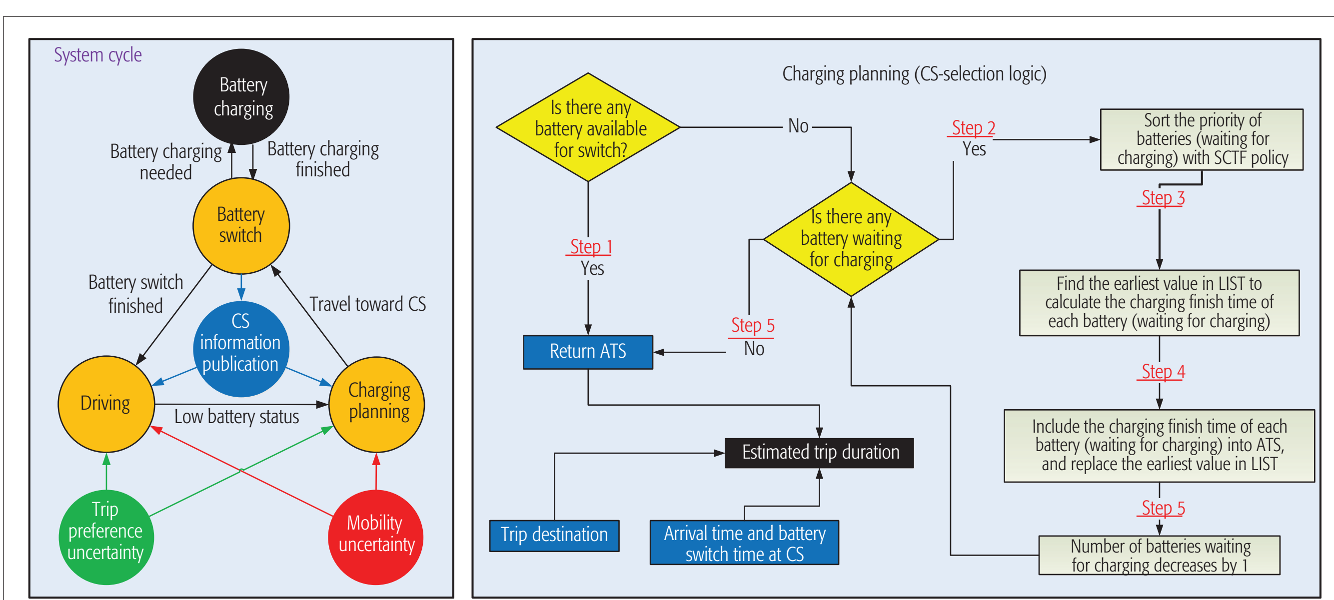


Figure 3. System cycle of on-the-move EV charging management via battery switch.

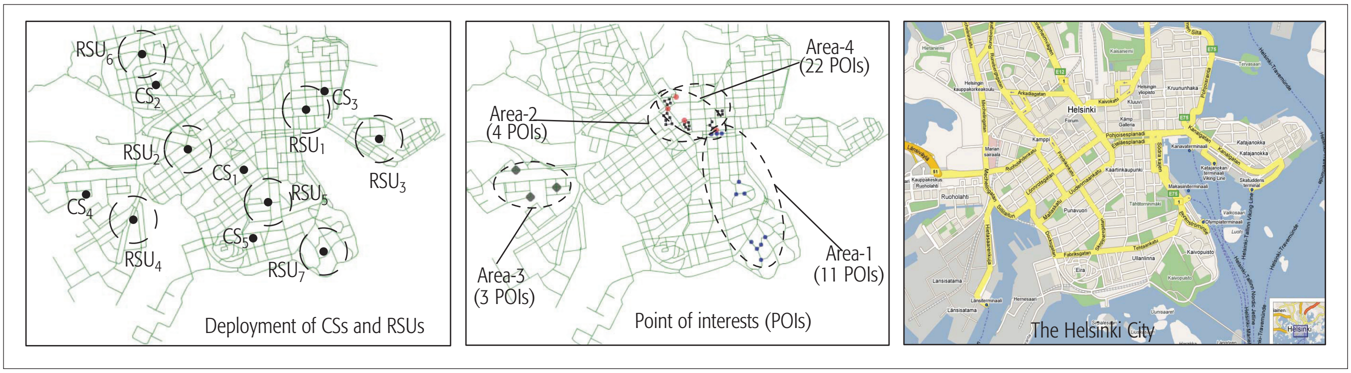


Figure 4. Simulation scenario of Helsinki city (5 CSs, 7 RSUs, 200 EVs).

is because the accuracy of ATS collected from CSs differs too much from the value when it is published, leading to suboptimal CS selection. Besides, setting a smaller communication range (100 m) and fewer RSUs (by removing RSU1, RSU2, and RSU5) also degrade performance. This is because EVs (far away from RSUs at a certain distance) cannot establish connections with RSUs, from which the ATS of CSs is less likely accessed.

Here, as decision making and information monitoring are simultaneous in CC, it has the best charging performance, but with the concern of high privacy sensitivity. Besides, PB always outperforms P/S, but at the expense of a higher number of accesses. By comparing the proposed P/S to [10], we observe a substantial improvement regarding number of accesses, thanks to the introduction of two dedicated topics involved in service discovery with RSUs.

The Influence of Urban Driving Uncertainties

Envisioning mobility uncertainty, 50 randomly generated traffic jams happen every 300 s, while its range is 300 m. Therefore, each EV will adjust its moving speed if the distance between its location and a traffic jam is smaller than 300 m. All traffic jams will last for 100 s since generation. Obviously, with an increased number of traffic jams in the city, the charging performance is inevitably degraded. This is because the uncertain arrival (due to speed reduction/stop) at CSs affects the accuracy of CS selection.

Envisioning trip preference uncertainty, we assign four types of POIs, of which the distribution influences the intentions of EVs' trip destinations. By setting potential trip preference, the intention-driven CS selection results in charging hotspots at those CSs around POIs (e.g., CS_1, CS_4, CS_5). Due to an increased number of EVs intending to switch batteries at these CSs, the average charging waiting time and trip duration are increased, while the distribution of charged EVs among CSs changes. We observe that the mobility uncertainty does not remarkably result in charging hotspots, in sharp contrast to that under the trip preference uncertainty.

The Influence of Charging System

Given the reduction of fully recharged batteries maintained at CSs, EVs have to wait for a longer time to get batteries switched, and their total trip duration is increased. Besides, the performance becomes worse, if charging depleted batteries using the first deplete first charge (FDFC) rather than STCF policy. This implies the importance of providing fast availability of switchable batteries.

Finally, the advantage of battery switching over plug-in charging technology is observed given 10 kW charging power. The former, with 60 kW charging power, can eventually achieve performance close to the latter. This reflects a realistic concern for which the battery switch system can alleviate the peak load in the power grid by running a lower charging power.

Open Research Issues

Heterogeneity of EVs

There have been many EV manufacturers, and each type of EV may only be compatible with one or a few types of batteries. Assuming each CS maintains different types of batteries, it then publishes the integrated information about distinguished ATS (in relation to a certain type of battery). Also, the underlying battery scheduling policy should adjust the availability of each type of battery, depending on demands from heterogeneous EVs.

Vehicle-to-Grid Operation

The CSs providing battery switch services can operate either as an energy source when their maintained batteries are fully charged, or as an electrical load when the depleted batteries need to be charged. The residual electricity of depleted batteries from EVs could be sold back to the grid. Instead, enabling the battery switch service in vehicle-to-grid (V2G) operation benefits both drivers and the grid due to short EV idle time at the CS as well as flexibility in controlling the depleted batteries.

The Business Model

Apart from strategically deploying CSs and V2G operation, the number of batteries to maintain at each CS should be decided prior to knowing the future demand for battery switching. Following our results, more batteries should be maintained at CSs that are in proximity to potential drivers' POIs. A dynamic pricing strategy to minimize congestion and maximize profit (by adjusting the switch price) is suggested.

Security

A malicious business may bombard an individual EV with unsolicited products or services, for example, attracting drivers using manipulated CS condition information. As such, security is required to maintain confidentiality, integrity, and availability

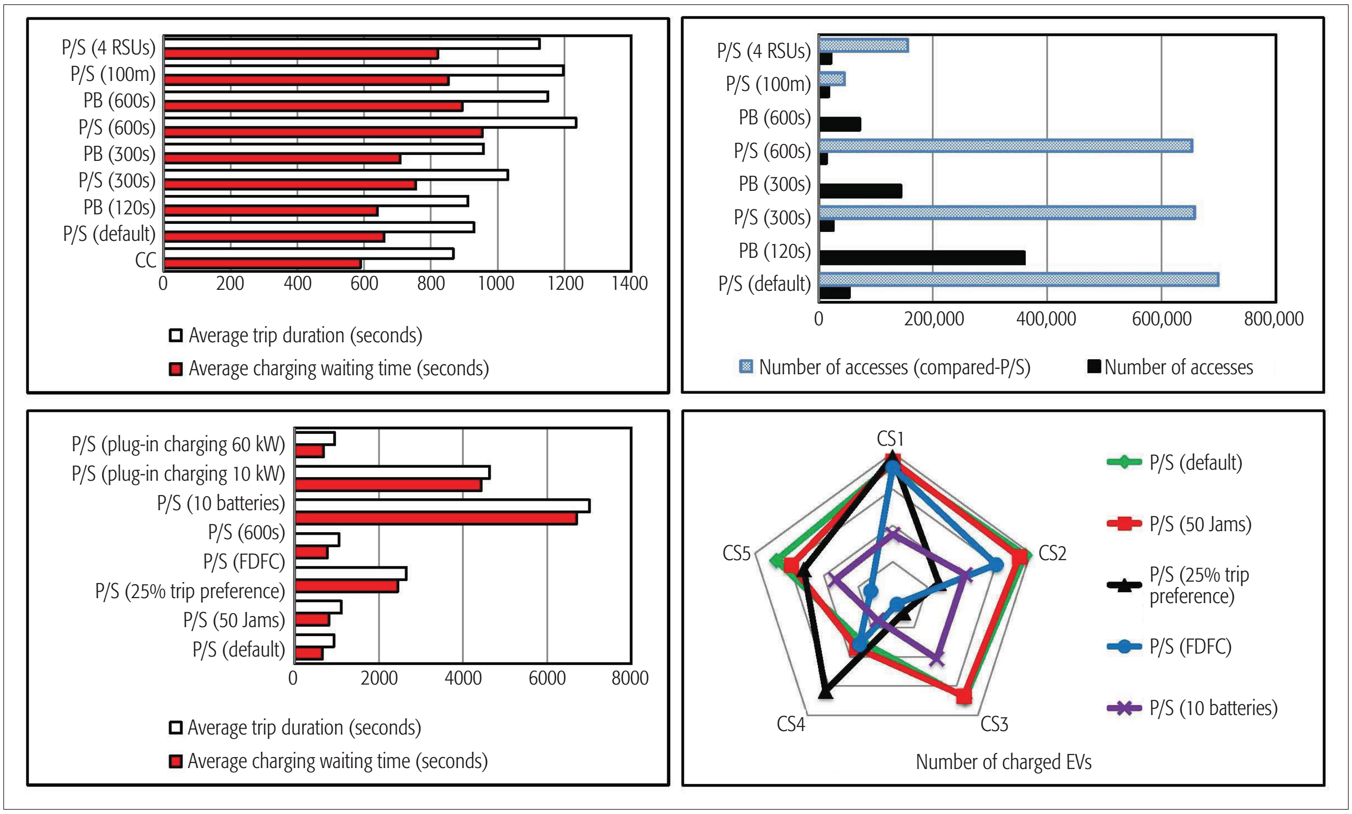


Figure 5. Evaluation results of case study.

of information exchange between CSs and EVs. The credibility of CSs' ATS is required for hazard-free decisions of EVs. Thus, all messages must be digitally signed by CSs and later verified by EVs before making their CS selection decisions.

Conclusion

The effort toward green communication and charging systems for EVs on the move has become important. Here, a cost-efficient P/S communication framework enabling battery switch service is provisioned to manage on-the-move EV charging, aiming to accelerate the charging process over traditional plug-in charging technology. Results of city-scale simulations demonstrate that a number of factors have an impact on the driving experience through battery switch services. Open research issues have also been discussed.

Acknowledgment

This work is supported by the NSFC 61372114 and NSFC 61631005, and the Beijing Nova Program Z151100000315077.

References

[1] A. Boulanger *et al.*, "Vehicle Electrification: Status and Issues," *Proc. IEEE*, vol. 99, no. 6, June 2011, pp. 1116–38.
[2] E. Rigas, S. Ramchurn, and N. Bassiliades, "Managing Electric Vehicles in the Smart Grid Using Artificial Intelligence: A Survey," *IEEE Trans. Intelligent Transportation Systems*, vol. 16, no. 4, Aug. 2015, pp. 1619–35.
[3] F. Pan *et al.*, "Locating PHEV Exchange Stations in V2G," *Proc. IEEE SmartGridComm '10*, MD, Oct. 2010.
[4] J. Gozalvez, M. Sepulcre, and R. Bauza, "IEEE 802.11p Vehicle to Infrastructure Communications in Urban Environments," *IEEE Commun. Mag.*, vol. 50, no. 5, May 2012, pp. 176–83.
[5] P. T. Eugster *et al.*, "The Many Faces of Publish/Subscribe," *ACM Computing Surveys*, vol. 35, no. 2, June 2003, pp. 114–31.
[6] J. C. Mukherjee and A. Gupta, "A Review of Charge Scheduling of Electric Vehicles in Smart Grid," *IEEE Systems J.*, vol. 9, no. 4, Dec. 2015, pp. 1541–53.
[7] R. Wang, P. Wang, and G. Xiao, "Two-Stage Mechanism for Massive Electric Vehicle Charging Involving Renewable Energy," *IEEE Trans. Vehicular Technology*, vol. 65, no. 6, June 2016, pp. 4159–71.
[8] W. Lee *et al.*, "Electric Vehicle Charging Stations with Renewable Power Generators: A Game Theoretical Analysis," *IEEE Trans. Smart Grid*, vol. 6, no. 2, 2015, pp. 608–17.
[9] T. Sweda and D. Klabjan, "Finding Minimum-Cost Paths for Electric Vehicles," *Proc. IEEE IEVC '12*, Mar. 2012.
[10] Y. Cao *et al.*, "An Electric Vehicle Charging Management Scheme Based on Publish/Subscribe Communication Framework," *IEEE Systems J.*, vol. PP, no. 99, 2015, pp. 1–14.
[11] Y. Cao and Z. Sun, "Routing in Delay/Disruption Tolerant Networks: A Taxonomy, Survey and Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 2, 2nd qtr., 2013, pp. 654–77.
[12] ETSI TS 101 556-1 v. 1.1.1, "Intelligent Transport Systems (ITS); Infrastructure to Vehicle Communication; Part 1: Electric Vehicle Charging Spot Notification Specification," tech. rep.
[13] M. Chiang and T. Zhang, "Fog and IoT: An Overview of Research Opportunities," *IEEE Internet of Things J.*, vol. 3, no. 6, 2016, pp. 854–64.
[14] O. Kaiwartya *et al.*, "T-MQM: Testbed-Based Multi-Metric Quality Measurement of Sensor Deployment for Precision Agriculture — A Case Study," *IEEE Sensors J.*, vol. 16, no. 23, Dec. 2016, pp. 8649–64.
[15] A. Keränen, J. Ott, and T. Kärkkäinen, "The ONE Simulator for DTN Protocol Evaluation," *Proc. ICST SIMUTools '09*, Rome, Italy, Mar., 2009.

Biographies

Yue Cao received his Ph.D. degree from the Institute for Communication Systems (ICS), formerly known as the Centre for Communication Systems Research, at the University of Surrey, Guildford, United Kingdom, in 2013. Further to his Ph.D. study, he was a research fellow at ICS. Since October 2016, he has been a lecturer in the Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, United Kingdom. His research interests focus on delay/disruption tolerant networks, e-mobility, information-centric networking, device-to-device communication, and mobile edge computing.

Shusen Yang [SM] received his Ph.D. in computing from Imperial College London in 2013. He is currently a professor at the

Institute of Information and System Science at Xi'an JiaoTong University (XJTU). Before joining XJTU, he was a lecturer at the University of Liverpool from 2015 to 2016, and a research associate at Intel Collaborative Research Institute (ICRI) from 2013 to 2014. He is a recipient of the China "1000 Young Talents Program" and an honorary research fellow at Imperial College London. He is a member of ACM.

GEYONG MIN is a professor of high performance computing and networking in the Department of Mathematics and Computer Science at the University of Exeter, United Kingdom. He received his Ph.D. degree in computing science from the University of Glasgow, United Kingdom, in 2003. His research interests include future Internet, computer networks, wireless communications, multimedia systems, information security, high performance computing, ubiquitous computing, modeling, and performance engineering.

XING ZHANG [SM] is a full professor with the School of Information and Communications Engineering, Beijing University of Posts and Telecommunications, China. His research interests are mainly in 5G wireless communications and networks, green communications, cognitive radio and cooperative communications, big data, and the Internet of Things. He is the author/coauthor of two technical books and more than 100 papers in top journals and international conferences, and has filed more than 30 patents. He has served as General Co-Chair of the 3rd IEEE International Conference on Smart Data (SmartData 2017), and as a TPC Co-Chair/TPC member for a number of major international conferences. He received the Best Paper Awards in the 9th International Conference on Communications and Networking in China, the 17th International Symposium on Wireless Personal Multimedia Communications, and the 8th IEEE International Conference on Wireless Communications and Signal Processing. He is a member of CCF.

HOUBING SONG received his Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, in August 2012. In August 2012, he joined the Department of Electrical and Computer Engineering, West Virginia University, Montgomery, where he is currently the Goden Bear Scholar, an assistant professor, and the founding director of the Security and Optimization for Networked Globe Laboratory (SONG Lab, www.SONGLab.us). His research interests lie in the areas of cyber-physical systems, the Internet of Things, edge computing, big data analytics, and communications and networking.

OMPRAKASH KAIWARTYA received his Ph.D. degree in computer science from the School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India, in 2015. He is currently a postdoctoral research fellow in the Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru. His research interests include vehicular ad hoc networks, mobile ad hoc networks, and wireless sensor networks.

NAUMAN ASLAM [M] is a reader in Department of Computer and Information Sciences at Northumbria University, Newcastle upon Tyne, United Kingdom. He received his Ph.D. in engineering mathematics from Dalhousie University, Halifax, Nova Scotia, Canada, in 2008. Prior to joining Northumbria University, he worked as an assistant professor at Dalhousie University from 2008 to 2011. Currently, he also holds an adjunct assistant professor position at Dalhousie University. His research interests include wireless ad hoc and sensor networks, fault-tolerant and reliable communication, and remote health monitoring application.

GREEN COMMUNICATIONS AND COMPUTING

On The Fundamental Energy Trade-offs of Geographical Load Balancing

Abbas Kiani and Nirwan Ansari

Geographical load balancing can optimize the utilization of green energy and the cost of electricity by taking the advantages of green and price diversities at geographical dispersed data centers. However, higher green energy utilization or lower electricity cost may actually increase the total energy consumption, and is not necessarily the best option.

ABSTRACT

Geographical load balancing can optimize the utilization of green energy and the cost of electricity by taking advantage of greenness and price diversities at geographically dispersed data centers. However, higher green energy utilization or lower electricity cost may actually increase the total energy consumption, and is not necessarily the best option. The achievable energy trade-offs can be captured by taking into consideration a defined service efficiency parameter for geo-dispersed data centers.

INTRODUCTION

The demand for online services including web search, online gaming, distributed file systems such as Google File System (GFS), and distributed storage systems such as BigTable and MapReduce is growing exponentially. This explosion of demand for online services has led to a multitude of challenges in data center networks (DCNs) from DCN architecture design, congestion notification, TCP Incast, virtual machine migration, to routing in DCNs [1].

Most importantly, data centers' electric power usage is growing at a rapid pace. In 2013, U.S. data centers consumed an estimated 91 billion kWh of electricity, and as the fastest growing consumer of electricity, they are estimated to consume roughly 140 billion kWh in 2020, which will cost $13 billion in electricity bills and emit 100 million metric tons of carbon pollution [2]. This huge average annual electricity consumption is due not only to the continuing explosion of Internet traffic but also to the gravity of preparing DCNs as a scalable and reliable computing infrastructure. Online services run on hundreds of thousands of servers spread across server farms provisioned for the peak load. In fact, to ensure user demand satisfaction, the servers run 24/7 and in vast underutilization the majority of the time. To put this in perspective, the total power consumption at a data center includes the base load and proportional load. The base load indicates the power consumption even when some of the turned on servers are idle. On the other hand, the proportional load is the extra power consumption that is proportional to the CPU utilization of the servers and accordingly to the load. Therefore, even idle, servers draw the base load power, thus incurring a substantial amount of annual energy use. However, in the past few years, more server capacities have been virtualized to facilitate multiple virtual machines (VMs) being run on a single physical machine (PM).

Complying with all of our online activities but limiting the increasing energy demand in an environmentally friendly manner calls for innovations across different disciplines. Recently, a great deal of research has been done to cut the data center's power consumption and accordingly the cost of electricity. A great part of the studies mainly aims to propose new power management techniques by investigating the CPU and memory power consumption of the servers. For examples, dynamic voltage/frequency scale (DVFS) schemes like [3] have been deployed to reduce the CPU power, and new techniques such as [4] have been proposed to adjust the power states of the memory devices in order to dynamically limit memory power consumption. However, the data center operators prefer to maintain a high level of reliability and uptime with their less expensive inefficient facilities rather than install energy-efficient devices at the cost of higher upfront cost [2].

Opportunities to improve data centers' energy efficiency are not limited to improvements in computing components. The energy consumption breakdown of data centers shows that a course of action is required to improve the energy consumption at other components like network equipment, electrical power delivery and conversion, cooling, and lighting. To this end, the power usage effectiveness (PUE) metric has been commonly adopted as a measure of data centers' efficiency, and is defined as the ratio of the total energy consumed by the data center to that consumed by the information technology (IT) equipment (EPA report on server and data center energy efficiency, Final Report to Congress, August 2007). Power delivery and cooling efficiency have been the subject of interest of many recent research papers, and a large number of studies have aimed at innovating networking components and topologies to shave the power consumed by IT network.

Another approach that addresses the energy consumption in all components is referred to as green data centers. The concept not only tries to cut down the electricity consumption and its cost but also integrates renewable energy resources such as solar panels and wind farms into data cen-

Digital Object Identifier: 10.1109/MCOM.2017.1600787

The authors are with the New Jersey Institute of Technology.

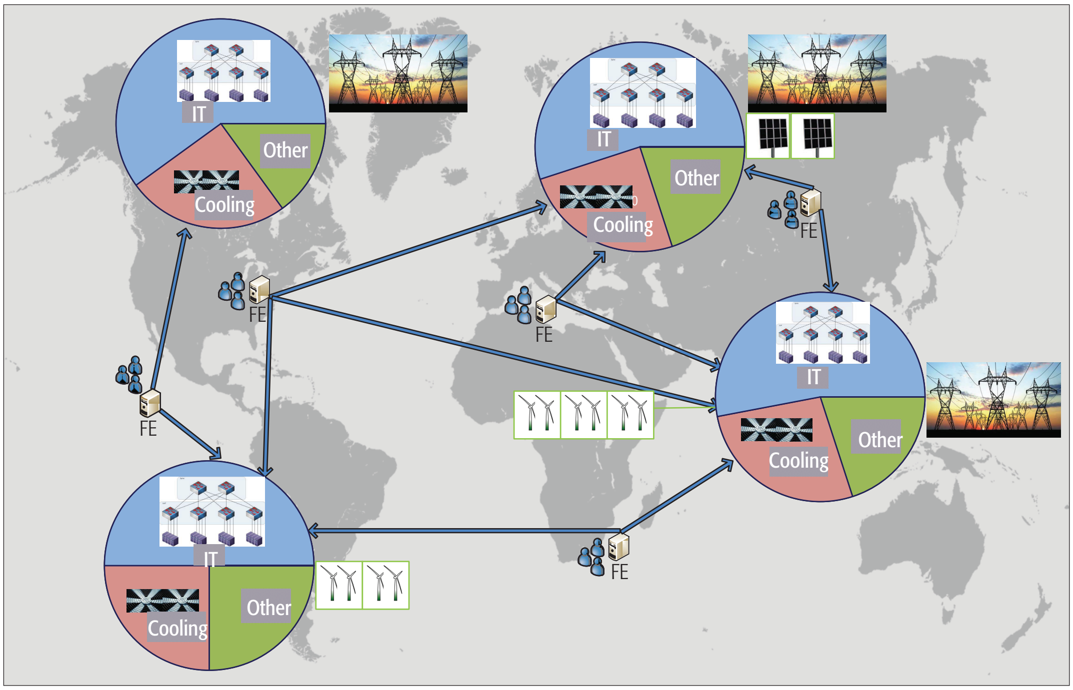


Figure 1. Geographical dispersed data centers.

The selection of the appropriate data centers can be based on different parameters like server or content availability, the network distance between FE and data center, the efficiency of the data centers, the cost of the electricity, and availability of the renewable energy.

ters, thereby promoting sustainability and green energy. Data center operators can assess the sustainability of their data centers using the carbon usage effectiveness (CUE) metric along with PUE. CUE is defined as the ratio of the total CO_2 emissions caused by the total data center energy consumption to IT equipment energy consumption. CUE has an ideal value of 0.0, which indicates that no carbon use is associated with the data center operations [5].

Shaving the energy consumption and its cost via load shedding and load shifting [6, references therein] is another approach. Load shedding is associated with quality of service (QoS) degradation where data centers, based on service level agreements (SLAs), decide to serve some types of the workload less effectively by utilizing less energy. On the other hand, load shifting algorithms investigate the possibility of shifting the load in time to run, for example, when cheaper electricity is available. Moreover, the effectiveness of the geographical load balancing on the energy costs and utilization of more renewable energy has been demonstrated in some studies. In so-called geographical load balancing (GLB), the workload is distributed among Internet scale data centers spread across geographical diversity [7]. In the following sections, we investigate the opportunities and challenges of geographical load balancing.

Green and Economic Geographical Load Balancing

Internet-scale powerful data centers are few because of the scale and cost of deployment and operation. These few data centers are generally dispersed in different geographical regions. The main goal of deploying such geo-dispersed data centers is not only to provide redundancy, scalability, and high availability but also to more efficiently employ global resources such as price diversity in electricity markets or location diversity in renewable power generation [7, 8] (Fig. 1). Therefore, the powerful data centers are generally deployed far away from a large majority of users. To this end, front-end (FE) servers are co-located with users. Each FE server receives requests from its nearby users and distributes the requests to the back-end servers at geo-dispersed data centers. In fact, each FE server functions as a workload distribution center that manages the workload by distributing user requests to the appropriate data centers.

The selection of the appropriate data centers can be based on different parameters like server or content availability, the network distance between FE and data center, the efficiency of the data centers, the cost of the electricity, and availability of renewable energy. Therefore, different workload distribution strategies can be adopted at each FE by considering different objectives like maximizing green energy utilization, minimizing the cost of electricity, or maximizing the profit gained by running DCNs. On the other hand, each service request has to be handled within a deadline determined by the SLA. Different parameters like the throughput of the connection between users and FE server, FE server and back-end servers at the data center, and the queuing and processing delay at the data center contribute to the end-to-end delay of a service request [9]. The QoS at a data center is generally ensured by implying an upper bound on the queuing delay at the data center, which has been commonly modeled as an M/GI/1 processor sharing (PS) queue or M/M/1 queue [7].

Green Diversity

To benefit from the energy efficiency and sustainability advantages of greening, data centers have recently been integrated with a green power source such as wind turbine or solar panel. There are three different ways to green a data center. The first approach, called behind the meter, is to install renewable power generators at the data center location. In this case, the data center operator can own the power generation system itself, or a third party can install the system and sell the generated power to the data center. However, the most efficient location to build a renewable power source is not always the same as the best location to build an efficient data center. Therefore, data center operators such as Google choose to either purchase renewable energy certificates (RECs) or make power purchase agreements (PPAs) to procure both power and RECs [10].

To maximize green energy utilization, one FE server can manage the distribution of its incoming workload to different data centers based on the availability of green energy. The available green energy at a data center can be determined by the green energy generation or storage at the data center. The generated on-site green energy at a data center can be predicted by taking into account the weather dependency of green energy. Specifically, when the renewable generator is a wind turbine, the prediction can rely on the foremost forecasting techniques, which are based on numeric weather prediction (NWP) of wind speed and power [11]. The prediction may include very-short-term forecasting, short-term forecasting, medium-term forecasting, and long-term forecasting techniques. If the case is solar generation, machine-learning-based prediction techniques can be employed. In the case of purchased green energy, although it is not possible to track the flows of green energy from grid, green energy generation can be estimated via a data center's RECs. Moreover, when extra green energy is available, each data center can store green energy at energy storage devices and draw the energy from the storage devices later.

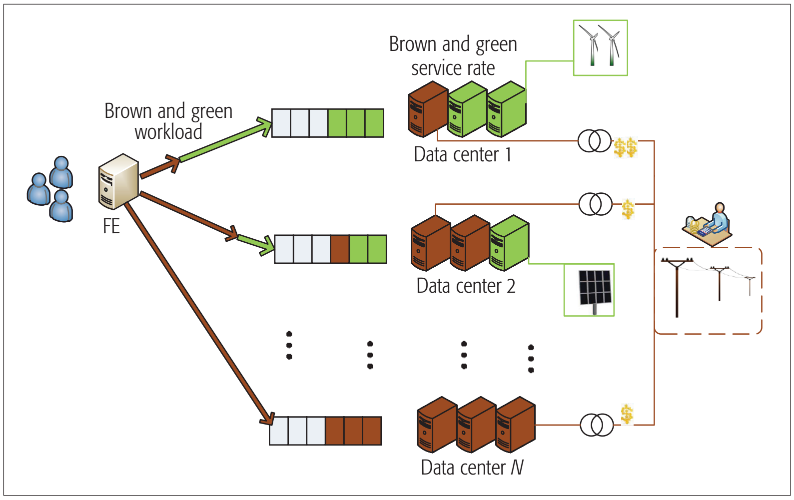


Figure 2. Green vs. brown.

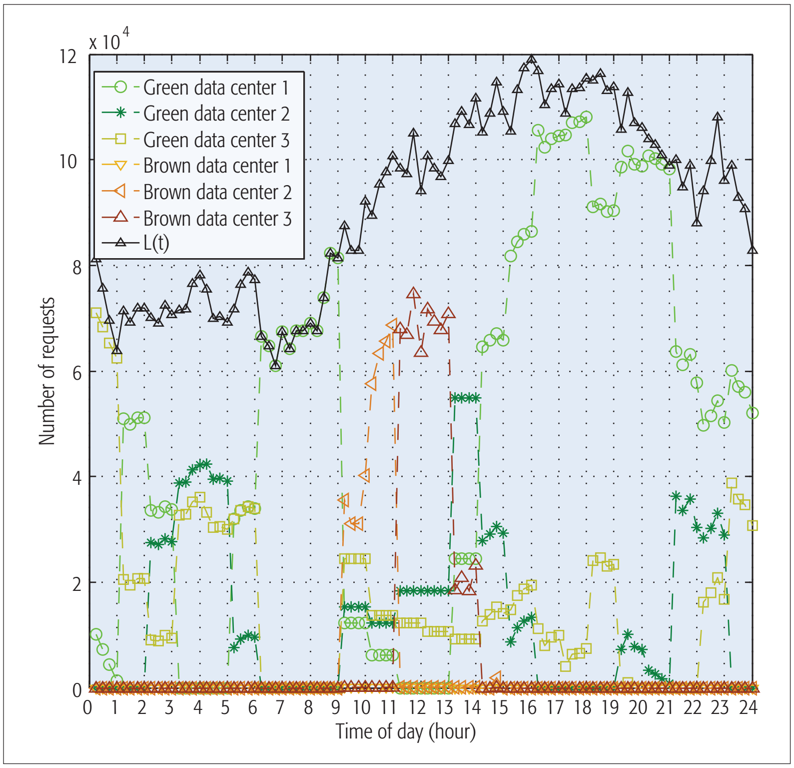


Figure 3. Total incoming workload and allocated green and brown workloads to the data centers.

Price-Diversity

While data centers operate 24/7, green energy is not a constantly available resource to power them. Therefore, data centers have to be connected to on-grid brown energy. In this case, we should note that brown energy is procured in deregulated electricity markets.

Unlike regulated electricity markets, in deregulated electricity markets such as day-ahead and real-time markets, the electricity prices vary during the day. The final prices are set based on the bidding process between the energy suppliers and consumers. Some studies also suggest that data centers can participate in the bidding process and procure electricity directly from the wholesale market [12]. However, the prices are not known to the data centers until operating time. For example, day-ahead prices are usually revealed several hours up to one day in advance, while real-time prices are known only a few minutes in advance. Therefore, the electricity price forecasting methods have to be employed when participating in bidding process.

GLB can be considered as an opportunity to reduce the cost of electricity by utilizing electricity price diversity at different locations. In other words, in order to minimize the electricity cost, each FE server can manage the workload by sending the requests to the data center locations with cheaper price of electricity.

Green vs. Brown

Green and price diversities are considered as an opportunity to design a green and low-cost GLB approach that not only can maximize the utilization of green energy but can also minimize the cost of electricity. However, due to the different costs and different environmental impacts of the renewable energy and brown energy, such a GLB approach should tap on the merits of the separation of the green energy utilization maximization and brown energy cost minimization problems, that is, the concept of decomposing the workload to the workloads served by green and brown energy [8]. In other words, it is the notion of green workload and green service rate vs. brown workload and brown service rate, respectively, to facilitate the separation of green energy utilization maximization and brown energy cost minimization problems.

The idea is to distinguish the servers at each data center based on the energy utilized to power them (Fig. 2). In fact, some servers are turned on and powered by the available green energy (green servers) and others if needed by purchasing brown energy (brown servers). Therefore, the distinction between green and brown workloads is made mainly based on the server utilized to serve the workload. Specifically, the workload served by a green server is defined as the green workload, and similarly the workload served by a brown server is defined as the brown workload. Moreover, using this idea, we can tackle the shortcoming in some studies, which propose an integrated optimization framework but under the assumption that local renewable generation is always less than the local power consumption. In fact, using the green vs. brown concept, each data center utilizes green energy as much as possible, and purchases brown energy only when the green energy generation is not adequate to serve all incoming workloads.

In terms of the optimization models, the optimization framework for green workload allocation has to maximize the utilization of green energy, and the constraints are to limit the allocated green workload by the available green resources. If the green energy is not adequate to satisfy the QoS requirements, another optimization framework is then designed to allocate the brown workload. While the objective of this optimization framework is to minimize the cost of electricity, the constraints are defined to enforce QoS requirements at each data center.

Figure 3 demonstrates the optimized allocated green and brown workload to $N = 3$ data centers. The simulation data are based on the trends of wind power at data centers 1, 2, and 3 shown in Fig. 4 and the electricity price used in [13]. We simulated the total incoming workload by scaling up the trends of a sample day of the requests made to the 1998 World Cup web site, which is also depicted in Fig. 3. In the simulations, we consider a time slotted system such that our optimization is applied in each time slot, and the QoS at each data center is enforced by maintaining an upper bound on our estimation of the data center's queue length at the next time slot.

For example, the trend of wind power indicates that after hour 14, more of the green workload is assigned to data center 1 where the highest wind power is available. However, from hours 10 to 13, the available wind power at data center 1 is lower than the other data centers, and thus less of the green workload is allocated to this data center. Moreover, as shown in Fig. 3, from hours 9 to 11, all of the leftover requests (brown workload) are allocated to data center 2, where the price of electricity is the lowest. Note that before hour 8 and from hours 15 to 24, the available wind power is adequate to serve all the service requests, and the brown workload is almost not allocated to the data centers. In other words, in these hours, the available wind power is the key decision factor to allocate workloads among the data centers.

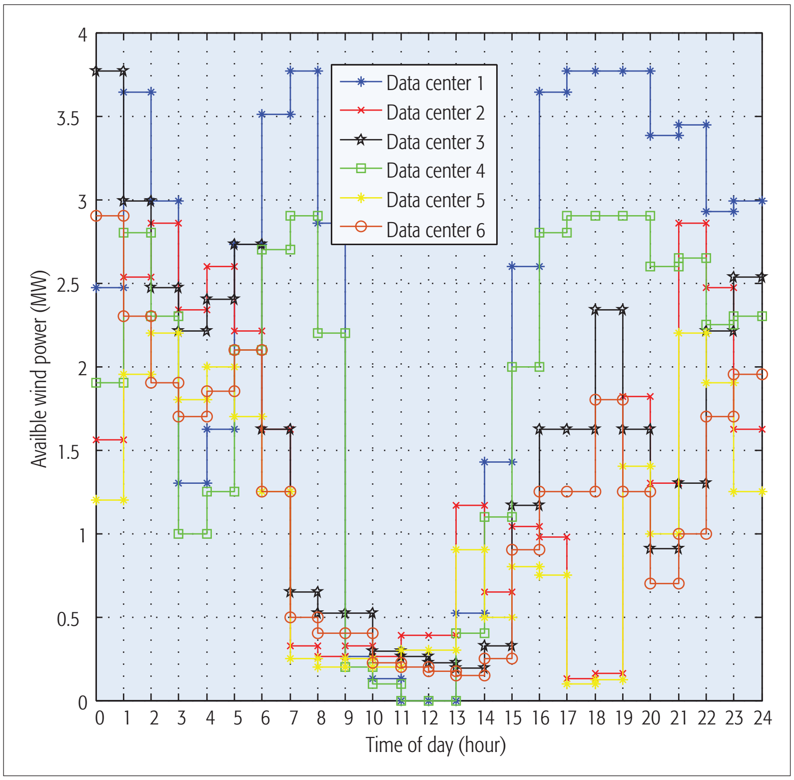


Figure 4. Wind power generation.

Fundamental Trade-offs

Most of the proposed GLB strategies aim at reducing the energy cost or brown energy consumption via distributing the requests to the locations with cheaper price of electricity or higher renewable energy generation. However, such strategies may increase the total power consumption due to the fact that different data centers have different servers with different service capabilities, and also a request sent to different data centers experiences different network delays. Therefore, consuming the same amount of energy or even more at one data center may handle fewer requests than another data center. In other words, the idea of sending a request to another data center with higher network delay or less service capability only in order to benefit from cheaper electricity or utilize more renewable energy may lead to a significant increase in total power consumption.

The extra green energy generation at a data center can be injected into the power grid, and the data center can receive compensation for the injected power. In the case of electricity, the cheap electricity at a data center can be stored at energy storage devices to be utilized later when the electricity becomes more expensive. Therefore, the higher green energy utilization or cheaper electricity at the expense of increasing the total energy consumption is not necessarily the best option.

Information Flow Graph-Based Model

To find the achievable trade-offs between total power consumption and green energy utilization, we propose to model geo-dispersed data centers with an information flow graph [14] (Fig. 5). Note that this idea may be adopted to capture the achievable trade-offs between total power consumption and the cost of electricity.

The information flow graph is a directed acyclic graph that includes three types of nodes:

- A single source node (S)
- Some intermediate nodes
- Data collector nodes [15]

As depicted in Fig. 5, an FE can be thought of as the source node, which is the source of original requests (FE node). Here, we only consider one FE, but the proposed model can be extended to multiple FEs as the geographically concentrated sources of requests. The intermediate nodes are also data centers, and the data collector node can correspond to users that receive processed requests.

The information flow graph, which models the geo-dispersed data centers, varies across time. At any given time, each node in the graph is either active or inactive. At the initial time of each time slot, the FE node as the only active node contacts all data center nodes and sorts them based on a defined service efficiency parameter. We define the service efficiency parameter based on an M/GI/1 processor sharing (PS) queue analysis by taking into consideration the network delay between FE and data centers. In other words, the allocated requests to a data center are first placed in a queue before they can be processed by any available server. We model the queue at each data center as an M/GI/1 PS queue, which has been commonly adopted in modeling the waiting time of the requests at a data center in many studies like [7]. The M/GI/1 PS queue is a single server queue with Poisson arrivals in which the service times form a sequence of i.i.d random variables with an arbitrary and general distribution function. Under PS, the processor is shared fairly among all jobs in the system. Therefore, the queuing delay at a data center can be computed as a function of the allocated requests to the data center, the service rate of a single server at the data center, and the total number of the turned on servers at the data center. The total number of turned on servers at a data center also depends on the power consumption of the data center, the average peak power of a turned on server in handling a service request, and the PUE of the data center [8, 12].

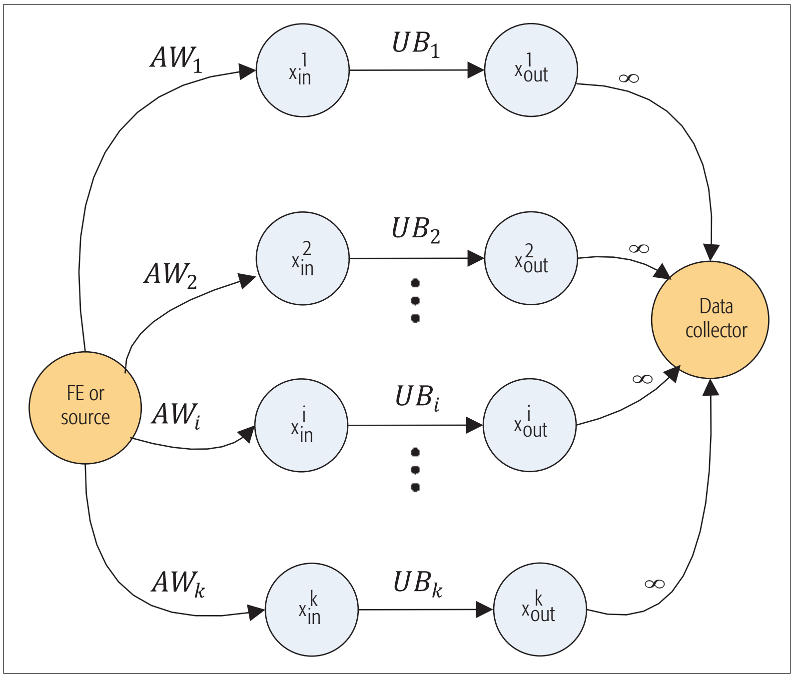


Figure 5. Information flow graph-based model for geo-dispersed data centers.

The FE node connects to a set of data center nodes with capacities of the edges equal to the allocated workloads (AWs) to these nodes. To satisfy the QoS requirements, the AW to a data center has to be upper bounded. In fact, the queueing delay for each service request should be limited by a given deadline determined by the SLA; this constraint is then translated to an upper bound (UB) for each data center. The value of this upper bound is varied at different data centers due to different service efficiency parameters.

Note that our trade-off parameters (total and brown power consumption) depend on the number of connected data center nodes to the FE node. In other words, connecting to a different number of data centers will result in a different amount of total and brown power consumption. After connecting to data center nodes, the FE node becomes and remains inactive, and selected data center nodes become active. Note that each data center node is represented by a pair of incoming and outgoing nodes, that is, x_{in} and x_{out}, connected by a directional edge whose capacity is the maximum number of requests that the data center can handle by the deadline (i.e., its UB). Finally, when the deadline comes, the data collector node becomes active and connects to the data center nodes to receive the processed requests. The edges that connect from the data center nodes to the data collector node are assumed to have infinite capacity, that is, users have access to all the processed requests.

Total-Brown Trade-off

The proposed information flow graph model can be used to capture the whole trade-off region between the total and brown power consumption. The main idea is that the FE has to distribute the workload to data centers such that the capacity of the FE-data collector minimum cut is larger than or equal to the total number of incoming requests at the FE. If so, the data collector node can ensure receiving all the processed requests by the deadline, and the workload distribution strategy can meet the SLA requirements. As a numerical example, we consider $k = 6$ data centers, each integrated with a wind farm as a renewable power source. Our simulation data are based on the trends of wind power and the total workload shown in Figs. 3 and 4, respectively. Figure 6 shows the trade-off curves between the total power consumption and green energy utilization at some sample hours of the day. For example, at time 10 a.m., if we utilize all available green energy at 6 data centers, that is, when the green energy utilization is 1, more than 35 MW of power is consumed totally. The trade-off curves in this figure confirm that we can increase the green energy utilization by increasing the total power consumption.

Summary

Due to the cost and scale, a few Internet-scale data centers are dispersed over the globe. To this end, geographical load balancing schemes are designed to leverage this geo-dispersion by efficient utilization of global resources such as price diversity in the electricity markets or locational diversity in renewable power generation, that is, distributing the requests to the locations with cheaper price of electricity or higher green power generation. Because of the different costs and different environmental impacts of the renewable energy and brown energy, geographical load balancing approaches can further benefit from the idea of green workload and green service rate vs. brown workload and brown service rate, respectively. In fact, the concept of decomposing the workload to the workloads served by green and brown energy facilitates the separation of green energy utilization maximization and brown energy cost minimization problems.

On the other hand, utilizing price and green energy diversities via geographical load balancing can increase the total power consumption due to the fact that different data centers have different servers with different service capabilities, and also a request sent to different data centers experiences different network delays. Therefore, there is a trade-off between total power consumption and green energy utilization or cost of the electricity. Such a trade-off can be captured by modeling the geo-dispersed data centers with an information flow graph and, more importantly, by defining a service efficiency parameter based on an M/GI/1 processor sharing queue to take into consideration the network delay between FE servers and data centers.

Extending the green vs. brown idea and information flow graph-based model by considering multiple workload distribution centers as the geographically concentrated sources of requests, the availability of energy storage devices at the data centers that may introduce new challenges like some battery related constraints, and elaborating on machine-learning-based energy prediction techniques will be an interesting future pursuit.

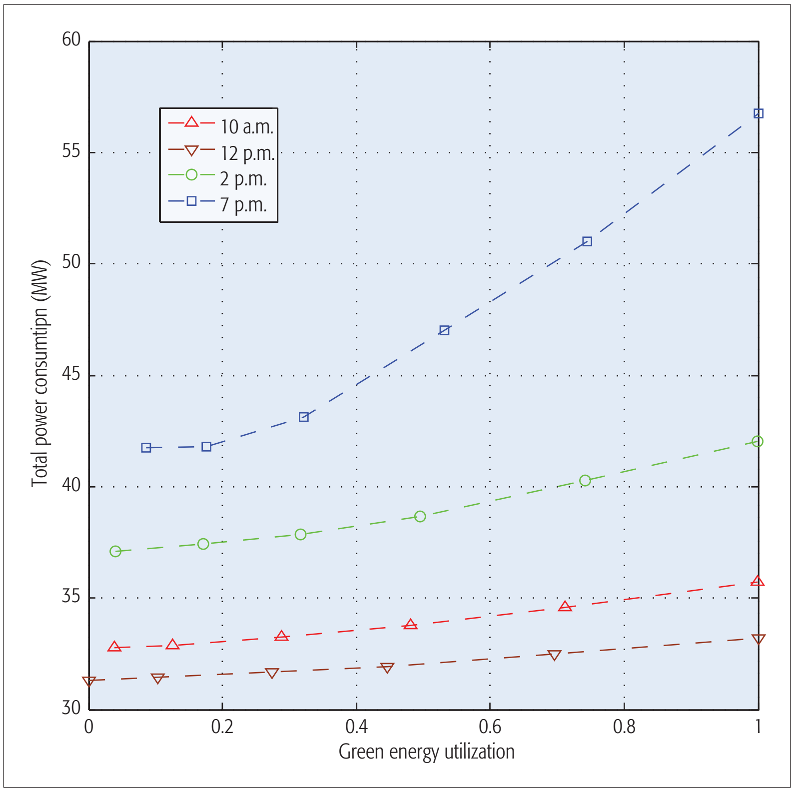


Figure 6. Green power utilization-total power consumption trade-off curves at different hours of day.

References

[1] Y. Zhang and N. Ansari, "On Architecture Design, Congestion Notification, TCP Incast and Power Consumption in Data Centers," *IEEE Communi. Surveys & Tutorials*, vol. 15, no. 1, 2013, pp. 39–64.

[2] P. Delforge, "America's Data Centers are Wasting Huge Amounts of Energy," Nat'l. Resources Defense Council, vol. Issue Brief 14-08-A, Aug. 2014.

[3] T. Pering, T. Burd, and R. Brodersen, "Dynamic Voltage Scaling and the Design of a Low-Power Microprocessor System," *Power Driven Microarchitecture Wksp.*, June 1998.

[4] B. Diniz et al., "Limiting the Power Consumption of Main Memory," *Proc. 34th Annual Symp. Computer Architecture*, San Diego, CA, 2007, pp. 290–301.

[5] C. Belady *et al.*, "Carbon Usage Effectiveness (CUE): A Green Grid Data Center Sustainability Metric," *The Green Grid White Paper*, 2010.

[6] A. Wierman *et al.*, "Opportunities and Challenges for Data Center Demand Response," *Proc. IEEE Green Computing Conf.*, 2014, pp. 1–10.

[7] Z. Liu *et al.*, "Greening Geographical Load Balancing," *IEEE/ACM Trans. Net.*, vol. 23, no. 2, 2015, pp. 657–71.

[8] A. Kiani and N. Ansari, "Towards Low-Cost Workload Distribution for Integrated Green Data Centers," *IEEE Commun. Letters*, vol. 19, no. 1, 2015, pp. 26–29.

[9] Y. Chen *et al.*, "Characterizing Roles of Front-End Servers in End-to-End Performance of Dynamic Content Distribution," *ACM Proc. 2011 ACM SIGCOMM Conf. Internet Measurement*, 2011, pp. 559–68.

[10] "Expanding Renewable Energy Options for Companies through Utility-Offered 'Renewable Energy Tariffs'," *Google White Paper*, April 19, 2013.

[11] S. S. Soman *et al.*, "A Review of Wind Power and Wind Speed Forecasting Methods with Different Time Horizons," *Proc. IEEE North American Power Symp.*, 2010, pp. 1–8.

[12] M. Ghamkhari, H. Mohsenian-Rad, and A. Wierman, "Optimal Riskaware Power Procurement for Data Centers in Day-Ahead and Real-Time Electricity Markets," *Proc. IEEE INFOCOM Wksps.*, Toronto, Canada, 2014.

[13] A. Kiani and N. Ansari, "Profit Maximization for Geographical Dispersed Green Data Centers," *IEEE Trans. Smart Grids*, DOI: 10.1109/TSG.2016.2562565, 2016.

[14] A. Kiani and N. Ansari, "A Fundamental Trade-Off between Total and Brown Power Consumption in Geographically Dispersed Data Centers," *IEEE Commun. Letters*, vol. 20, no. 10, 2016, pp. 19–55.

[15] A. G. Dimakis *et al.*, "Network Coding for Distributed Storage Systems," *IEEE Trans. Info. Theory*, vol. 56, no. 9, 2010, pp. 4539–51.

Biography

Abbas Kiani [S'14] received his B.Sc. degree in electrical engineering from Imam Khomeini International University, Qazvin, Iran, and his M.Sc. in communication engineering from Shahed University, Tehran, Iran. He is currently working toward a Ph.D. degree in electrical engineering at the New Jersey Institute of Technology, Newark. His research interests include data centers, cloud computing, network optimization, and green communications.

Nirwan Ansari [S'78, M'83, SM'94, F'09] is Distinguished Professor of Electrical and Computer Engineering at the New Jersey Institute of Technology. He wrote *Green Mobile Networks: A Networking Perspective* (Wiley-IEEE, 2017) with T. Han, and co-authored two other books. He has also (co-)authored more than 500 technical publications, over 200 in widely cited journals/magazines. He has guest-edited a number of Special Issues covering various emerging topics in communications and networking. He has served on the Editorial/Advisory Boards of over 10 journals. His current research focuses on green communications and networking, cloud computing, and various aspects of broadband networks. He was elected to serve on the IEEE Communications Society (ComSoc) Board of Governors as a Member-at-Large, has chaired ComSoc technical committees, and has actively organized numerous IEEE international conferences/symposia/workshops. He has frequently delivered keynote addresses, distinguished lectures, tutorials, and invited talks. Some of his recognitions include several Excellence in Teaching Awards, some best paper awards , the COMSOC TCGCC Distinguished Technical Achievement Recognition Award, the NCE Excellence in Research Award, the ComSoc AHSN TC Technical Recognition Award, the NJ Inventors Hall of Fame Inventor of the Year Award, the Thomas Alva Edison Patent Award, Purdue University Outstanding Electrical and Computer Engineer Award, and designation as a ComSoc Distinguished Lecturer. He has also been granted over 30 U.S. patents. He received a Ph.D. from Purdue University in 1988, an M.S.E.E. from the University of Michigan in 1983, and a B.S.E.E. (summa cum laude with a perfect GPA) from New Jersey Institute of Technologyå in 1982.

Multi-Method Data Delivery for Green Sensor-Cloud

Chunsheng Zhu, Victor C. M. Leung, Kun Wang, Laurence T. Yang, and Yan Zhang

The authors discuss the potential applications and recent work on SC, and observe two issues regarding green SC. Further, motivated by solving these two issues, they propose a multi-method data delivery scheme for SC users. MMDD strategically incorporates four kinds of delivery: delivery from cloud to SC users; delivery from WSN to SC users; delivery from SC users to SC users; and delivery from cloudlet to SC users.

Abstract

Delivering sensory data to users anytime and anywhere if there is network connection, sensor-cloud (SC), which integrates WSNs and cloud computing, is attracting growing interest from both academia and industry. This article discusses the potential applications and recent work about SC and observes two issues regarding green SC. Further, motivated by solving these two issues, this article proposes a Multi-Method Data Delivery (MMDD) scheme for SC users. MMDD strategically incorporates four kinds of delivery: delivery from cloud to SC users; delivery from WSN to SC users; delivery from SC users to SC users; and delivery from cloudlet to SC users. Compared to exclusive data delivery from cloud to SC users, evaluation results show that MMDD could achieve lower delivery cost or less delivery time for SC users.

Introduction

Recently, motivated by incorporating the ubiquitous data gathering ability of wireless sensor networks (WSNs) as well as the powerful data storage and data processing capabilities of cloud computing (CC), sensor-cloud (SC) [1] [2] is receiving growing attention from both the academic and industrial communities. Basically, integrating WSNs and CC, as shown in Fig. 1, the sensory data is gathered by the ubiquitous sensor nodes (e.g., temperature sensor nodes, humidity sensor nodes, motion sensor nodes) in WSNs and transmitted to the powerful data centers in the cloud. Then the sensory data is stored and processed by the cloud and further delivered to SC users on demand.

With such integration, from the perspective of users, SC enables them to obtain their required sensory data anytime and anywhere if there is network connection. From the view of WSNs and the cloud, SC complements them. For example, in the cloud, the service the cloud offers can be greatly enriched by providing the services (e.g., environmental monitoring, healthcare monitoring, landslide detection, forest fire detection) that WSN provides [3]. Regarding WSNs, the utility of WSNs could be enhanced, by being able to serve multiple applications via the cloud. Generally, in such integration, there are three main entities: a sensor network provider (SNP), which enables the WSN; the cloud service provider (CSP), which offers the cloud, and the SC user.

Discussing the potential applications and recent work regarding SC, this article observes two issues concerning green SC. Then, triggered by solving these two issues, this article proposes a Multi-Method Data Delivery (MMDD) mechanism for SC users. Particularly, MMDD strategically combines four kinds of delivery: delivery from the cloud to SC users; delivery from the WSN to SC users; delivery from SC users to SC users; and delivery from a cloudlet to SC users. In contrast to exclusive data delivery (EDD) from the cloud to SC users, evaluation results show that MMDD could obtain lower delivery loss or less delivery time for SC users.

For the rest of this article, the potential applications of SC are presented in the next section. Following that we review the recent work on SC and present the two issues regarding green SC. The proposed MMDD scheme is then introduced. We next perform the evaluation with respect to MMDD and EDD. This article is concluded in the final section.

Potential Applications of SC

SC has a lot of exciting potential applications [4]. For instance, concerning *real-time agriculture monitoring*, WSNs comprise a variety of sensor nodes (e.g., soil moisture sensor nodes, air sensor nodes, temperature sensor nodes, CO_2 concentration sensor nodes, and camera sensor nodes) that can be arranged to collect various information about the crops on a farm. These data can be further analyzed by the cloud in real time in order to track the health of the crops. With respect to *real-time transportation monitoring*, WSNs that include different kinds of sensor nodes (e.g., pressure sensor nodes, image sensor nodes, video sensor nodes, and alcohol gas sensor nodes) can be used for gathering vehicle and driver information. After the cloud incorporates the collected information in real time, the level of fuel and the vehicle arrival time as well as the status of the driver can be tracked, predicted, and observed, respectively. Regarding *real-time tunnel monitoring*, WSNs including light sensor nodes can be utilized to sense the light levels inside a tunnel. Meanwhile, the cloud can analyze the sensed light levels in real time so that the light intensity can be automatically adjusted to save energy spent unnecessarily for lighting throughout the day. About *real-time wildlife monitoring*, WSNs consisting of various types of sensor nodes (e.g., video sensor nodes) can be deployed into a wide

Digital Object Identifier:
10.1109/MCOM.2017.1600822

Chunsheng Zhu and Victor C. M. Leung are with The University of British Columbia; Kun Wang is with Nanjing University of Posts and Telecommunications; Laurence T. Yang is with St. Francis Xavier University. Yan Zhang is with the University of Oslo, 0373 Oslo.

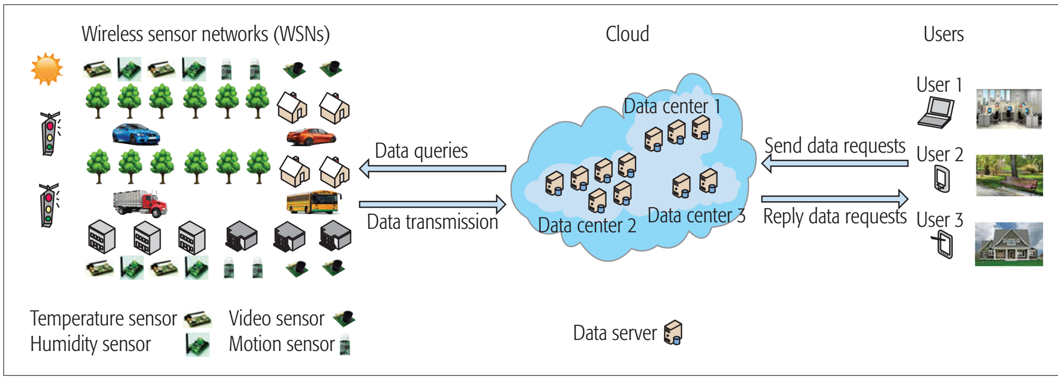


Figure 1. An instance of SC.

field, collecting information about the wildlife sanctuaries, activities, and so on. With the powerful cloud, which stores and processes the gathered information, real-time monitoring and further protecting the wildlife (e.g., endangered species) can be achieved.

Recent Work about SC

Focusing on *cloud-based WSNs*, the aim of [5] is enhancing the lifetime of the WSN integrated with the cloud. Specifically, two collaborative location-based sleep scheduling approaches are introduced for WSNs. The strategy is to dynamically determine the awake or asleep state of each sensor node to decrease energy consumption of the integrated WSN, considering the locations of mobile cloud users. With respect to the *WSN-based cloud*, the purpose of [6] is reducing the expected completion time for the CC integrated with WSN. Particularly, two job scheduling schemes (i.e., priority-based two-phase Min-Min and priority-based two-phase Max-Min) are described for CC. The technique executes WSN related cloud tasks in phase 1, while executing other ordinary cloud tasks in phase 2.

Concerning *SC integration*, a sensory data processing framework is shown in [7], aimed at transmitting desirable sensory data to the mobile cloud users in a fast, reliable, and secure manner. The mechanism incorporates the WSN gateway, the cloud gateway, the cloud, and the mobile users to perform various functions (e.g., data traffic monitoring, data filtering, data prediction, data recommendation, data compression, data decompression, data security). Another sensory data delivery scheme is presented in [8], toward offering more useful data reliably to the mobile cloud from a WSN. The idea is making the WSN gateway selectively transmit the sensory data to the cloud based on the time and priority features of the data requested by the mobile user, while utilizing the priority-based sleep scheduling to save the energy consumption of the WSN. An authenticated trust and reputation calculation and management system is exhibited in [9], targeted at helping a cloud service user (CSU) choose a desirable CSP and assisting the CSP in selecting an appropriate SNP. The scheme incorporates the authenticity of the CSP and SNP; the attribute requirement of the CSU and CSP; and the cost, trust, and reputation of the service of the CSP and SNP. A trust-assisted SC is designed in [10], devoting effort to improve the quality of service (QoS) of the sensory data experienced by SC users. The method utilizes trusted sensors (i.e., sensors with trust values surpassing a threshold) in the WSN for gathering and transmitting sensory data, while using trusted data centers (i.e., data centers with trust values surpassing a threshold) in the cloud for storing, processing, and delivering the sensory data to users. Five pricing models are devised in [11], induced by offering guidance for future research regarding SC pricing. The pricing designs consider the following factors: the lease period of the SC user; the required working time of SC; the SC resources utilized by the SC user; the volume of sensory data obtained by the SC user; and the SC path that transmits the sensory data from the WSN to the SC user, respectively.

To the best of our knowledge, the sensory data delivery of SC in all the above work is EDD from the cloud to SC users. With such delivery, the following two issues regarding green SC could exist.

Issue 1: Since SC users might request the same data from the cloud, the cloud might deliver large amounts of the same data to SC users. A large number of repeated data transmissions from the cloud to SC users exclusively increases the demand regarding the energy and resources as well as the bandwidth of SC.

Issue 2: When multiple SC users request data from the cloud simultaneously, a lot of data needs to be delivered from the cloud to multiple SC users at the same time. Substantial data delivery from the cloud to multiple SC users exclusively also increases the requirement with respect to the energy and resources as well as the bandwidth of SC.

In both of the above cases, we can observe that the delivery cost (e.g., utilized energy and resources as well as bandwidth) for providing data to SC users is increased. Furthermore, in terms of an SC with certain energy and resources as well as bandwidth, the delivery time for offering data to SC users is also increased.

The Proposed MMDD Scheme

Overview

Motivated by solving the above observed two issues, the MMDD scheme is proposed. Particularly, as shown in Fig. 2, the following four methods are incorporated by MMDD for delivering sensory data to SC users:

1. MMDD1 (delivery from cloud to SC users)
2. MMDD2 (delivery from WSN to SC users)

Two collaborative location-based sleep scheduling approaches are introduced for WSNs. The strategy is dynamically determining the awake or asleep state of each sensor node to decrease energy consumption of the integrated WSN, considering the locations of mobile cloud users.

In terms of a certain delivery method in MMDD, the delivery cost and delivery time probably are also various in different situations. Thus, the appropriate delivery method (s) in different conditions need (s) to be used, to better satisfy the SC user's requirement regarding delivery cost or delivery time.

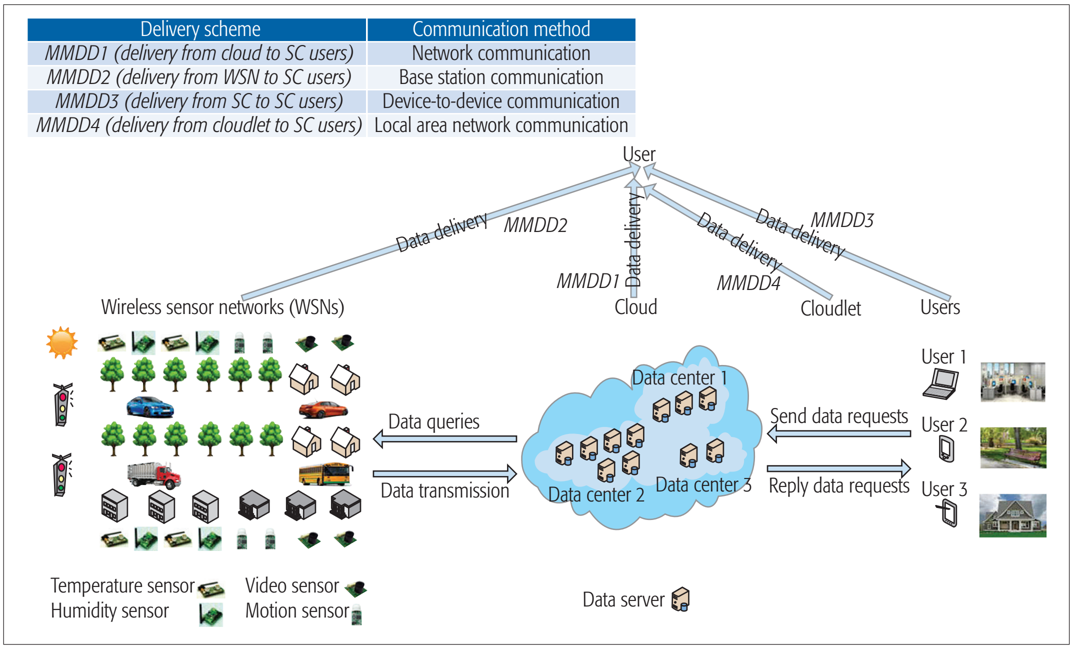


Figure 2. An instance of MMDD.

3. MMDD3 (delivery from SC users to SC users)
4. MMDD4 (delivery from cloudlet to SC users)

Regarding item 1, MMDD1 (delivery from cloud to SC users), the sensory data is delivered via the network communication, for example, wideband code-division multiple access (WCDMA), LTE, or WiMAX. About item 2, MMDD2 (delivery from WSN to SC users), the sensory data is delivered via base station communication. With respect to item 3, MMDD3 (delivery from SC users to SC users), the sensory data is delivered via device-to-device communication. For item 4, MMDD4 (delivery from cloudlet to SC users), the sensory data is delivered via the local area network communication. Here, available for use by nearby mobile devices, a cloudlet [12] is a resource-rich and trusted computer or cluster of computers well connected to the Internet.

Delivery Rules

The delivery rules consider the following elements:

- The delivery methods that are available for the SC user might be different when the SC user is in various locations.
- The sensory data requested by the SC user needs to be delivered through the delivery method(s).
- For each delivery method in MMDD, there is an associated delivery cost and delivery time.

Particularly, regarding different delivery methods in MMDD, the delivery cost and delivery time probably vary in different conditions. In terms of a certain delivery method in MMDD, the delivery cost and delivery time probably also vary in different situations. Thus, the appropriate delivery method(s) in different conditions need(s) to be used, to better satisfy the SC user's requirement regarding delivery cost or delivery time.

Specifically, the rules to utilize the delivery method(s) are shown as follows.

Rule 1: Based on the location of the SC user, the delivery method(s) available to the SC user is (are) determined by the SC.

Rule 2: Taking into account the data required by the SC user, the delivery method(s) that can deliver the needed sensory data is (are) decided by the SC.

Rule 3: Considering the delivery cost and delivery time of the delivery method(s), the specific delivery method(s) is (are) utilized by the SC based on the service level agreement of the SC user.

Regarding rule 1, the locations of SC users are obtained by SC through a mobile application that dynamically uploads the location of the SC user to SC [13]. About rule 2, if the sensory data required by the SC user needs to be delivered by multiple delivery methods, multiple delivery methods become the candidates. Otherwise, only one delivery method is used. With respect to rule 3, if the SC user wants to obtain the data with minimum delivery time, the delivery method(s) with the minimum delivery time is (are) utilized. If the SC user wants to achieve the data with a minimum delivery cost, the delivery method(s) with the minimum delivery cost is (are) used. Otherwise, the delivery method(s) with a satisfactory delivery time (i.e., the delivery time is less than a threshold) and a minimum delivery cost is (are) utilized, or the delivery method(s) with a satisfactory delivery cost (i.e., the delivery cost is less than a threshold) and a minimum delivery time is (are) used.

Delivery and Application Analysis

The purpose of MMDD is to better cater for the SC user's delivery requirement, since the delivery cost or delivery time with EDD might not satisfy the SC user's delivery requirement in some cases. For example, for issues 1 and 2 discussed above, in the case that SC users request the same data from the cloud or multiple SC users request the data from the cloud simultaneously, the data could be offered to the SC users by intelligently utilizing different delivery method (s) in MMDD, instead of EDD from the cloud to SC users. Since

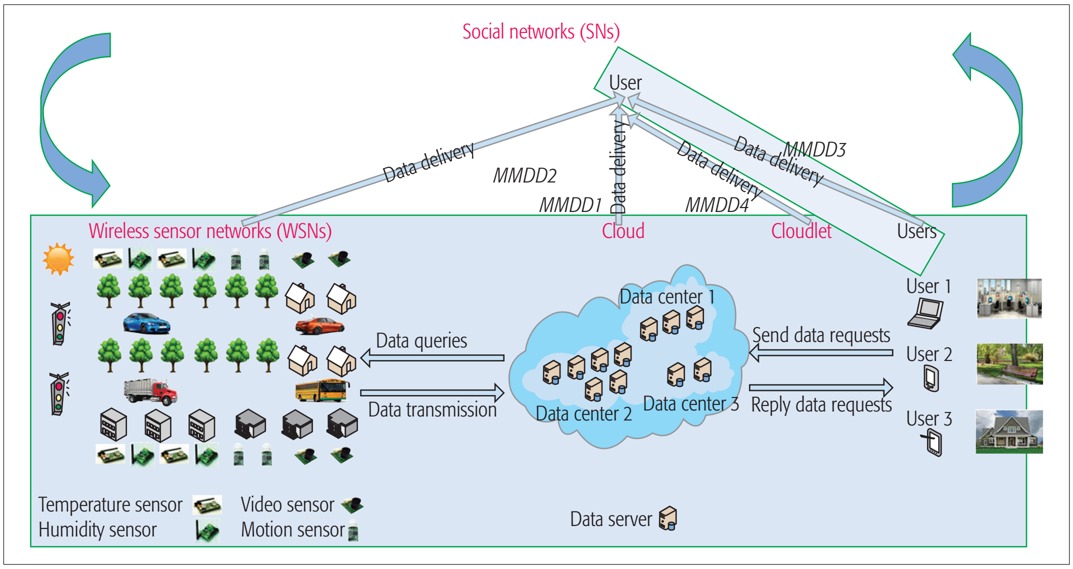


Figure 3. MMDD application in SSC.

The aim of MMDD is to strategically incorporate the four delivery methods to generate lower delivery cost or less delivery time for the SC user, while offering the needed data to the SC user. As a result, different MMDD (s) might be utilized for delivering the sensory data to the SC user in different conditions.

different delivery methods in MMDD have varying delivery cost and delivery time in different conditions, the delivery cost for providing data to SC users with MMDD might be lower than that with EDD. Similarly, the delivery time for offering data to SC users with MMDD might be less than that with EDD.

In other words, the aim of MMDD is to strategically incorporate the four delivery methods (i.e., MMDD1, MMDD2, MMDD3, and MMDD4) to generate lower delivery cost or less delivery time for the SC user, while offering the needed data to the SC user. As a result, different MMDDs might be utilized for delivering the sensory data to the SC user in different conditions. For instance, MMDD4 might be used when the SC user is in a coffee shop in which a cloudlet could offer the sensory data, while MMDD3 probably could be used when the SC user is in a classroom where other SC users could provide the sensory data. MMDD2 might be utilized when the SC user is very close to the WSN offering the sensory data, while MMDD1 probably could be utilized when the SC user is very far from home and the cloud could provide the sensory data. Moreover, multiple delivery methods in MMDD could be used if one delivery method (e.g., MMDD4) is not sufficient to offer all the sensory data that the SC user requests.

Furthermore, regarding MMDD, it could also be utilized in the future *social sensor cloud* (SSC) [14]. As shown in Fig. 3 about SSC, the SC users form social networks (SNs) [15], which connect and complement the WSNs, cloud, and cloudlet. In particular, SNs are networks formed by social members (e.g., individuals, organizations) who interact with each other. With SNs, the resources and services of the WSNs, cloud, and cloudlet could be shared, leveraging the relationships established between members of an SN. In such a manner, the resources and services requested by the SC users could be substantially reduced. Then the delivery cost and delivery time with MMDD might be further decreased. With further decreased delivery cost or delivery time, the SC user's delivery requirement could be even better satisfied.

Evaluation

Evaluation Setup

To evaluate the delivery time and delivery cost of EDD and MMDD for SC, the following two evaluations are performed as case studies. In both evaluations, there is an SC consisting of a WSN, a cloud, a cloudlet, and a number of SC users. In terms of EDD, the sensory data is delivered to the SC users from the cloud exclusively. For MMDD, the sensory data is delivered to the SC users cooperatively with MMDD1, MMDD2, MMDD3, and MMDD4.

For *evaluation 1*, it is assumed that the delivery times to an SC user with MMDD1, MMDD2, MMDD3, and MMDD4 are 0.4 s, 0.3 s, 0.2 s, and 0.1 s, respectively. The delivery cost to an SC user with MMDD1, MMDD2, MMDD3, and MMDD4 are US$0.8, US$0.6, US$0.4, and US$0.2, respectively. Evaluation 1 is conducted in the two scenarios below.

Scenario 1: MMDD1, MMDD2, MMDD3, and MMDD4 are each used to serve 25 percent of SC users, respectively. The number of SC users ranges from 100 to 1000 (increased by units of 100). This scenario is for evaluating the impacts of the number of SC users on SC's delivery time and SC's delivery cost with EDD and MMDD.

Scenario 2: There are 500 SC users. The percentage of SC users MMDD1 serves ranges from 10 to 90 percent (increased by units of 10 percent). MMDD2, MMDD3, and MMDD4 are utilized to serve the remaining users equally. This scenario is to evaluate the impacts of the utilization rate of MMDD1 on SC's delivery time and SC's delivery cost with EDD and MMDD.

Concerning *evaluation 2*, there are 500 SC users, served by MMDD1, MMDD2, MMDD3, and MMDD4 equally. Moreover, it is given that the delivery times to an SC user with MMDD3 and MMDD4 are 0.4 s and 0.2 s, respectively. The delivery cost to an SC user with MMDD3 and MMDD4 are US$0.8 and US$0.4, respectively. Evaluation 2 is implemented in the following two scenarios.

Scenario 3: The delivery time and delivery cost to an SC user with MMDD1 are 0.8 ss and

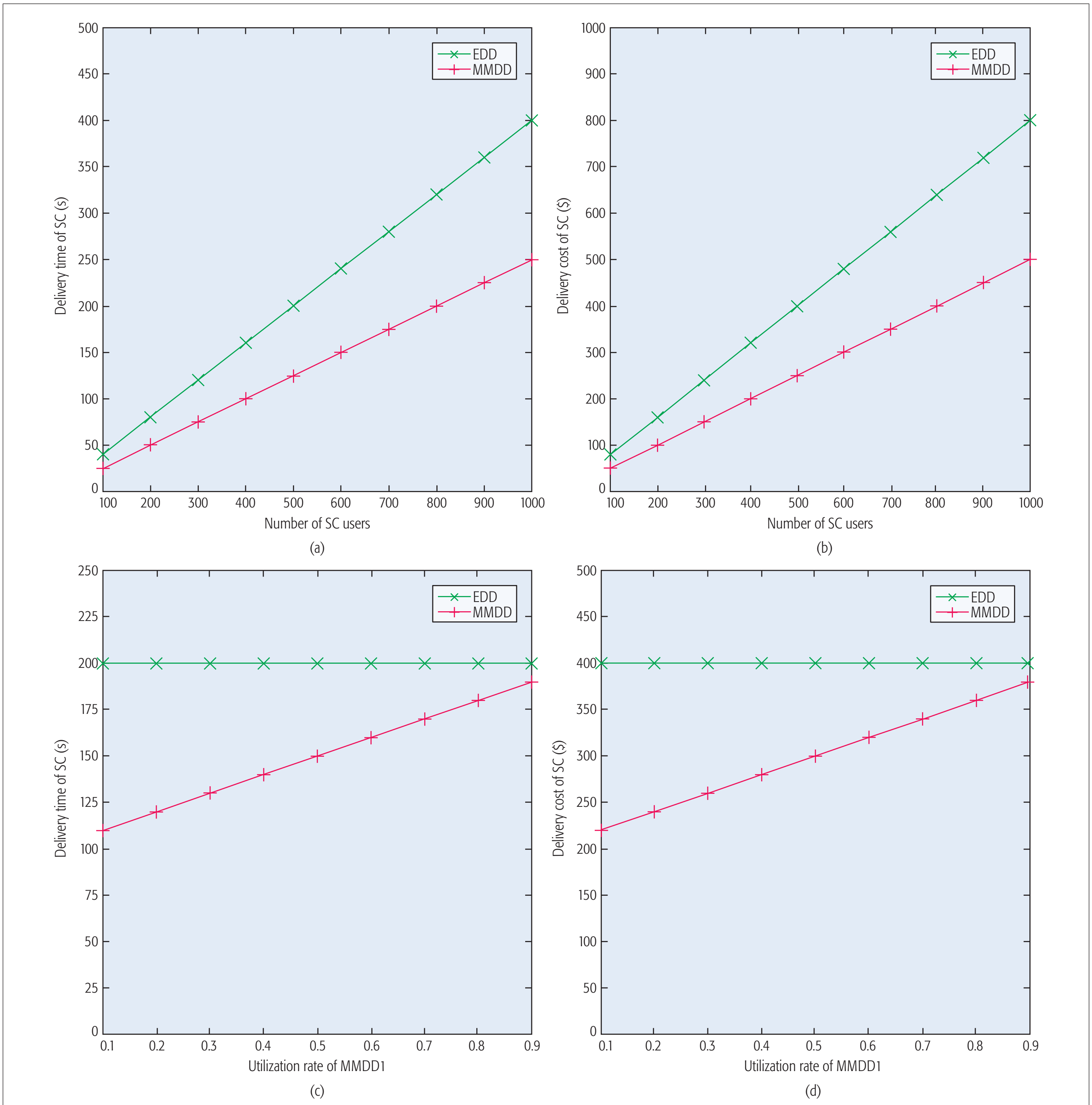


Figure 4. Evaluation 1: comparison of a) delivery time and b) delivery cost b) with EDD and MMDD in scenario 1; comparison of c) delivery time and d) delivery cost with EDD and MMDD in scenario 2.

US$1.6, respectively. The delivery time to an SC user with MMDD2 ranges from 0.6 to 1.4 s (increased in units of 0.1 s). The delivery cost to an SC user with MMDD2 ranges from US$1.2 to US$2.0, increased by units of US$0.1). This scenario is for evaluating the impact of MMDD2's delivery time on SC's delivery time and the impact of MMDD2's delivery cost on SC's delivery cost with EDD and MMDD.

Scenario 4: The delivery time and delivery cost to an SC user with MMDD2 are 0.6 s and US$1.2, respectively. The delivery time to an SC user with MMDD1 ranges from 0.8 to 1.6 s (increased by units of 0.1 s). The delivery cost to an SC user with MMDD1 ranges from US$1.6 to US$2.4 (increased by units of US$0.1). This scenario is to evaluate the impact of MMDD1's delivery time on SC's delivery time and the impact of MMDD1's delivery cost on SC's delivery cost with EDD and MMDD.

Evaluation Results

Figures 4 and 5 present the evaluation 1 results and evaluation 2 results about the delivery time and delivery cost with EDD and MMDD for SC in the two different scenarios, respectively. As shown in these figures, it can be clearly obtained that the delivery time or delivery cost of MMDD is better than that of

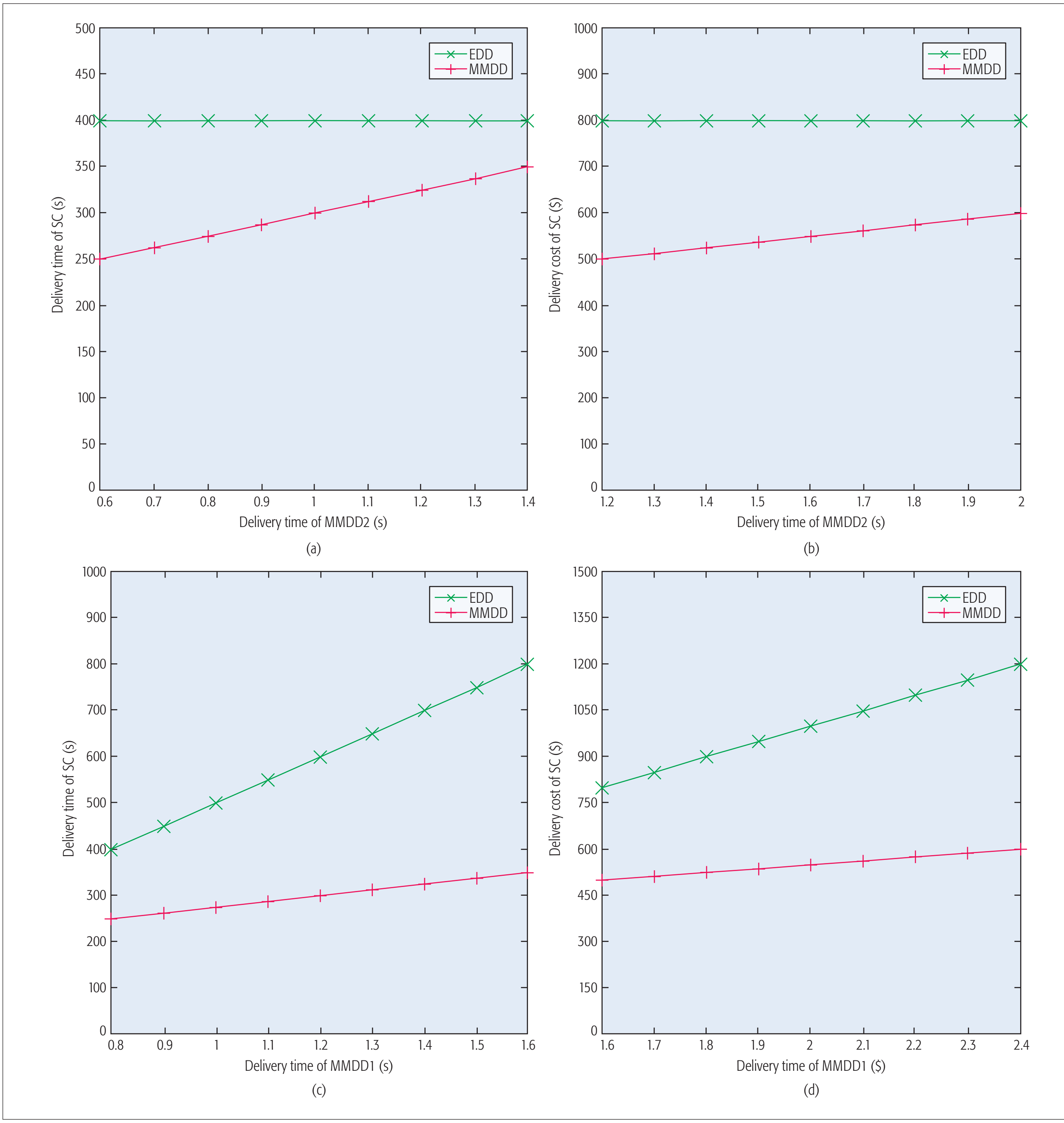


Figure 5. Evaluation 2: comparison of a) delivery time and b) delivery cost with EDD and MMDD in scenario 3; comparison of c) delivery time and d) delivery cost with EDD and MMDD in scenario 4.

EDD in terms of the above case studies. Particularly, from Figs. 4a and 4c and 5a and 5c, it can be observed that the SC's delivery times with MMDD are always less than those with EDD in scenarios 1, 2, 3, and 4. In addition, based on Figs. 4b and 4d and 5b and 5d, it can be perceived that the SC's delivery costs with MMDD are also always lower than those with EDD in all four scenarios.

Conclusion

Attracting growing interest from both the academic and industrial communities by integrating WSNs and CC, SC delivers sensory data to users anytime and anywhere if there is network connection. In this article, the potential applications and recent work with respect to SC have been discussed, and two research issues about green SC have been identified. Further, the MMDD scheme has been proposed, induced by solving the observed two research issues. Specifically, the following four methods that deliver sensory data to SC users are strategically incorporated in MMDD:

- Delivery from cloud to SC users
- Delivery from WSN to SC users
- Delivery from SC users to SC users
- Delivery from cloudlet to SC users

Evaluation results have also been presented about

With SNs, the resources and services of the WSNs, cloud and cloudlet could be shared, leveraging the relationships established between members of a SN. In such a manner, the resources and services requested by the SC users could be substantially reduced. Then the delivery cost and delivery time with MMDD might be further decreased.

MMDD and EDD, demonstrating that MMDD could achieve lower delivery cost or less delivery time for SC users.

Acknowledgment

This work was partially supported by a Four Year Doctoral Fellowship from The University of British Columbia and funding from the Natural Sciences and Engineering Research Council of Canada, the ICICS/TELUS People & Planet Friendly Home Initiative at The University of British Columbia, TELUS and other industry partners. This work was partially supported by the National Natural Science Foundation of China (NSFC) under Grant 61572262, the NSF of Jiangsu Province under Grant BK20141427. This work was partially supported by the project IoTSec - Security in IoT for Smart Grids, with number 248113/O70 part of the IKTPLUSS program funded by the Norwegian Research Council. This research is partially supported by the projects 240079/F20 funded by the Research Council of Norway.

References

[1] C. Zhu *et al.*, "Sensor-Cloud and Power Line Communication: Recent Developments and Integration," *Proc. 14th IEEE Int'l. Conf. Depend., Autonomic, Secure Comp.*, 2016, pp. 302–08.

[2] C. Zhu *et al.*, "Towards Integration of Wireless Sensor Networks and Cloud Computing," *Proc. 7th IEEE Int'l. Conf. Cloud Comp. Tech. Sci.*, 2015, pp. 491–94.

[3] C. Zhu *et al.*, "Insights of Top-K Query in Duty-Cycled Wireless Sensor Networks," *IEEE Trans. Ind. Electron.*, vol. 62, no. 2, Feb. 2015, pp. 1317–28.

[4] A. Alamri *et al.*, "A Survey on Sensor-Cloud: Architecture, Applications, and Approaches," *Int'l. J. Distrib. Sensor Net.*, vol. 9, no. 2, Feb. 2013, pp. 1–18.

[5] C. Zhu *et al.*, "Collaborative Locationbased Sleep Scheduling for Wireless Sensor Networks Integrated with Mobile Cloud Computing," *IEEE Trans. Comp.*, vol. 64, no. 7, July 2015, pp. 1844–56.

[6] C. Zhu *et al.*, "Job Scheduling for Cloud Computing Integrated with Wireless Sensor Network," *Proc. 6th IEEE Int'l. Conf. Cloud Comp. Tech. Sci.*, 2014, pp. 62–69.

[7] C. Zhu *et al.*, "A Novel Sensory Data Processing Framework to Integrate Sensor Networks with Mobile Cloud," *IEEE Sys. J.*, vol. 10, no. 3, Sept. 2016, pp. 1125–36.

[8] C. Zhu *et al.*, "Towards Offering More Useful Data Reliably to Mobile Cloud from Wireless Sensor Network," *IEEE Trans. Emerging Topics Comp.*, vol. 3, no. 1, Mar. 2015, pp. 84–94.

[9] C. Zhu *et al.*, "An Authenticated Trust and Reputation Calculation and Management System for Cloud and Sensor Networks Integration," *IEEE Trans. Info. Forensics Security*, vol. 10, no. 1, Jan. 2015, pp. 118–31.

[10] C. Zhu *et al.*, "Trust Assistance in Sensor-Cloud," *Proc. IEEE Conf. Comp. Commun. Wksps.*, 2015, pp. 342–47.

[11] C. Zhu *et al.*, "Pricing Models for Sensor-Cloud," *Proc. 7th IEEE Int'l. Conf. Cloud Comp. Tech. Sci.*, 2015, pp. 454–57.

[12] M. Satyanarayanan *et al.*, "The Case For VM-Based Cloudlets in Mobile Computing," *IEEE Pervasive Comp.*, vol. 8, no. 4, Oct.-Dec. 2009, pp. 14–23.

[13] G. Ananthanarayanan *et al.*, "Startrack: A Framework for Enabling Track-Based Applications," *Proc. 7th Int'l. Conf. Mob. Sys., Appl., Serv.*, 2009, pp. 207–20.

[14] C. Zhu *et al.*, "Green Internet of Things for Smart World," *IEEE Access*, vol. 3, Nov. 2015, pp. 2151–62.

[15] Y. Jiang and J. C. Jiang, "Understanding Social Networks from a Multiagent Perspective," *IEEE Trans. Parallel Distrib. Sys.*, vol. 25, no. 10, Oct. 2014, pp. 2743–59.

Biographies

Chunsheng Zhu is a postdoctoral research fellow in the Department of Electrical and Computer Engineering, University of British Columbia, Canada. He received his Ph.D. degree in electrical and computer engineering from the University of British Columbia in 2016. His current research interests mainly include wireless sensor networks, cloud computing, Internet of Things, social networks, and security.

Victor C. M. Leung [S'75, M'89, SM'97, F'03] is a professor in the Department of Electrical and Computer Engineering and holder of the TELUS Mobility Research Chair, University of British Columbia. His research is in the areas of wireless networks and mobile systems. He is a Fellow of the Royal Society of Canada, a Fellow of the Canadian Academy of Engineering, and a Fellow of the Engineering Institute of Canada.

Kun Wang [M'13] is an associate professor in the School of Internet of Things, Nanjing University of Posts and Telecommunications, China. He received his Ph.D. degree from the School of Computing, Nanjing University of Posts and Telecommunications, in 2009. In 2016, he was a research fellow with the School of Computer Science and Engineering, University of Aizu, Fukushima, Japan. His current research interests include information security, ubiquitous computing, and wireless communications technologies.

Laurence T. Yang [M'97, SM'15] is a professor in the Department of Computer Science, St. Francis Xavier University, Canada. His research interests include parallel and distributed computing, embedded and ubiquitous/pervasive computing, and big data. He has published more than 220 papers in various refereed journals (around 40 percent in top IEEE/ACM transactions and journals). His research has been supported by the National Sciences and Engineering Research Council, and the Canada Foundation for Innovation.

Yan Zhang [SM'10] is a professor in the Department of Informatics, University of Oslo, Norway. He is also a chief research scientist at Simula Research Laboratory, Norway. He is an Associate Technical Editor of *IEEE Communications Magazine*, an Editor of *IEEE Transactions on Green Communications and Networking*, and an Editor of *IEEE Communications Surveys & Tutorials*. His current research interests include next-generation wireless networks leading to 5G, green, and secure cyber-physical systems.

Energy Efficiency Challenges of 5G Small Cell Networks

Xiaohu Ge, Jing Yang, Hamid Gharavi, and Yang Sun

The authors investigate the computation power based on the Landauer principle. Simulation results reveal that more than 50 percent of the energy is consumed by the computation power at 5G small cell BSs. Moreover, the computation power of a 5G small cell BS can approach 800 W when massive MIMO (e.g., 128 antennas) is deployed to transmit high volume traffic.

Abstract

The deployment of a large number of small cells poses new challenges to energy efficiency, which has often been ignored in 5G cellular networks. While massive MIMO will reduce the transmission power at the expense of higher computational cost, the question remains as to which (computation or transmission power) is more important in the energy efficiency of 5G small cell networks. Thus, the main objective in this article is to investigate the computation power based on the Landauer principle. Simulation results reveal that more than 50 percent of the energy is consumed by the computation power at 5G small cell BSs. Moreover, the computation power of a 5G small cell BS can approach 800 W when massive MIMO (e.g., 128 antennas) is deployed to transmit high volume traffic. This clearly indicates that computation power optimization can play a major role in the energy efficiency of small cell networks.

Introduction

With the anticipated high traffic, small cell networks are emerging as an inevitable solution for fifth generation (5G) cellular networks [1]. In particular, the massive multiple-input multiple-output (MIMO) and millimeter-wave technologies are expected to be deployed toward improving the transmission rate and reducing the transmission power of 5G mobile communication systems [2]. On the other hand, more computation power will be required to process anticipated heavy traffic at small cell base stations (BSs). Under these conditions, a trade-off between computation and transmission power needs to be thoroughly evaluated in order to achieve energy efficiency optimization for 5G small cell networks.

This has been widely investigated in [3–6]. Compared to transmission power, computation power was obviously smaller and usually fixed as a constant in a traditional energy efficiency evaluation of BSs [4]. As a consequence, the energy efficiency investigation of small cell networks has focused on the optimization of transmission power at BSs [5]. Furthermore, the BS sleeping scheme has been considered to improve energy efficiency where the RF chains and transmitters of BSs are closed to save transmission power [6]. In addition, the computation power of small cell BSs has been improved by the volume and complexity of signal processing, which is weighted by massive MIMO and millimeter-wave technologies [7].

When small cell BSs are ultra-densely deployed in 5G cellular networks [8], there are scenarios in which the computation power of BSs will become larger than the transmission power of BSs despite lower power transmission requirements for small cell BSs.

The transmission rate of 5G mobile communication systems is expected to reach an average of 1 Gb/s (10 Gb/s at peak rate) [2]. Hence, the huge traffic has to be handled at the baseband units (BBUs) of small cell BSs, and then the computation power of signal processing has to be accordingly improved at the BBUs. Moreover, the cache communications and cloud computing network architecture will strengthen functions of signal processing and computing at small cell BSs. Nonetheless, the computation power of 5G small cell networks could be predicted to increase in the near future. All the above factors trigger us to rethink the roles of computation and transmission power in 5G small cell networks.

Based on the Landauer principle, we first propose a computation power model for 5G small cell networks. Considering that the massive MIMO and millimeter-wave technologies are adopted at small cell BSs, the impact of the number of antennas and bandwidths on the computation power of 5G small cell networks is investigated. Simulation results indicate that the computation power will consume more than 50 percent of the energy at 5G small cell BSs. It is a surprising result for the energy efficiency optimization of 5G small cell networks. Finally, future challenges of energy efficiency optimization are discussed for 5G small cell networks, and conclusions are drawn in the last section.

Power Consumption at BSs

To evaluate the roles of computation and transmission power for BSs, the total BS power consumption needs to be analyzed in detail. Therefore, in this section 5G transmission technologies, such as massive MIMO and millimeter-wave, will be incorporated for analyzing the power consumption of small cell BSs.

BS Power Consumption Types

Considering functions and architectures of BSs, the power consumption at BSs is typically classified into three types: transmission power, com-

Xiaohu Ge, Jing Yang, and Yang Sun are with Huazhong University of Science and Technology; Hamid Gharavi is with the National Institute of Standards and Technology.

Digital Object Identifier: 10.1109/MCOM.2017.1600788

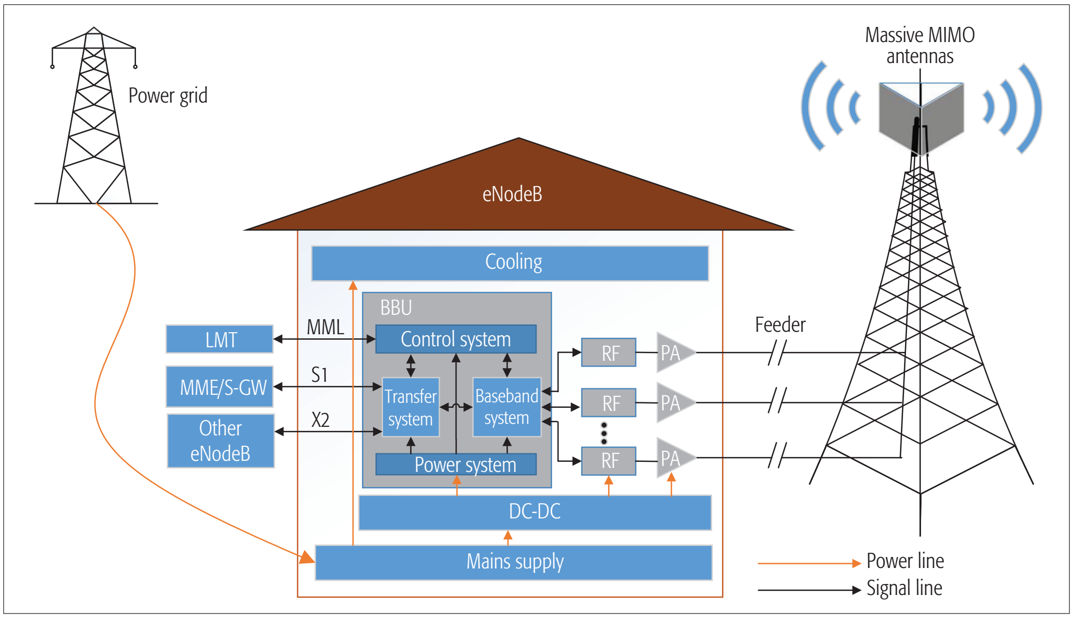


Figure 1. Logistical architecture of eNodeB BS.

Because of the extensive traffic processing at 5G small cell BSs, the volume of data processing at 5G small cell BSs is evaluated by the operation per second at BBUs. Furthermore, Landauer's principle is used to estimate the computation power consumed for data processing in this section.

putation power, and additional power, which are described as follows.

Transmission Power: corresponds to the energy used by power amplifiers (PAs) and RF chains, which perform the wireless signals change, that is, signal transforming between the baseband signals and the wireless radio signals. Also, the power consuming at feeders is included as a part of the transmission power.

Computation Power: represents the energy consumed at BBUs, which includes digital signal processing functions, management and control functions for BSs and the communication functions among the core network and BSs. All these operations are executed by software and realized at semiconductor chips.

Additional Power: represents the BS power, except for the transmission and computation power, for example, the power consumed for maintaining the operation of BSs. More specifically, additional power includes the power lost at the exchange from the power grid to the main supply, at the exchange between different DC-DC power supply, and the power consumed for active cooling at BSs.

The values of the three types of consumed power are different depending on the types of BS. For example, unlike the macrocell BS, the small cell BS normally does not have an active cooling system.

The Total BS Power Consumption Model

The EARTH project has promoted energy efficiency optimization for wireless access networks and proposes a framework for the power consumption at BSs [9]. Based on this energy efficiency framework, the BS is divided into five parts (Fig. 1): the antenna interface, the power amplifier, the RF chains, the BBU, the mains supply, and cooling and DC-DC. The power consumed at the power amplifier and the antenna interface occupy the largest portion of the total power consumption in macrocell BSs (i.e., 57 percent). The portion of RF chains and BBUs is about 10 and 13 percent, respectively, and the portion of remaining parts is about 20 percent. To analyze this in greater detail, the total BS power consumption model is presented as follows: When the BS is equipped with N_{TRX} antennas, the total BS power consumption P_{in} is calculated by

$$P_{in} = \frac{P_{PA} \cdot N_{TRX} + P_{RF} \cdot N_{TRX} + P_{BB}}{(1-\sigma_{DC})(1-\sigma_{MS})(1-\sigma_{cool})},$$

where P_{PA} is the power of PA per antenna, P_{RF} is the RF chain power per antenna, P_{BB} is the power consumed at the BBU, σ_{DC} is the power loss rate of the DC-DC converter, σ_{MS} is the power loss rate of the alternating current supply, and σ_{cool} is the power loss rate of cooling. Based on the expression of the total BS power consumption, $P_{PA} \cdot N_{TRX} + P_{RF} \cdot N_{TRX}$ is the transmission power, P_{BB} is the computation power, and $(1 - \sigma_{DC})(1 - \sigma_{MS})(1 - \sigma_{cool})$ is the relationship between the power loss rate and the total power consumption. The power of PA per antenna is calculated by

$$P_{PA} = \frac{P_{out}}{\eta_{PA} \cdot (1-\sigma_{feed})},$$

where P_{out} is the transmission power at every antenna, η_{PA} is the exchange efficiency of PA, and feeder loss is configured as $\sigma_{feed} = -3$ dB. For macrocell BSs and small cell BSs, the values of σ_{DC}, σ_{MS}, and σ_{cool} are configured as as 6, 7, 9 and 8, 10, and 0 percent, respectively [9]. To simplify calculation, in this article the RF chain power per antenna is usually fixed as different constants corresponding to different types of BSs. Since P_{BB} is obviously less than the power consumed at other parts of BSs, the power consumed at the BBU is fixed as constant in a traditional BS power model [10]. Bear in mind that, as small cells are expected to be widely deployed in 5G cellular networks, the distance between BSs and users will be much shorter, resulting in a considerable reduction of transmission power. Under these conditions, the BBU becomes the dominant source of power consumption.

With the recent advances of 5G of the massive MIMO and millimeter wave technologies, small cell BSs are replacing macro cell BSs to perform the function of wireless data transmission in 5G cellular networks. Moreover, the power consumed at BBUs is expected to gradually increase because of the massive traffic in 5G small cell BSs.

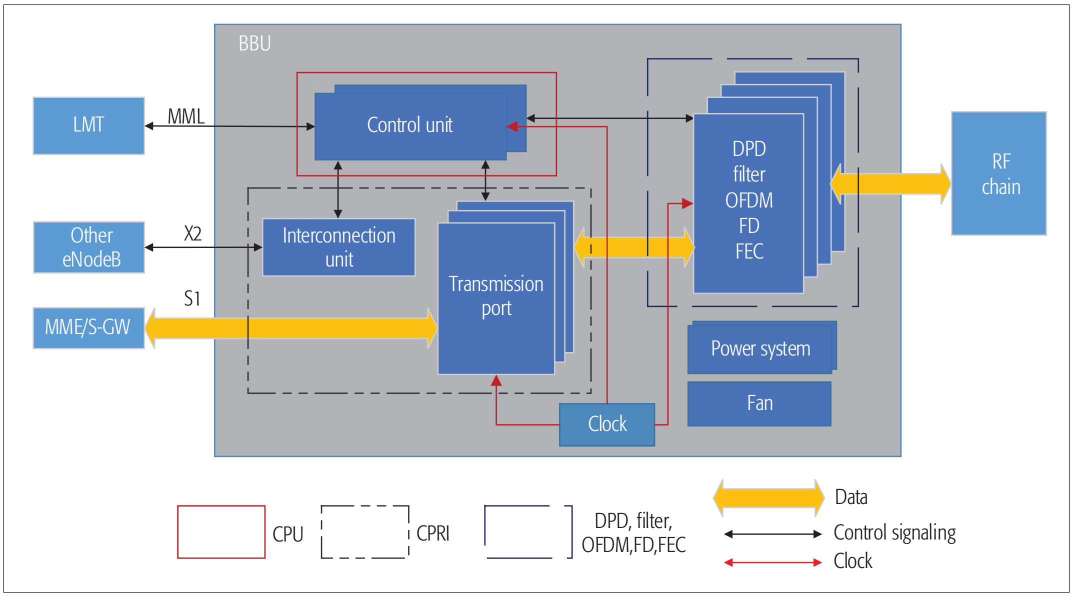


Figure 2. Hardware architecture of BBU.

The Computation Power Model

Because of the extensive traffic processing at 5G small cell BSs, the volume of data processing at 5G small cell BSs is evaluated by the operation per second at BBUs. Furthermore, Landauer's principle is used to estimate the computation power consumed for data processing in this section. In this section we also study the impact of massive MIMO and millimeter-wave technologies on the computation power of 5G small cell BSs.

Computation Power Types

In traditional macrocell BSs the power used at BBUs (the BBU is the core unit of a BS) is small compared to the power consumed by PAs. With the recent advances for 5G of the massive MIMO and millimeter-wave technologies, small cell BSs are replacing macrocell BSs to perform the function of wireless data transmission in 5G cellular networks. Moreover, the power consumed at BBUs is expected to gradually increase because of the massive traffic in 5G small cell BSs.

Figure 1 is a typical logistical architecture of an eNodeB BS, that is, a macrocell BS in a cellular network. Without loss of generality, the BBU of a macrocell BS includes four systems: the baseband system, the control system, the transfer system, and the power system. The detailed functions of these systems in a BBU are described as follows.

Baseband System: Functions include signal filtering, fast Fourier transform/inverse fast Fourier transform (FFT/IFFT), modulation and demodulation, digital pre-distortion (DPD) processing, signal detection, and wireless channel coding/decoding. Note that the function of signal processing used for transmitters and receivers is performed by the BBU.

The Control System: takes charge of controlling and managing resource allocation at BSs in order to provide control interface between the BS and other network units. Moreover, communication control protocols are run at the control system. The control system also provides an interface of man-machine language (MML) for the local maintain terminal (LMT) to configure the resource allocation of BBUs.

The Transfer System: connects with the mobility management entity/serving gateway (MME/S-GW) of the core network by the S1 interface (Fig. 1). Moreover, the control and management information among BSs are forwarded by the X2 interface of the transfer system in BBUs.

The Power System: is responsible for power supply, cooling, and monitoring at BBUs.

For small cell BSs, most functions are integrated into a few semiconductor chips, and there is not a single power system. Therefore, the systems of BBUs at small cell BSs is simpler than the systems of BBUs at macrocell BSs.

The Computation Power Model

Based on the four systems in the logistical architecture shown in Fig. 1, the main difficulty is how to calculate the computation power for every logistical system in BBUs. To achieve this, we partition a BBU into different parts based on the hardware architecture as shown in Fig. 2. These consist of DPD, filter, CPRI, OFDM, FD, FEC, and CPU where DPD is the digital pre-distortion processing part, the filter is the hardware used for up/down signal sampling and filtering, CPRI is the common public radio interface for connecting to the core network and RF chains by serial links, orthogonal frequency-division multiplexing (OFDM) is the hardware used for FFT and OFDM-specific signal processing, FD is the frequency-domain processing part, which includes symbol mapping/demapping and MIMO equalization, FEC is the forward error correction, which includes the channel coding and decoding, and CPU is the BBU platform control processor. Based on Landauer's principle, we estimate the computation power of semiconductor chips using giga operations per second (GOPS) and considering different semiconductor chip techniques. The computation power of a BBU is summed up by the computation power of every hardware part (i.e., every semiconductor chip at the BBU).

Landauer's principle was proposed in 1961 by

Rolf Landauer, who attempted to apply the thermodynamic theory to digital computers. Landauer's principle elaborates the relationship between the information process and energy consumption from the viewpoint of a microscopic degree of freedom in statistical physics. This is based on a physical principle pertaining to the lower theoretical limit of energy consumption that corresponds to the computation. Bear in mind that the concept of entropy in information theory introduced by Claude Shannon is borrowed from thermodynamic theory. Similarly, Landauer's principle connects these two concepts of information and energy by using thermodynamic theory and statistical physics. Therefore, in this article Landauer's principle is first used to analyze the computation power consumption in 5G small cell networks. More specifically, Landauer's principle points out that any logically irreversible manipulation of information, such as the erasure of a bit or the merging of two computation paths, must be accompanied by a corresponding entropy increase in non-information-bearing degrees of freedom of the information processing apparatus or its environment [11]. In other words, erasing a bit of information will consume more than $kT\ln(2)$ energy in a computing system, where k is the Boltzmann constant, that is, 1.38×10^{-23} J/K, and T is the Kelvin temperature [12]. According to Landauer's principle, the lower bound of computation power for a computing system can be obtained. Compared to the value of computation power at real semiconductor chips, there is a difference of three orders of magnitude for the values of computation power derived by Landauer's principle [13]. Moreover, the values of computation power are different when different semiconductor chip techniques are adopted at BBUs. Under these conditions, the main difficulty is how to accurately calculate the computation power of small cell BSs using Landauer's principle.

To overcome the gap of computation power estimated by Landauer's principle and real semiconductor chips, we propose a power coefficient ε is that can represent the level of the semiconductor chip technique in BBUs. Moreover, the power coefficient ε is defined as the ratio of the active switching power of a transistor and the limit of Landauer's principle. From Fig. 3, the power coefficient ε reflects the distance between semiconductor chip techniques and the limit of Landauer's principle. Bear in mind that up until now, the development of semiconductor chip techniques still follows Moore's law. However, the international technology roadmap for semiconductors (ITRS) predicts that the development of semiconductor chip techniques will deviate from Moore's law when the power coefficient approaches the limit of Landauer's principle. For example, when nanomagnetic logic is used for chips, the computation power is expected to approach the limit of Landauer's principle [14]. Considering the development of current chip techniques, we focus our attention on the computation power of semiconductor chips.

Without loss of generality, in this article the power coefficient is configured as $\varepsilon = 10^3$ when the 22 nm semiconductor technique is assumed to be adopted for chip manufacture in BBUs. Moreover, the active switching power of a transistor is approximated by $E_{FET} \approx \varepsilon kT\ln(2)$, which is used to calculate the power for operating 1 bit of information at the semiconductor chip of BBUs.

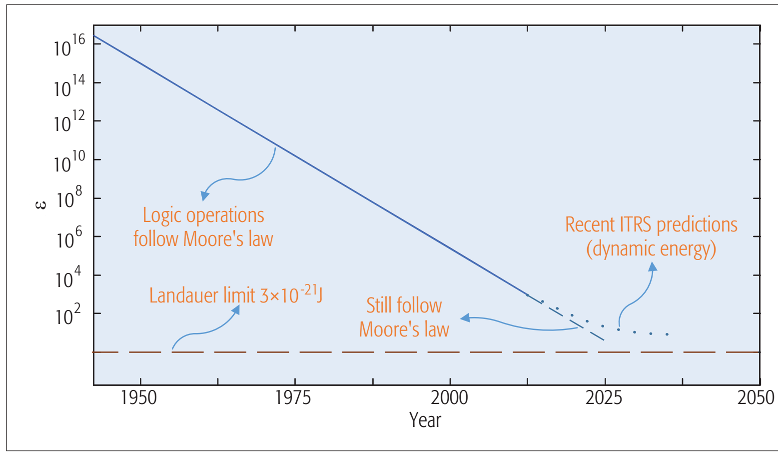


Figure 3. The power coefficient with respect to the development of chip techniques.

In general, the data processing rate of semiconductor chips is represented by the instructions per second (IPS). Based on the definition of GOPS, in this article the relationship between the IPS and the GOPS is expressed by

$$IPS = \frac{GOPS \times 10^9}{64}$$

when the logistical architecture of semiconductor chips is assumed to be 64 bits. According to the experimental results in [15], the information throughput of semiconductor chips is denoted by

$$\rho = \left(\frac{IPS}{\omega}\right)^{\frac{1}{\gamma}}$$

where ω and γ are configured as 0.1 and 0.64, respectively. As a consequence, the computation power of different parts of a BBU is calculated by the product of the information throughput of semiconductor chips and the active switching power of transistors considering different values of GOPS at different parts of the BBU.

Since different types of BSs have different hardware components at BBUs, it is difficult to directly build a uniform model to evaluate the computation power of BBUs in different types of BSs. Therefore, we first build a reference BS with typical parameters. By comparing different types of BSs with the reference BS, we can derive the computation power of different BBUs for different types of BSs. Without loss of generality, the system parameters are represented by $i \in \{BW, Ant, M, R, dt, df\}$, where BW is the bandwidth parameter, Ant is the number of antennas parameter, M is the modulation coefficient parameter, R is the parameter of coding rate, dt is the parameter of time-domain duty cycling, and df is the parameter of frequency-domain duty cycling. To simplify symbols in this article, X_i^{ref} is denoted as the reference BS. When the subscript i of X_i^{ref} is replaced by different symbols, the new variable represents the corresponding system parameter in the reference BS, for example, X_{BW}^{ref} is the bandwidth of the reference BS. Similarly, X_i^{real} is denoted for a real BS. When the subscript i of X_i^{real} is replaced by

BBU parameters	GOPS of macrocell	GOPS of small cell	S_{BW}	S_M	S_R	S_{Ant}	S_{dt}	S_{df}
DPD	160	0	1	0	0	1	1	0
Filter	400	250	1	0	0	1	1	0
CPRI/SERDES	720	0	1	1	1	1	1	1
OFDM	160	120	1	0	0	1	1	0
FD (linear)	90	50	1	0	0	1	1	1
FD (nonlinear)	30	15	1	0	0	2	1	1
FEC	140	130	1	1	1	1	1	1
CPU	400	40	0	0	0	1	0	0

Table 1. Configuration parameters of the BBU.

different symbols, the new variable represents the corresponding system parameter in the real BS, for example, X_{BW}^{real} is the bandwidth of the real BS.

Considering different computation powers at different hardware parts of a BBU, the computation power of the DPD filter, CPRI, OFDM, frequency domain (FD), FEC, and CPU at reference BS are denoted by P_{DPD}^{ref}, P_{Filter}^{ref}, P_{CPRI}^{ref}, P_{OFDM}^{ref}, P_{FD}^{ref}, P_{FEC}^{ref}, and P_{CPU}^{ref}, respectively. The different hardware parts of a BBU depend on the different system parameters of BS. S_i, $i \in \{BW, Ant, M, R, dt, df\}$ signifies the ratio of the different hardware parts of the BBU and the system parameters of the BS. When the relationship between the hardware part of the BBU and the system parameter of the BS is linear, the corresponding S_i is configured as 1. If such a relationship is nonlinear, the corresponding S_i is set to 2. When the relationship between the hardware part of the BBU and the system parameter of the BS is independent, the corresponding S_i is configured as 0. The detailed configuration parameters of S_i are illustrated in Table 1.

Based on measurement results from the reference BS, the computation power of the reference BBU can be obtained by $P_{BB}^{ref} = P_{DPD}^{ref} + P_{Filter}^{ref} + P_{CPRI}^{ref} + P_{OFDM}^{ref} + P_{FD}^{ref} + P_{FEC}^{ref} + P_{CPU}^{ref}$. To calculate the computation power of real BBUs, the reference coefficient α is defined by

$$\alpha = \prod_i \left(\frac{X_i^{real}}{X_i^{ref}}\right)^{S_i} = \left(\frac{X_{BW}^{real}}{X_{BW}^{ref}}\right)^{S_{BW}} \cdot \left(\frac{X_{Ant}^{real}}{X_{Ant}^{ref}}\right)^{S_{Ant}} \cdot \left(\frac{X_M^{real}}{X_M^{ref}}\right)^{S_M} \cdot \left(\frac{X_R^{real}}{X_R^{ref}}\right)^{S_R} \cdot \left(\frac{X_{dt}^{real}}{X_{dt}^{ref}}\right)^{S_{dt}} \cdot \left(\frac{X_{df}^{real}}{Xdf_i^{ref}}\right)^{S_{df}}.$$

Finally, the computation power of real BBUs is calculated by $P_{BB}^{realL} = \alpha \cdot P_{BB}^{ref}$.

Evaluations of Computation Power

Considering that 5G small cell networks with massive MIMO and millimeter-wave techniques have not yet been commercially deployed, it is difficult to compare our simulation results with real 5G small cell networks. To validate the performance of the proposed power consumption model, we first compare the results of the proposed model with those of the EARTH project [9], which measures the power consumption of macrocell and small cell BSs from real wireless networks. Without loss of generality, the two wireless communication systems are configured with 10 MHz and 2 × 2 antennas at BSs and terminals. Based on the results from the EARTH project, the total power consumption and computation power of a macrocell BS are 321.6 W and 29.68 W, respectively. Similarly, for our proposed power consumption model the total power consumption and the computation power of a macrocell BS are 317.84 W and 24.78 W, respectively. In the case of a small cell BS for the EARTH project, the total power consumption and computation power of a small cell BS are 6.2 W and 2.4 W, respectively. For the proposed power consumption model, the total power consumption and computation power are 7.22 W and 3.6 W, respectively. Compared to the above, the results of the proposed power consumption model are in agreement with the results of real wireless networks. Therefore, our proposed power consumption model is shown to be capable of estimating the power consumption of 5G small cell networks.

Without loss of generality, the system parameters of the reference BS are configured as X_{BW}^{ref} = 20 MHz, X_{Ant}^{ref} = 1, X_M^{ref} = 6, X_R^{ref} = 1, X_{dt}^{ref} = 100 percent, X_{df}^{ref} = 100 percent. Based on the configuration parameters of the BSs in Table I, the computation power of BSs is simulated for 5G small cell networks. Since the massive MIMO and millimeter-wave technologies are the core technologies for 5G mobile communication systems, in this section the impact of the number of antennas and bandwidths on macrocell BSs and small cell BSs are simulated in detail.

Generally speaking, the PAs of a macrocell BS and a small cell BS are configured as 102.6 W and 1.0 W. Figure 4 illustrates the computation power of a BS with respect to the number of antennas and bandwidths. The default system parameters of real BSs are configured as follows: the bandwidth is 20 MHz, the modulation is 64-quadrature amplitude modulation (QAM), the coding rate is (5/6), the time-domain duty cycling is 100 percent, and the FD duty cycling is 100 percent. Figure 4a shows the computation power of BSs with respect to the number of antennas. Based on the results in Fig. 4a, the computation power of BSs quickly increases with the increase in the number of antennas. The reason is that the computation power consumed for FD processing is in proportion to the square of the number of antennas. Moreover, the computation power of macrocell BSs is always larger than the computation power of small cell BSs when the number of antennas is increased. When the number of antennas is equal to 128, that is, adopting massive MIMO technology, the computation power of a macrocell BS is larger than 3000 W, and the computation power of a small cell BS is larger than 800 W.

In general, with the adaptation of millimeter-wave techniques, 5G communication systems will be able to support large bandwidths (e.g., 400 MHz) or, more precisely, high transmission rates. Consequently, this would require more processing at the BBU, hence further increasing the computation power at BSs. Therefore, in this article, the impact of a millimeter-wave technique on the computation power of BSs is based on a wireless communication bandwidth. When the number of antennas is configured as 4, Fig. 4b depicts the

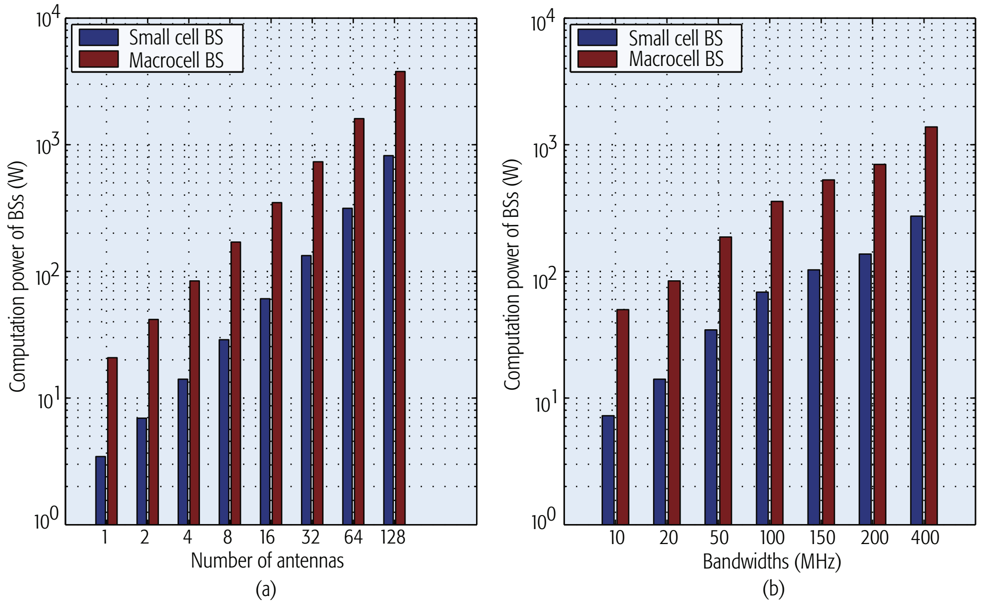


Figure 4. Computation power of BSs with respect to the number of antennas and bandwidths.

It is possible to predict that computation power, no matter what the power consumption level of the absolute volume and the ratio will be shall further improve for 5G cellular networks. In this case, the energy efficiency optimization of 5G cellular networks will not only consider the transmission power, it will also consider the power consumed for data computation and storage at BSs.

computation power of BSs with respect to bandwidths. Based on the results in Fig. 4b, the computation power of a BS increases with the increase of bandwidths. Moreover, the computation power of macrocell BSs is always larger than the computation power of small cell BSs when the bandwidth is increased. When the bandwidth is 400 MHz (i.e., adopting the millimeter-wave technology), the computation power of a macrocell BS is larger than 1000 W, and the computation power of a small cell BS is larger than 200 W. Based on results in Fig. 4, small cell BSs can save more computation power for BBUs than macrocell BSs in 5G mobile communication systems.

To evaluate the role of computation power in the BS, the computation power ratio is defined by the computation power over the total power at a BS. Figure 5 illustrates the computation power ratio with respect to the number of antennas and bandwidths for small cell BSs and macrocell BSs. Figure 5a shows the computation power with respect to the number of antennas. Based on the results in Fig. 5a, the computation power ratio increases with the increased number of antennas. Moreover, the computation power ratio of small cell BSs is always larger than the computation power ratio of macrocell BSs. In addition, the computation power ratio of small cell BSs is obviously larger than 50 percent. Figure 5b depicts the computation power ratio with respect to bandwidths. Based on the results in Fig. 5b, the computation power ratio increases with the increase of bandwidths. Moreover, the computation power of small cell BSs is always larger than the computation power of macrocell BSs. When millimeter-wave technology is adopted, the bandwidth is larger than or equal to 20 MHz, and the computation power ratio of small cell BSs is obviously larger than 50 percent.

Future Challenges

Based on the results in Figs. 4 and 5, the computation power will play a more important role than other power consumptions, including the transmission of power at 5G small cell BSs, no matter what the level of the absolute volume and the ratio for 5G small cell networks is. On the other hand, energy efficiency of 5G mobile communication systems is expected to improve 100 to 1000 times compared to the energy efficiency of 4G mobile communication systems. However, most studies involving the energy efficiency of 5G cellular networks still focus on the transmission power optimization of BSs. To face the role of computation power in 5G small cell networks, some potential challenges are presented here.

The first challenge is the impact of 5G network architectures on the computation power in 5G small cell networks. Based on the results in Fig. 5, the importance of computation power is improved for energy efficiency optimization of 5G small cell networks. One obvious reason is that the transmission power is reduced in 5G small cell networks that adopt the massive MIMO and millimeter-wave technologies. With cloud/fog computing and cache communications emerging for 5G networks, more and more data storage and computation will be performed at 5G small cell BSs. Therefore, it is possible to predict that computation power, no matter what the power consumption level of the absolute volume and the ratio will be, shall further improve for 5G cellular networks. In this case, the energy efficiency optimization of 5G cellular networks will not only consider the transmission power, but will also consider the power consumed for data computation and storage at BSs.

The second challenge is optimization of computation power at BSs with massive MIMO and millimeter-wave transmission technologies. Existing studies usually fix the value of computation power at BSs. Moreover, the impact of 5G wireless transmission technologies such as the massive MIMO and millimeter-wave technologies on the computation power is ignored at BSs. Based on the results in Fig. 4, the massive MIMO and millimeter-wave technologies have a greater impact

The new energy efficiency model of 5G small cell networks considering computation and transmission power needs to be investigated. Moreover, the software-defined networks (SDN) could be used to trade off computation and transmission powers at 5G small cell BSs with cloud/fog computing functions.

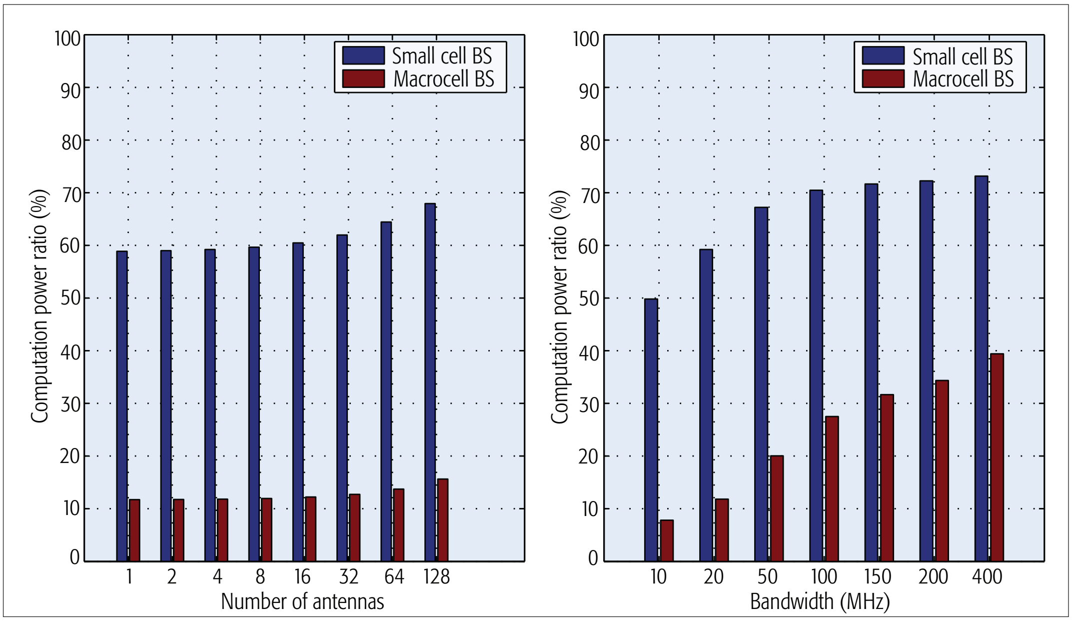


Figure 5. Computation power ratio with respect to the number of antennas and bandwidths.

on the computation power of 5G small cell BSs. Considering the role of computation power at 5G small cell BSs, it is inadvisable to ignore the impact of 5G transmission technologies on the computation power of 5G small cell BSs. When massive MIMO and millimeter-wave technologies are adopted by 5G small cell BSs, a large number of antennas and bandwidths can be scheduled for resource optimization in 5G small cell networks. An interesting question is how to schedule the number of antennas and bandwidths for the optimization of computation power at 5G small cell BSs.

The third challenge is the trade-off between computation power and transmission power in 5G networks. Based on the analysis above, the additional power of BSs depends on the computation and transmission powers of BSs. When the additional power of BSs is combined with the computation and transmission power of BSs, the energy efficiency of 5G networks can be calculated by the energy efficiency of computation and transmission powers at BSs. However, 5G transmission technologies have different effects on the energy efficiency of computation and transmission power of small cell BSs. In some specific scenarios, the effects on energy efficiency of computation and transmission powers are contradictory at 5G small cell BSs. Hence, the relationship between the computation and communication powers needs to be further investigated for 5G networks. Moreover, the trade-off between computation and transmission power needs to be optimized for 5G small cell BSs.

To face the above challenges in the energy efficiency optimization of 5G small cell networks, some potential research directions are summarized to solve these issues:

The new energy efficiency model of 5G small cell networks considering computation and transmission power needs to be investigated. Moreover, software-defined networks (SDNs) could be used to trade off computation and transmission powers at 5G small cell BSs with cloud/fog computing functions.

To improve the energy efficiency of 5G small cell BSs, joint optimization schemes and algorithms should be developed to save computation and transmission power at BBUs and RF chains together.

Based on the simulation results in Fig. 4, a lot of the computation power of BBUs has to be changed into heat, and more cooling systems need to be designed to support computation functions at BBUs. To save energy at BBUs, we should take the energy cycle into account, and some potential technologies are expected to change the heat from BBUs into electrical energy based on the pyroelectric effect.

Conclusions

Until recently, the computation power of BSs was ignored or just fixed as a small constant in the energy efficiency evaluation of cellular networks. In this article, the power consumption of BSs is analyzed for 5G small cell networks adopting massive MIMO and millimeter-wave technologies. Considering the massive traffic in 5G small cell networks, the computation power of 5G small cell BSs is first estimated based on Landauer's principle. Moreover, simulation results show that the computation power of BSs increases as the number of antennas and bandwidths increases. Compared to transmission power, computation power will play a more important role in the energy efficiency optimization of 5G small cell networks. Therefore, we conclude that the energy efficiency optimization of 5G small cell networks should consider computation and transmission power together. How to converge computation and transmission technologies to optimize the energy efficiency of 5G networks is still an open issue. If this is accomplished, a different challenge would indeed emerge in the next round of the transmission and computation revolution.

Acknowledgment

The authors would like to acknowledge the support from the NSFC Major International Joint Research Project (Grant No. 61210002) and the Fundamental Research Funds for the Central Universities under grant 2015XJGH011. This research is partially supported by the EU FP7-PEOPLE-IRSES, project acronym CROWN (grant no. 610524), and the China international Joint Research Center of Green Communications and Networking (No. 2015B01008).

References

[1] X. Ge *et al.*, "5G Ultra-Dense Cellular Networks," *IEEE Wireless Commun.*, vol. 23, no. 1, Feb. 2016, pp. 72–79.

[2] J. G. Andrews *et al.*, "What Will 5G Be?" *IEEE JSAC*, vol. 32, no. 6, June 2014, pp. 1065–82.

[3] X. Ge *et al.*, "Energy Efficiency of Small Cell Backhaul Networks Based on Gauss-Markov Mobile Models," *IET Networks*, vol. 4, Mar. 2015, pp. 158–67.

[4] S. Samarakoon *et al.*, "Ultra Dense Small Cell Networks: Turning Density Into Energy Efficiency," *IEEE JSAC*, vol. 34, May 2016, pp. 1267–80.

[5] X. Ta, G. Mao, and B. D. O. Anderson, "On the Giant Component of Wireless Multihop Networks in the Presence of Shadowing," *IEEE Trans. Vehic. Tech.*, vol. 58, no. 9, Nov. 2009, pp. 5152–63.

[6] C. Liu, B. Natarajan, and H. Xia, "Small Cell Base Station Sleep Strategies for Energy Efficiency," *IEEE Trans. Vehic. Tech.*, vol. 65, Mar. 2016, pp. 1652–61.

[7] J. Choi, "Energy Efficiency of A Heterogeneous Network Using Millimeter-Wave Small-Cell Base Stations," *Proc. IEEE 26th Annual Int'l. Symp. Personal, Indoor, and Mobile Radio Commun.*, 2015, pp. 293–97.

[8] S. Chen *et al.*, "User-Centric Ultra-Dense Networks for 5G: Challenges, Methodologies, and Directions," *IEEE Wireless Commun.*, vol. 23, no. 2, Apr. 2016, pp. 78–85.

[9] M. Imran *et al.*, "Energy Efficiency Analysis of the Reference Systems, Areas of Improvements and Target Breakdown," tech. rep. ICT-EARTH deliv., 2011.

[10] C. Desset *et al.*, "Flexible Power Modeling of LTE Base Stations," *Proc. IEEE WCNC*, 2012, pp. 2858–62.

[11] R. Landauer, "Irreversibility and Heat Generation in The Computing Process," *IBM J. Research and Development*, vol. 5, 1961, pp. 183–91.

[12] A. Berut *et al.*, "Experimental Verification of Landauer's Principle Linking Information and Thermodynamics," *Nature*, vol. 483, Mar. 2012, pp. 187–89.

[13] W. P. Cockshott *et al.*, *Computation and Its Limits*, Oxford Univ. Press, 2012.

[14] B. Lambson, D. Carlton, and J. Bokor, "Exploring the Thermodynamic Limits of Computation in Integrated Systems: Magnetic Memory, Nanomagnetic Logic, and the Landauer Limit," *Physical Review Letters*, vol. 107, no. 1, July 2011, p. 010604.

[15] L. Zhirnov, R. Cavin, and L. Gammaitoni, "Minimum Energy of Computing, Fundamental Considerations," *ICT-Energy-Concepts Towards Zero-Power Info. and Commun. Technology*, vol. 7, 2014.

How to converge computation and transmission technologies to optimize the energy efficiency of 5G networks is still an open issue. If this is accomplished, a different challenge would indeed emerge in the next round of the transmission and computation revolution.

Biographies

XIAOHU GE [M'09, SM'11] (xhge@mail.hust.edu.cn) is currently a full professor with the School of Electronic Information and Communications at Huazhong University of Science and Technology (HUST), China, and an adjunct professor with the Faculty of Engineering and Information Technology at the University of Technology Sydney (UTS), Australia. He received his Ph.D. degree in communication and information engineering from HUST in 2003. His research interests include green communications, vehicular communications, and wireless networks. He is the director of the China International Joint Research Center of Green Communications and Networking. He serves as an Associate Editor for *IEEE Transaction on Greening Communications and Networking*, *IEEE Access*, and so on.

JING YANG (yang_jing@mail.hust.edu.cn) received his B.E. degrees in communication engineering from HUST in 2014. He is currently pursuing a Ph.D. degree in the School of Electronic Information and Communications at HUST. His research interests mainly include computation power of wireless communication systems, green communication, and energy efficiency of wireless cellular networks.

HAMID GHARAVI [LF] (hamid.gharavi@nist.gov) received his Ph.D. degree from Loughborough University, United Kingdom, in 1980. He joined the Visual Communication Research Department at AT&T Bell Laboratories, Holmdel, New Jersey, in 1982. He was then transferred to Bell Communications Research (Bellcore), where he became a Distinguished Member of Research Staff. In 1993, he joined Loughborough University as a professor and chair of communication engineering. Since September 1998, he has been with the National Institute of Standards and Technology (NIST). His research interests include smart grid, wireless multimedia, mobile communications and wireless systems, mobile ad hoc networks, and visual communications. He served as a member of the Editorial Board of *Proceedings of the IEEE* from January 2003 to December 2008. From January 2010 to December 2013 he served as Editor-in-Chief of *IEEE Transactions on CAS for Video Technology*. He is currently serving as Editor-in-Chief of *IEEE Wireless Communications*.

YANG SUN (tylzzq@mail.hust.edu.cn) received his B.S. degrees with honors in electronic and information engineering from HUST and now is pursuing his Master's degree in electronic information and communications, HUST. His research interests are focused on green communications.

On Measuring the Geographic Diversity of Internet Routes

Attila Csoma, András Gulyás, and László Toka

Route diversity in networks is elemental for establishing reliable, high-capacity connections with appropriate security between endpoints. As for the Internet, route diversity has already been studied at both the autonomous system and router level topologies by means of graph theoretical disjoint paths. In this article the authors complement these approaches by proposing a method for measuring the diversity of Internet paths in a geographical sense.

The work of Laszló Toka and András Gulyás was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (MTA).

Digital Object Identifier: 10.1109/MCOM.2017.1100111

Editor's Note: The IEEE Communications Society has a Sister Society agreement with the Hungarian Association for Infocommunications (HTE). The terms of the agreement include republication of articles of HTE's *Infocommunications Journal* in IEEE Communications Society publications. The article below has already appeared in *Infocommunications Journal*. The citation is: *Infocommunications Journal*, no 4, pp. 34–39, Dec. 2015.

Osman Gebizlioglu, Editor-in-Chief
IEEE Communications Magazine

Abstract

Route diversity in networks is elemental for establishing reliable, high-capacity connections with appropriate security between endpoints. As for the Internet, route diversity has already been studied at both the autonomous system and router level topologies by means of graph theoretical disjoint paths. In this article we complement these approaches by proposing a method for measuring the diversity of Internet paths in a geographical sense. By leveraging the recent developments in IP geolocation we show how to map the paths discovered by `traceroute` into geographically equivalent classes. This allows us to identify the geographical footprints of the major transmission paths between end-hosts, and building on our observations, we propose a quantitative measure for geographical diversity of Internet routes between any two hosts.

Introduction

The value of knowledge of the Internet topology is arguably immense. In the last decades we have witnessed a myriad of stories in which topology-related information about the Internet was directly compiled into more efficient architectures, services and more appropriate business decisions. content delivery networks (CDNs) [1], in which global topological peculiarities are highly exploited for, say, surrogate server and cache placement strategies or request routing mechanisms, are just a narrow segment of the whole spectrum. Peer-to-peer networks [2, 3], data center placement [4], traffic engineering [5], business-based autonomous system (AS) peering strategies [6], just to mention a few, are all receivers of Internet topology related knowledge. With this non-comprehensive list of receivers in mind, it should come as no surprise that many researchers have contributed to our current understanding of the topology of the Internet.

An elemental metric of Internet topology is the diversity of routes between end hosts, as multiple uncorrelated routes can provide better throughput, resiliency and security. In [7–9] authors describe how to increase the resilience of future networks and the role of multipath communication in it. A detailed description about network security can be found in [10], and authors in [11] propose a method which improves network security; however, it requires multiple path between end-hosts.

Existing studies of IP-level route diversity usually focus on extracting routes between end hosts, for example, by using `traceroute`, and on computing their diversity by means of edge or node disjointedness in a graph theoretical sense [12–14]. Such analysis provides an interpretation of route diversity in a microscopic level where each node in the route is a router interface having a particular IP address. In [15] the authors propose to interpret route diversity at a higher level, namely at the level of points of presence (PoPs): the interfaces residing in the same building or campus are grouped together, forming a PoP, and finally, route diversity is computed at the level of these PoPs. In this article we interpret route diversity at an even higher, geographical level. We propose grouping routers in a given geographic vicinity and compute route diversity at the level of geographical regions (e.g., the level of cities) independently from ASs. Our contribution is threefold: *first, we describe a method for identifying the geographically equivalent routes in* `traceroute` *outputs; second, we show the efficiency of the method in terms of successfully merged* `traceroute` *routes and present their possible applications; finally, we define a metric which can capture the geo-diversity of Internet routes between endpoints and compute this metric for our measurement dataset*. Such characterization of routes' "geo-diversity" is clearly beneficial if one is curious about connectivity between end hosts in case of, say, power outages affecting larger geographical areas.

The rest of the article is organized as follows. We overview the corresponding related work. We describe our `traceroute` measurements and our algorithms for extracting geographically equivalent routes from those. We define and evaluate a metric called Geographic Diversity Index (GDI) that captures how Internet routes differ from each

Attila Csoma is with Budapest University of Technology and Economics and with HSNLab; András Gulyas and László Toka are with Budapest University of Technology and Economics, with HSNLab, and with MTA-BME Information Systems Research Group.

other. We validate our framework and present its performance. Finally, we draw the conclusions and list the possible applications and future work.

Related Work

Numerous existing studies apply geographical information to uncover non-trivial aspects of the Internet topology. In [16] the authors use the geographical positions of routers to estimate route circuitousness, route length distribution and geographic fault tolerance from an end-to-end and from an ISP perspective. However, the DNS name-based geolocation method used in their work has its own limitations and may create a false spatial distribution as described in [17]. The distance and angle between consecutive IP hops are investigated in [18]. In [19] the authors use geographic information to construct the hyperbolic map of the Internet and prove the navigability of the AS level topology using greedy algorithms. Points of presence are detected using delay constraints on an IP interface graph, and the distribution of PoPs around the globe is visualized in [15]. The authors of [20] used PoP detection to evaluate the accuracy of an IP geolocation database. Inter-AS route diversity is examined through the network of Sprint in [14]. In [21] the authors study the route diversity of multi-homed and overlay networks as seen from multiple vantage points using graph theoretical methods exclusively. However, none of these works capture route diversity in the pure geographical sense on the router-level.

Measurement Framework

We built a framework that determines the extent of geographical heterogeneity of end-to-end routes that are being used to carry traffic between any two points in the Internet. The method that we implemented is the following: first, we run `traceroute` measurements to collect the IP-level routes between the selected endpoints; second, we use *MaxMind* [22], a localization tool that determines the geographical position of the recorded IP addresses; third, we group those IP-level routes that we consider equivalent from a geographical perspective; finally, we calculate a route diversity index for the selected endpoints. In this section we present these steps in details.

Route Measurements With Traceroute

Usually `traceroute` is used to discover end-to-end routes between two endpoints in the Internet. Network operators use it for detecting network errors; researchers use it to build Internet topology models. Although most in-network routers and endpoints support its operation, `traceroute` has a number of well-known shortcomings. On one hand, it can be easily deceived by load balancers; on the other hand, it is an active measurement tool and due to the extra data traffic it generates, certain network equipment are configured to disable reactions to `traceroute` (and `ping`).

Several projects exist that collect `traceroute` measurements and make them publicly available. These data sources differ significantly based on their vantage point types, their vantage point location, the `traceroute` implementation they run and their endpoint selection methods. Two such projects are IPlane [23] and CAIDA's ITDK (Internet Topology Data Kit) [24]. IPlane offers a route performance prediction service and periodically runs `traceroute` measurements from PlanetLab [25] nodes to a set of endpoints changed biweekly. ITDK datasets are produced from measurements collected by CAIDA's Archipelago project: `paris-traceroute` [26] is run from 89 vantage points spread over 37 countries to randomly selected endpoints. In order to measure route diversity between two endpoints, we need to detect as many routes between those two endpoints as possible. Although the `paris-traceroute` output of ITDK is more reliable than that of IPlane's `traceroute`, the random selection of endpoints implemented by CAIDA hinders the collection of routes between the same vantageand endpoints. Therefore we used the data of IPlane's `traceroute` measurements.

IP Localization and Filtering of Routes

Once we have the IP-level routes, we determine the geographic position of each node appearing in them. Naturally the accuracy of the positioning is of paramount importance. As also noted in [27, 17], the use of DNS names and contents of various registries leads to unacceptable inaccuracy. Instead, we use the freely available geolocation tool MaxMind GeoLite [22] in order to establish the position for the IP addresses recorded in the measurements. As pointed out in [28], it is one of the most reliable, freely available geolocation database.

We filter the `traceroute` dataset from IPlane in order to remove measurements of vantage and endpoint pairs between which only 1 IP-level route was observed. After localizing the IP hops, we further removed those vantage and endpoint pairs between which only 1 geographic path was observed. With this second filter we eliminated traces differing only due to IP-aliasing and load balancing. The remaining set of traces contained ~0.5 million discovered routes; interestingly, ~80 percent of vantage and endpoint pairs had only 1 geographic path in the measurements.

Clustering of Routes

After establishing *geo-paths* for the remaining routes by localizing their IP hops, we set out to decide which routes can be considered to be the same and which ones are different in a geographical sense. We make this decision by clustering the geo-paths on a hop-by-hop basis and by defining geodiversity, a mutual metric, between geo-paths. Iterating through all the geo-paths from a given vantage point to a given endpoint, we choose an appropriate cluster for each geo-path: if the geo-path satisfies the geographical equality with all cluster members, it is assigned to that cluster; if not, a new cluster is created for it. At the end, all geo-paths in each cluster are considered to be geographically the same.

Two geo-paths are geographically equal if they are not farther from each other than 50 km: the distance between any of their nodes and the closest one of the straight lines determined by consecutive nodes of the other geo-path is not larger than 50 km. We used the arbitrary threshold of 50 km to reflect a large city's diameter [29]. Note that this threshold also allows for fiber duct curves in the physical net-

Two geo-paths are geographically equal if they are not farther from each other than 50 km: the distance between any of their nodes and the closest one of the straight lines determined by consecutive nodes of the other geo-path is not larger than 50 km.

Our final step to capture the geographic route diversity is to define and evaluate a Geographic Diversity Index (GDI). We require GDI to produce values to a given route set between a source and destination pair such that multiple geo-paths spanning over large geographical areas get a higher GDI.

Figure 1. Route merging example.

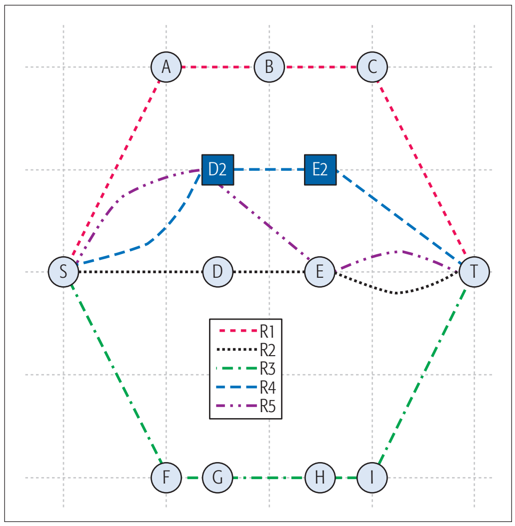


Figure 2. Geo-diversity example.

work, alleviating the mismatch between the location of IP-level nodes of `traceroute` measurements and the actual trail of the underlying physical links.

As a demonstrative example, we drew two geo-paths that are grouped in the same cluster in Fig. 1. Those routes differ at the IP level and their geo-paths are also different, but since the distance (marked with dashed blue line) of the intermediate node of one route from the geo-path of the other route is smaller than 50 km, our clustering algorithm rules the geographic difference between these two routes negligible.

Calculation of Geographic Route Diversity

Our final step to capture the geographic route diversity is to define and evaluate a Geographic Diversity Index (GDI). We require a GDI to produce values to a given route set between a source and destination pair such that multiple geo-paths spanning over large geographical areas get a higher GDI.

Requirements

Before describing the computation of GDI in detail, we highlight the key attributes that make a route set "more" diverse in a geographical sense against another. Let us assume a source (S) and a target (T) node as endpoints and two "cover" routes S-A-B-C-T (R1) and S-F-G-H-I-T (R3) as shown in Fig. 2.[1] Let us assume that nodes are positioned according to their geographic positions. Let us also assume that the GDI for this setting is $\hat{r}$. Our first goal is to reward higher route count. Therefore, a route set with two routes must

[1] Lines are curved to distinguish different routes using the same link.

r	2182
$\hat{r}_4$	3299
$\hat{r}_5$	3467
$\hat{r}_2$	3729

Table 1. GDI values for the geo-diversity example.

achieve lower GDI than the same route set extended with another arbitrary route. That is, if we add R2 to our theoretical example and obtain $\hat{r}_2$ as the GDI for this amended route set, we expect $\hat{r} < \hat{r}_2$. Our second goal is to reward higher geographic distance between routes. Therefore, if we modify the route set by reducing the distance between the routes, the GDI of the new route set must be lower: by replacing R2 with R4 (new GDI is $\hat{r}_4$), we require the GDI to fall. It follows that geo-paths with varying distances (from other routes) increase the GDI more (e.g., R5 and $\hat{r}_5$) than parallel geo-paths (e.g., R4 and $\hat{r}_4$) closer than R2. Therefore, we require the order between GDI values to be

$$\hat{r} < \hat{r}_4 < \hat{r}_5 < \hat{r}_2 \quad (1)$$

Geographic Diversity Index

We model geo-paths of Internet routes as collections of sections between their consecutive nodes. Let us assume two routes defined by their nodes: $P = \{a, b, c, d\}$ and $L = \{e, f, g\}$. Naturally, the distance $\delta(a, L)$ between node a and route L translates to the distance between a and the closest point (not necessarily a node) of route L to a. Let $\Delta(P, L) = \{\forall u \in P \,|\, (\delta(u, L), \forall u \in L \,|\, (\delta(u, P)\}$ be a vector containing all possible node distances between P and L. In the toy example of Fig. 3 $\Delta(P, L) = \{\delta(a, L), \delta(b, L), \delta(c, L), \delta(d, L), \delta(e, P), \delta(f, P), \delta(g, P)\}$.

In order to satisfy the requirements listed above, we define GDI for a given set of routes as follows. First, we define a diversity index between two routes as

$$d(P, L) = \left(1 - \mathrm{Var}'\left(\Delta(P, L)\right)\right) \mathrm{Mean}(\Delta(P, L)), \quad (2)$$

where Var′ denotes the variance of its arguments normalized to the interval [0..1]. Second, we define the diversity between a single route P and a set of routes $\mathbb{V}$ as

$$\mathcal{D}(P, \mathbb{V}) = \min_{L \in V} d(P, L) \quad (3)$$

Finally, we quantify the overall GDI for a given set of routes. In order to calculate this, we use a step-by-step method. Let us assume a set $\mathbb{V}$ containing the routes. The process starts from an empty set $\mathbb{U} = \{\}$. In the 0. step we search in set $\mathbb{V}$ for the two routes having the highest diversity score $d_0 = \max_{P,L \in \mathbb{V}}(d(P, L))$ and move paths $\mathrm{argmax}_{P,L \in \mathbb{V}}(d(P, L))$ to $\mathbb{U}$. In the ith step we compute $\mathcal{D}_i = \max_{P \in \mathbb{V}}(\mathcal{D}(P, \mathbb{U}))$ and move the path $\mathrm{argmax}_{P \in \mathbb{V}}(\mathcal{D}(P, \mathbb{U}))$ to $\mathbb{U}$. The process terminates when $\mathbb{V}$ is empty. Finally, we compute GDI as

$$\mathrm{GDI} = d_0 + \sum \mathcal{D}_i . \quad (4)$$

To demonstrate that the proposed method of GDI calculation satisfies the requirements we

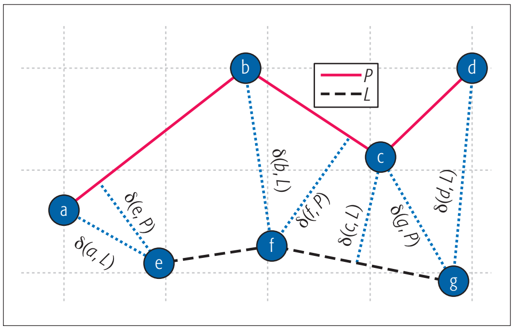


Figure 3. Distance example.

set, we show the GDIs of the route sets defined above (with a grid cell size of ~84 km in Fig. 2) in Table 1. Note, that the produced GDI values fulfill Eq. 1.

Results

In this section we show the geodiversity results we achieved from the measurement data set. First, we show through an illustrative example the difference between the raw `traceroute` outputs and the geographically equivalent routes achievable after applying our route clustering algorithm. Second, we present the compression rates that we were able to attain on the whole measurement set. Finally, we show the GDI results that we calculated for the already clustered route set.

An Example of Route Clustering

An example of the results of our route clustering algorithm is shown in Fig. 4. Arcs represent hops between the localized IP nodes obtained from `traceroute` output. Note, that arcs do not indicate a real link trajectories, merely distinguish routes on the same intermediate links. On the left-hand side one can observe all the routes that the `traceroute` measurements yielded between a source node located in Poland and a destination node located in India (Fig. 4a). Between these two hosts there are seven different routes on the IP-level, but only three geographically different routes (i.e., geo-paths) (Fig. 4b).

Compression Ratio of Route Counts

Stepping up from one example, here we show the overall results in terms of route clustering on the whole measurement dataset. We call the ratio of the number of original routes and the resulting clusters the geo-compression ratio. In Fig. 5 we plot the empirical distribution of this geo-compression ratio for all source and destination host pairs in our dataset.

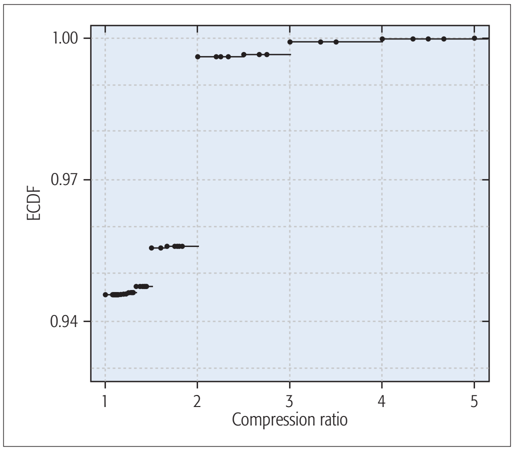


Figure 5. ECDF of compression ratios for all source-destination pairs.

The fraction of source-destination pairs for which the route clustering does not decrease the number of geographically equivalent routes is large (around 80 percent). In these cases the main reason for the poor compression performance is the fact that only one route is yielded by `traceroute`. Another reason, in a less significant number of cases, is that the multiple recorded routes run through nodes at exactly the same locations, as seen earlier. On the remaining 20 percent of source-destination pairs we applied our clustering algorithm with remarkable results: this is shown in Fig. 5. In more than 4 percent of all cases we could obtain a geo-compression ratio higher than 2.

Geo-Diversity Results

Using the GDI metric that we defined earlier to characterize the geo-diversity of routes between a source and destination pair, we show how the calculated values compare to the theoretical maximum of the same metric, taking only the number of geographically equivalent routes and the length of the longest one into account (not their actual trajectories). For this hypothetical maximal value, denoted as MGDI, we place a number of routes forming triangle shapes, the longest one reaching to the highest, so that their GDI would be the largest. In Fig. 6 we plot the distribution of the ratios of GDI over MGDI for those source-destination pairs between which we found at least 2 geographically different routes (20 percent of all pairs, as mentioned above). The results show that

Our goal was to find out where on Earth the packets travel exactly when traffic is carried over the Internet. We discovered that between two given points packet flows are not so scattered as the diversity of traceroute results suggest. We actually saw very few host pairs between which more than 1 geographically different routes exist.

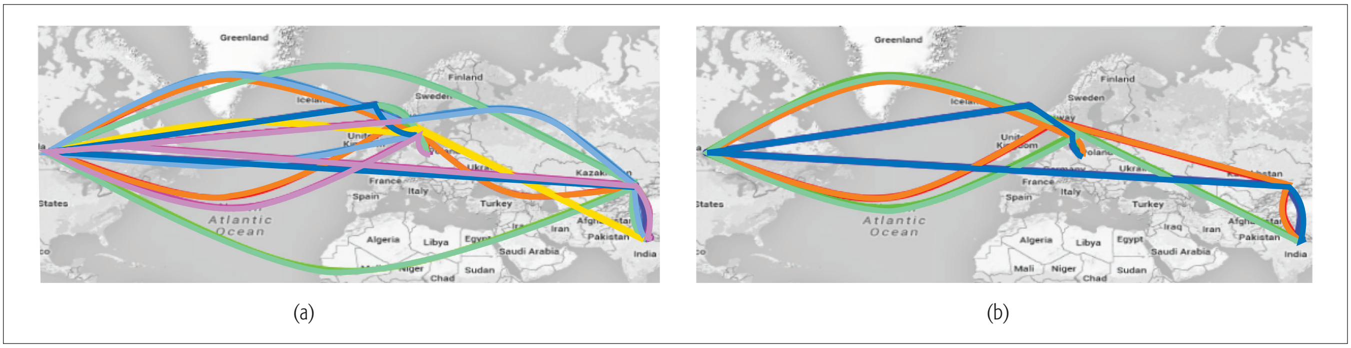


Figure 4. Route count comparison: a) `traceroute` output, route count: 7; b) geodiverse routes, route count: 3.

When network links are not going down individually, but instead are affected en masse due to a regional catastrophe, then the geographic topology of the Internet suffers correlated link failures. In order to be ready for this, planning geographic redundancy of Internet paths can use the results of our method as an input.

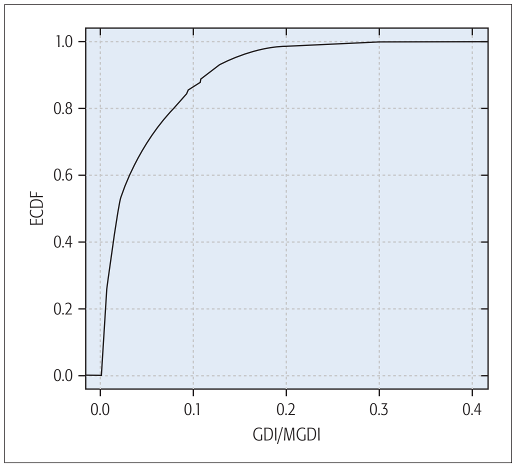


Figure 6. Geo-diversity results given by GDI/MGDI.

for 80 percent of these cases the GDI of routes is less than 10 percent of their MGDI, that is, the theoretically maximum diversity given the number of routes and the length of the longest route.

Conclusion

Our goal was to find out where on Earth the packets travel exactly when traffic is carried over the Internet. We discovered that between two given points, packet flows are not as scattered as the diversity of `traceroute` results suggest. We actually saw very few host pairs between which more than one geographically different routes exist. We showed that 80 percent of endpoints with more than one route have less than a 10 percent MGDI value, which indicates low diversity in terms of geographic distance. The knowledge we gained from this study about the geographical diversity of Internet routes is useful for several applications. In this section we give a few examples of these, and we discuss the weaknesses of our method.

Applications

Estimating Bufferbloat: In order to measure Internet delay correctly, one must fight many sources of inaccuracy: if one-way latency measurement is possible between two hosts, their clocks must be synchronized; if not, several issues come up: misleading `traceroute` results due to load balancing and MPLS tunnels, different return path of the ICMP REPLY when using `ping`, and so on. Indeed, even if the topology is correctly discovered, many aspects of the actual operation of the network equipment pieces affect the measured delay. In order to somehow infer the impact of bufferbloat from the total delay, a very hot topic nowadays, it is important to have an idea about the propagation delay of the packets. Since the propagation delay is closely related to the traveled geographic distance, the sets of geographically equivalent paths discovered by our method provide important input to the analysis of the bufferbloat phenomenon.

Network Resilience: Network resilience, in its classical sense, is a well-studied research domain [30]. When network links are not going down individually, but instead are affected en masse due to a regional catastrophe, whether it be a natural disaster, a power blackout, or an EMP attack, the geographic topology of the Internet suffers correlated link failures. In order to be ready for this, planning geographic redundancy of Internet paths can use the results of our method as an input.

Discussion

After discussing the potential role of our method in various use cases, we turn to the weaknesses of it of which we are aware. First, by applying measurements created by the relatively simple tool `traceroute`, we do not tackle IP-aliasing when building the paths before clustering nodes close to each other. One could argue that performing an already documented merging method targeting IP aliases might yield the same result as the geomerging we do. Second, one might question the accuracy of MaxMind, the tool we use to position the nodes. However, its accuracy is explored in detail in [28], and in long run measurements, we plan to use active measurement based geolocation tools, such as Spotter [31]. Third, it can be argued that estimating geo-paths using straight lines between IP-level nodes may be misleading. However, our city-sized threshold ensures that as long as there are IP-level nodes in close proximity of fiber ducts' ends, this is avoided with high probability. Finally, the proposed GDI and MGDI metrics might seem simplistic, but we argue that to define a diversity metric between routes, many key attributes must be taken into account, and a trade-off between various features and scenarios must be accepted.

References

[1] M. Pathan *et al.*, *A Taxonomy of CDNs*, Springer, 2008.
[2] M. Castro *et al.*, "Topology-Aware Routing in Structured Peer-To-Peer Overlay Networks," *Future Directions in Distributed Computing*, Springer, 2003.
[3] E.-K. Lua *et al.*, "A Survey and Comparison of Peer-to-Peer Overlay Network Schemes," *IEEE Commun. Surveys & Tutorials*, vol. 7, no. 1-4, 2005, pp. 72–93.
[4] A. Greenberg *et al.*, "The Cost of A Cloud: Research Problems in Data Center Networks," *ACM SIGCOMM CCR*, vol. 39, no. 1, 2008 , pp. 68–73.
[5] D. Awduche *et al.*, "Overview and Principles of Internet Traffic Engineering," IETF RFC 3272, tech. rep., 2002.
[6] D. Clark *et al.*, "Interconnection in the Internet: The Policy Challenge," *Research Conf. Commun., Info. and Internet Policy*, 2011.
[7] J. P. Sterbenz *et al.*, "Resilience and Survivability in Communication Networks: Strategies, Principles, and Survey of Disciplines," *Computer Networks*, vol. 54, no. 8, 2010, pp. 1245–65.
[8] J. P. Sterbenz *et al.*, "Evaluation of Network Resilience, Survivability, and Disruption Tolerance: Analysis, Topology Generation, Simulation, and Experimentation," *Telecommun. Systems*, vol. 52, no. 2, 2013, pp. 705–36.
[9] J. Rak, *Resilient Routing in Communication Networks*, Springer, Computer Communications and Networks series, 2015.
[10] C. Kaufman, R. Perlman, and M. Speciner, *Network Security: Private Communication in a Public World*, Prentice Hall Press, 2002.
[11] W. Lou, W. Liu, and Y. Fang, "Spread: Improving Network Security by Multipath Routing," *IEEE MILCOM '03*, 2003, vol. 2, 2003, pp. 808–13.
[12] Y. Schwartz *et al.*, "On the Diversity, Stability and Symmetry of End-to-End Internet Routes," *IEEE INFOCOM Wksps.*, 2010.
[13] W. Mühlbauer *et al.*, "Building an AS-Topology Model that Captures Route Diversity," *ACM SIGCOMM CCR*, vol. 36, no. 4, 2006, pp. 195–206.
[14] R. Teixeira *et al.*, "In Search of Path Diversity in ISP Networks," *ACM SIGCOMM IMC*, 2003.
[15] Y. Shavitt *et al.*, "Geographical Internet Pop Level Maps," *Traffic Monitoring and Analysis*, Springer, 2012, pp. 121–24.
[16] L. Subramanian *et al.*, "Geographic Properties of Internet Routing," *Annual USENIX Tech. Conf.*, 2002.
[17] M. Zhang *et al.*, "How DNS Misnaming Distorts Internet Topology Mapping," *USENIX Annual Tech. Conf.*, General Track, 2006, pp. 369–74.

[18] S. Laki *et al.*, "A Detailed Path-Latency Model for Router Geolocation," *IEEE TridentCom*, 2009; http://ieeexplore.ieee. org/xpls/abs all.jsp?arnumber=4976258
[19] M. Boguná, F. Papadopoulos, and D. Krioukov, "Sustaining the Internet with Hyperbolic Mapping," *Nature Commun.*, vol. 1, 2010, p. 62.
[20] Y. Shavitt *et al.*, "A Structural Approach for PoP Geo-Location," *IEEE INFOCOM Wksps*, 2010.
[21] J. Han *et al.*, "An Experimental Study of Internet Path Diversity," *IEEE Trans. Dependable and Secure Computing*, vol. 3, no. 4, 2006, pp. 273–88.
[22] "Geolocation and Online Fraud Prevention from MaxMind"; http://www. maxmind.com/, accessed Dec. 22, 2014.
[23] H. V. Madhyastha *et al.*, "iplane: An Information Plane for Distributed Services," *Proc. 7th USENIX Symp. Operating Systems Design and Implementation*, 2006, pp. 367–80.
[24] CAIDA, "Macroscopic Internet Topology Data Kit (ITDK)"; http://www. caida.org/data/internet-topology-data-kit.
[25] B. Chun *et al.*, "Planetlab: An Overlay Testbed for Broad-Coverage Services," *ACM SIGCOMM Computer Commun. Review*, vol. 33, no. 3, 2003, pp. 3–12.
[26] B. Augustin *et al.*, "Avoiding Traceroute Anomalies with Paris Traceroute," *ACM IMC*, 2006; http://dl.acm. org/citation.cfm?id=1177100.
[27] S. Laki *et al.*, "Spotter: A Model Based Active Geolocation Service," *IEEE INFOCOM*, 2011; http://ieeexplore.ieee.org/xpls/abs all.jsp?arnumber=5935165
[28] B. Huffaker *et al.*, "Geocompare: A Comparison of Public and Commercial Geolocation Databases Technical Report ," CAIDA tech. rep., May 2011.
[29] Y. Shavitt and N. Zilberman, "A Study of Geolocation Databases, CoRR, vol. abs/1005.5674, 2010; http://arxiv.org/abs/1005.5674.
[30] W. Najjar *et al.*, "Network Resilience: A Measure of Network Fault Tolerance," *IEEE Trans. Computers*, vol. 39, no. 2, 1990, pp. 174–81.
[31] P. Máttray *et al.*, "On the Spatial Properties of Internet Routes," *Computer Networks*, vol. 56, no. 9, 2012, pp. 2237–48; http://linkinghub.elsevier.com/retrieve/pii/S1389128612000989.

Biographies

Attila Csoma (csoma@tmit.bme.hu) received his B.Sc. and M.Sc degrees in informatics from Budapest University of Technology and Economics (BME), Hungary, in 2010 and 2012 respectively. Since 2012, he has been a Ph.D. student at the High Speed Networks Laboratory at the Department of Telecommunications and Media Informatics, BME. He conducts his Ph.D. research on network topology and complex networks. His research interests also include network coding and IoT.

László Toka (toka@tmit.bme.hu) is currently a postdoctoral researcher at the Hungarian Academy of Sciences, and works as an assistant professor in the Department of Telecommunications at BME. He received his Ph.D. in 2011 from Telecom Paris, and afterwards worked at Ericsson Research for three years before rejoining academia. His main research topics lie within software defined networking, data analytics, and economic modeling of distributed systems.

András Gulyás (gulyas@tmit.bme.hu) received M.Sc. and Ph.D. degrees in informatics from BME in 2002 and 2008, respectively. Currently he is a research fellow in the Department of Telecommunications and Media Informatics. His research interests are complex and self-organizing networks and software defined networking.

Cyber Meets Control: A Novel Federated Approach for Resilient CPS Leveraging Real Cyber Threat Intelligence

Elias Bou-Harb, Walter Lucia, Nicola Forti, Sean Weerakkody, Nasir Ghani, and Bruno Sinopoli

While almost all works in the literature exclusively tackled the security of one independent aspect of CPS (i.e., cyber or physical), the authors argue that these systems cannot be decoupled. In this context, they present what they believe is a first attempt ever to tackle the problem of CPS security in a coupled and a systematic manner. To this end, this article proposes a novel approach that federates the cyber and physical environments to infer and attribute tangible CPS attacks.

Abstract

Cyber-physical systems (CPS) embody a tight integration between network-based communications, software, sensors, and physical processes. While the integration of cyber technologies within legacy systems will most certainly introduce opportunities and advancements not yet envisioned, it will undoubtedly also pave the way to misdemeanors that will exploit systems' resources, causing drastic and severe nationwide impacts. While almost all works in the literature exclusively tackled the security of one independent aspect of CPS (i.e., cyber or physical), we argue that these systems cannot be decoupled. In this context, we present what we believe is a first attempt ever to tackle the problem of CPS security in a coupled and a systematic manner. To this end, this article proposes a novel approach that federates the cyber and physical environments to infer and attribute tangible CPS attacks. This is achieved by

- Leveraging real cyber threat intelligence derived from empirical measurements.
- Capturing and investigating CP data flows by devising an innovative CPS threat detector.

An added value of the proposed approach is rendered by physical remediation strategies, which are envisioned to automatically be invoked as a reaction to the inferred attacks to provide CPS resiliency. We conclude this article by discussing a few design considerations and presenting three case studies that demonstrate the feasibility of the proposed approach.

Introduction

Critical infrastructure systems are indispensable to the broader health, safety, security, and economic well-being of modern society and governments. In recent years, many of these systems, such as smart grids, nuclear plants, and automated highway systems, have been undergoing large-scale transformations with the infusion of new "smart" cyber-based technologies to improve their efficiency and reliability. These transitions are being driven by continual advances and cost-efficiencies in areas such as integrated networking, information processing, sensing, and actuation. Hence, increasingly, physical infrastructure devices and systems are being tasked to co-exist and seamlessly operate in cyber-based environments. Indeed, tightly coupled systems that exhibit this level of integrated intelligence are often referred to as Cyber-physical systems (CPS) [1].

Nowadays, CPS can be found in significantly diverse industries, including, but not limited to, aerospace, automotive, energy, healthcare, and manufacturing. Undeniably, the development and adoption of such CPS will generate unique opportunities for economic growth and improvement of quality of life. While CPS presents great opportunities, their complexity, which arises from the fusion of computational systems with physical processes, indeed poses substantial security challenges. To this end, novel vulnerabilities will manifest themselves, leading to attack models that are fundamentally new and hard to infer, attribute, and analyze.

Indeed, historical events confirm that industrial control systems have long been the target of disruptive cyber attacks. For instance, in 2010, the prominent Stuxnet malware was employed to target the SCADA control system of a critical uranium enriching facility, which triggered immense plant damage and even endangered human life [2]. Most recently, in March 2016, another malware was inferred to be responsible for the massive power outage that struck Ukraine in December 2015. Given the rapid transformation of industrial control systems toward CPS-based setups, attacks are anticipated to increase in frequency, sophistication, and target diversity. In fact, the latter trend was recently confirmed by the U.S. Department of Homeland Security (DHS), when they published the statistics in Fig. 1, revealing thousands of CPS attacks targeting diverse sectors [3].

Motivated by the imminent threats targeting CPS in addition to the lack of security approaches that tackle both aspects of such systems in a coupled and a coherent manner [4, 5], we frame the contributions of this article as follows:

•Proposing a new multidisciplinary approach that strives to diminish the gap between cyber security and control systems' science for securing CPS. Contrary to theoretical approaches that only consider the physical aspects of CPS in which some assumption is made regarding an attack

Elias Bou-Harb is with Florida Atlantic University and the National Cyber Forensic and Training Alliance (NCFTA) of Canada.
Walter Lucia is with Concordia University. Nicola Forti is with the University of Florence. Sean Weerakkody and Bruno Sinopoli are with Carnegie Mellon University.
Nasir Ghani is with the University of South Florida, and the Florida Center for Cybersecurity.

Digital Object Identifier:
10.1109/MCOM.2017.1600292CM

and its corresponding countermeasure, the proposed approach uniquely exploits tangible cyber threat intelligence (CTI) to infer real attack scenarios that could realistically affect CPS. Further, the proposed approach also considers the dynamics of CPS by triggering prompt remediation strategies in the physical realm as a reaction to the inferred attacks. To the best of our knowledge, the devised capability is of high impact in the CPS literature and has never been attempted before.

•Generating insightful CTI related to CPS by employing approximately 40 million real malware samples and undergoing dynamic binary analysis to capture CPS insider threats. The latter CTI is further expanded by characterizing and attributing real CPS external attacks by means of designing and deploying a high-interactive CPS honeypot.

•Evaluating and validating the feasibility and effectiveness of the integrated proposed approach by presenting three case studies that depict three different CPS attack scenarios. To this end, we also discuss some insightful design considerations that we expect to be guiding and helpful in the realization of future CPS security approaches.

The organization of this article is as follows. In the next section, we review some CPS security related works. We present the proposed approach by thoroughly discussing each of its components and pinpointing several design considerations. We demonstrate the effectiveness and practicality of the proposed approach by means of three different proof of concept case studies. Finally, closing remarks and future research directions conclude this article.

Related Work

In this section, we review a number of related works in the context of security challenges in CPS, with special emphasis on control-theoretic, cyber, and hybrid approaches for securing CPS.

Research challenges related to CPS security have been addressed in prior works [4], where unprecedented CPS vulnerabilities and threats were investigated, and new directions for securing control systems were presented. Moreover, the authors of [6] highlighted the need for collaborative approaches that better integrate security into the core design of CPS.

From a control-theoretic perspective, Teixeira *et al.* [7] have introduced and modeled different attack scenarios such as false data injections, replay, and zero-dynamics attacks, where adversarial activities attempt to cause damage to the controlled system while remaining stealthy. Furthermore, Mo *et al.* [5] proposed an active detection method, known as physical watermarking, to authenticate the nominal behavior of a control system. To this end, a known noisy control input is injected to detect replay attacks by analyzing the output of the system. In another closely related work, Weerakkody *et al.* [8] introduced time-varying dynamics, acting as a moving target, to detect integrity attacks. Additionally, Fawzi *et al.* [9] focused on the design, implementation, analysis, and characterization of robust estimation and control in CPS when they are affected by corrupted sensors and actuators.

From a cyber perspective, Caselli *et al.* [10] presented a sequence-aware intrusion detection system that aims at detecting CPS semantic attacks using real empirical measurements from

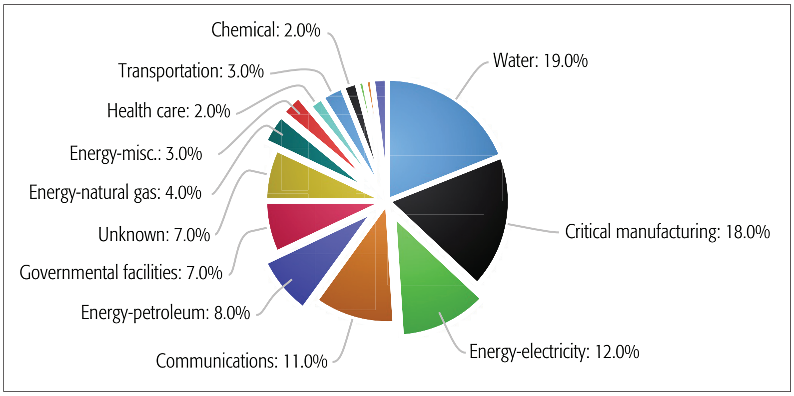


Figure 1. Recent attacks on cyber-physical systems as reported by DHS [3].

a water treatment facility. In an alternative work, Zonouz *et al.* [11] investigated malware that specifically target programmable logical controllers (PLCs) in CPS environments. The authors proposed a set of big data analytics rooted in symbolic execution that aim at capturing the behavior of malicious code.

Complementary hybrid security approaches that attempt to systematically combine cyber and control capabilities are particularly rare. One interesting research specimen was proposed by Zonouz *et al.* [12]. In this work, the authors present an approach that aims at performing detection of corrupted measurements in power grids. Specifically, the proposed approach exploits alert notifications from intrusion detection and firewall systems to generate attack graphs providing an estimate of the compromised set of power grid hosts. Although such an approach leverages information from both the cyber and physical realms, it is neither capable of inferring specific types of attacks nor attributing the attack. Moreover, the proposed method is primarily focused on detection, and thus does not provide any tangible approach on how to provide CPS resiliency during or immediately after an attack.

Proposed Architecture

In this section, we present and elaborate on the components of our proposed approach as depicted in Fig. 2. The core intuition behind the devised architecture is the unique introduction of the notion of *true maliciousness* to CPS security research. This notion embodies tangible malicious CPS attacks that could realistically affect the stability and security of CPS. The rationale behind this concept is threefold. First, the majority of literature approaches that solely focus on the physical aspects of CPS tend to characterize anomalies by a *deviation* of observed data in comparison to its expected value, generating a significant amount of false positives. Second, it is known that malicious empirical CPS security data from within operational CPS settings is extremely rare [10]. Third, it would be desirable to have an architecture that also provides insights and inferences that would aid in attribution. Undeniably, such attribution evidence will be leveraged to build highly-effective countermeasures to provide CPS resiliency. In the following section, we present and discuss the components of such an architecture that strives to exploit tangible CTI to infer and mitigate real CPS attack scenarios.

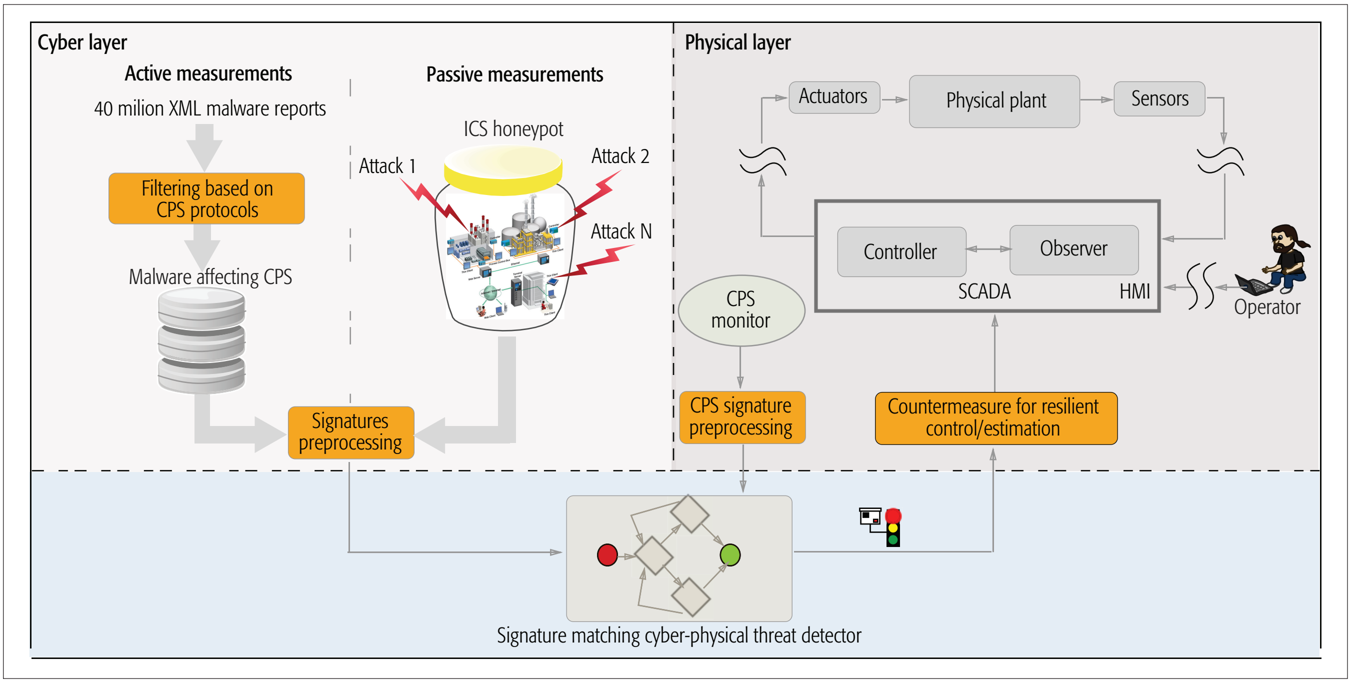


Figure 2. A holistic perspective of the proposed architecture.

Cyber Layer

Threats toward CPS could arise from external as well as internal entities. On one hand, an example of an external threat could be rendered by scanning activities [13], originating from the Internet, in an attempt to probe and quantify CPS vulnerabilities, to be exploited in a subsequent directed attack. On the other hand, the Stuxnet malware would be an accurate example of an insider CPS threat, which leveraged system specific knowledge to execute a stealthy attack from within the boundaries of the CPS. Motivated by such diverse threats, this component aims at capturing real malicious CPS attack signatures for both internal and external threats. In this context, we define attack signatures by a series of consecutive attackers' steps that constitute a well-defined attack scenario. In the following, we describe how the latter CTI is generated.

Active Measurements: It is known that malware attacks similar to Stuxnet pose one of the most debilitating internal threats to CPS. To this end, we resort to dynamic malware binary analysis as depicted in Fig. 3 to generate real attack signatures that aim at capturing malware attack scenarios targeting CPS. We are fortunate to have access to malware data provided by Team Cymru Research. Consistent with Fig. 3, each malware binary sample is executed in a controlled client environment. During execution, the client would monitor and record the executed activities by the malware sample at the network and system levels. Consequently, the server processes such received information by producing an XML report, which summarizes the activities of the executed malware. Moreover, to select only those malware that specifically target CPS, we further execute a filtering mechanism on the indexed XML reports. Such filtering mechanisms can be easily modified to match CPS protocols and systems used in particular sectors.

Passive Measurements: It is also desirable to possess attack signatures related to external CPS attacks. Undeniably, the optimal approach to achieve this is to employ traffic capturing, measurement and analysis from the external boundaries of an operational CPS. However, due to legal, logistic, and privacy constraints, the latter is not always feasible. For such reasons, we resort to the concept of a honeypot: a set of software modules that can realistically imitate the components and the operations of any CPS. Further, a honeypot could be easily modified and tuned to generically exploit any CPS role within a specialized sector, allowing the capturing of a wide range of tailored external attacks. To this end, we design and implement a honeypot based on the open source industrial control system project dubbed as Conpot. To provide more realism to the CPS honeypot, we have implemented a custom capability that emulates the plant dynamics and configured the honeypot to operate several generic CPS protocols, with a human machine interface (HMI) and Simple Network Management Protocol (SNMP) capabilities. We deployed the honeypot online for a specific duration and enabled high levels of logging. To infer external attacks, we exploited the log files to build the attack strategies by tracking one attacker at a time, throughout its interaction with the honeypot. Please note that there might exist zero-day or unknown attacks that possibly would not be captured by the active and passive measurement modules. Future work will address how to remedy this issue.

Physical Layer

This module deals with the physical (i.e., control) aspects of CPS. These include hardware and software modules capable of:

- Characterizing the physical process
- State estimation and reconstruction
- Stabilizing the physical plant
- Monitoring and managing the system

It is worthy to note that this module is generic to any CPS environment, including those operating in diverse domains such as power, wireless sensor networks, or manufacturing. To support CPS attack detection and mitigation, we extend this layer by introducing an additional component that we refer to as the CPS monitor, as depicted in Fig. 2. The monitor exploits CPS communication channels and protocols to tap, gather, and amalgamate cyber-physical (CP) data flows that are circulating through a CPS. Moreover, the monitor coordinates with the CP threat detector to generate attack-resilient estimation and control remediation strategies, which are invoked as a reaction to an inferred attack.

Cyber-Physical Threat Detector

This component lies at the intersection between the cyber and physical layers. Indeed, the core aim of the CP threat detector module is to investigate whether the CP data flows extracted from the CPS communication channels are susceptible to any tangible external or internal threats. An imperative auxiliary aim of this module is to characterize the severity of any inferred attack.

To achieve the intended tasks, the CP threat detector initially models both the malicious attack signatures generated from the cyber layer and the CP data flows extracted from the physical layer, into a common structure that we refer to as semantic behavioral graphs. Such directed graphs capture the activity performed as well as the semantics of such actions in the context of the observed protocol. To clarify the notion of semantic behavioral graphs, consider Fig. 4, which depicts a miniature specimen of such a graph capturing the dynamics of a benign CP data flow of the Siemens proprietary communication s7comm.

The CP threat detector proceeds in an attempt to infer any similarities between semantic graphs, as an indicator of an ongoing malicious activity on the CPS. However, given the fact that such graphs could possibly be of large scale due to the excessive captured/modeled network activity, and of high dimensionality due to the appended feature vectors, computing graph similarities in practice would indeed be challenging. In an attempt to overcome these issues, we devise a twofold approach.

First, the CP threat detector applies the notion of graph kernels borrowed from [14] on the formed behavioral graphs (discarding any semantics at this stage). The rationale here is to transform the similarity computation procedure of complex graphs into a linear space. To achieve this, sub-graphs from the behavioral graphs are initially extracted based on a certain criterion. Subsequently, compact representations of the created sub-graphs are generated based on a specific fingerprinting approach. Lastly, a mapping technique is employed to transform the latter representations into a linear space defined by a kernel matrix [14, 15], where the similarity computation is executed in linear time.

Second, we employ a threshold mechanism to deem when there is a significant similarity between those two types of graphs in order to flag that an ongoing malicious activity might be occurring on the physical plant. In this work, we set 60 percent as a conservative similarity threshold, to indicate a possible attack. To further reduce any other false positive cases and to confirm the attack, the CP threat detector proceeds by investigating the semantics of those graphs that exceed the similarity threshold. To this end, the CP threat detector performs binary comparisons between each corresponding pair of features of the semantic vectors. Currently, the header layer is only used in the comparisons for efficiency purposes; however, this can be easily extended to include the protocol layer and data layer. Any similarity between the semantics of the graphs of the CP data flows and the semantics of the graphs of the malicious signatures will likely confirm the existence of an ongoing attack. In this context, it is important to note that:

- We employ the similarity measure as a severity score.
- We are capable of providing tangible evidence attributing the attack to a specific malware specimen, in case the matching semantic behavioral graph was captured from the active measurements of the cyber layer.

Please note that the components of the proposed architecture that were discussed are fully automated, from an implementation perspective.

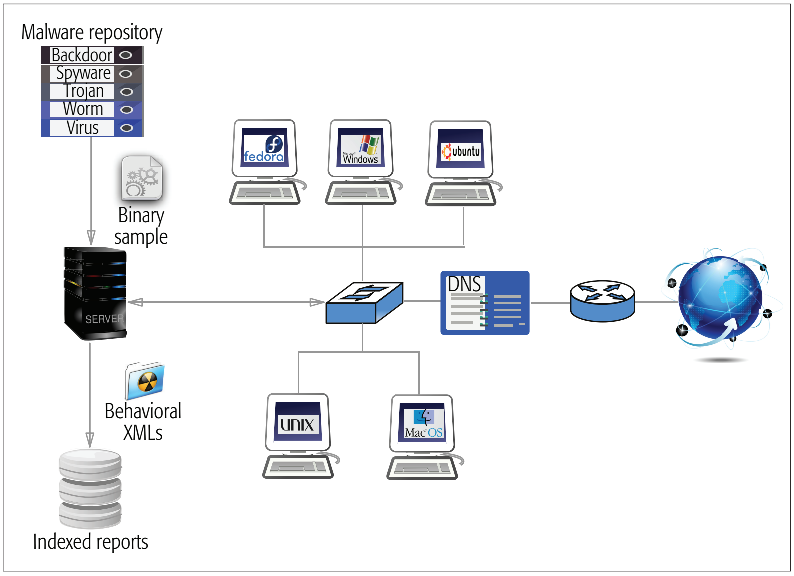


Figure 3. Dynamic malware binary analysis to capture real internal CPS threats.

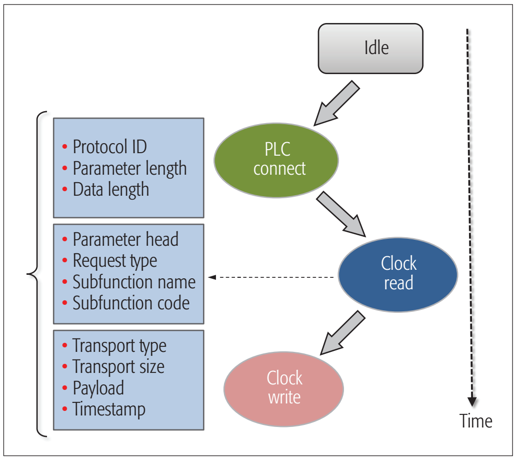


Figure 4. A high-level depiction of a semantic behavioral graph of a Siemens communication protocol.

While CP data flows in the physical layer are characteristically in the order of milliseconds or seconds, attacks in the cyber realm require time intervals of larger magnitude. Thus, when designing the CP threat detector, as detailed earlier, this issue ought to be taken into account.

Design Considerations

The proposed architecture of Fig. 2 introduces a new paradigm that is rendered by the tight coupling and systematic fusion of the cyber and physical layers in the context of CPS security. Given that such an approach has never been attempted before, we thought that it would be of added value in this article to provide some design blueprints that elaborate on a few considerations related to the proposed scheme.

On Time-Scale Discrepancy: Architectures that aim at inferring, characterizing, and mitigating tangible attacks on CPS should pay special attention to time-scale discrepancies. While CP data flows in the physical layer are characteristically in the order of milliseconds or seconds, attacks in the cyber realm require time intervals of larger magnitude. Thus, when designing the CP threat detector, as detailed earlier, this issue ought to be taken into account. A feasible solution could be rendered by an approach that operates in a sliding time-window fashion; sampling periods from the physical realm could be employed in conjunction with attack signatures derived within specific time-window cycles. In this context, the time-window duration would be treated as a system parameter that would be tuned in order to achieve a balance between attack inference rates and computational load.

On Detection Practicality: Depending on the quantity and type of the derived CTI, a general design hurdle would be how to build effective and efficient models of such CTI that can be employed for detection. Commonly, attack detection and remediation in CPS realms require some near real-time requirements; any significant detection delay or poorly-timed reactions might result in cascading failures or damage of system components. Although in this work, we devised a CP threat detector based on efficient graph models, other approaches that have yet to be investigated could be more practical, easier to manage, and provide stronger analytical capabilities. Further, given that the extracted CTI could be extensive, as we expect and have observed throughout this work, one should initially model it offline and subsequently incrementally enrich it as new CTI becomes available.

On Cyber-Physical Countermeasures: Any approach that aims at achieving CPS resiliency should endeavor to provide cyberphysical remediation strategies to combat the effects of the inferred malicious activity. In this context, we advocate and stress the importance of designing countermeasures that cooperatively leverage information and capabilities from both the cyber and physical realms. Indeed, if only cyber mitigation strategies are executed, then the safety of the CPS under an attack cannot be guaranteed. Conversely, if only physical countermeasures are adopted, an attack-free CPS environment would be impossible to achieve. In this work, we attempted to capture the latter desired features by introducing the notion of semantic behavioral graphs as a modeling and a detection approach in the context of CP data flows and malicious attack signatures. Another desirable design goal to be considered would be the ability to characterize attacks through severity metrics, which would be highly beneficial from two perspectives, namely, situational awareness and prioritized mitigation. Additionally, one should postulate CTI approaches that not only can infer attacks, but more imperatively, can generate attack signatures that aim at disclosing attackers' strategies, aims, and intentions. Indeed, the active measurements approach described earlier, which exploits real malware strategies, attempts to achieve exactly the latter. In this case, the concrete knowledge of the disruptive resources available to the attacker allows the design of more effective cyber-physical countermeasures. Specifically, the observer and control modules could be re-designed by discarding the information originating from the corrupted CPS channels. Moreover, if an estimate of the attack vector is available, a resilient control solution can be achieved by means of an adaptive compensator, which simply adds a compensating term to the nominal control signal, avoiding redesign of the controller. Simultaneously, in the cyber realm, various information technology (IT) security countermeasures can be rapidly deployed to minimize attackers resources, and thereby limit their damage to targeted CPS assets. Such strategies could include dropping network traffic originating from attackers' IP addresses or dynamically adapting firewall rules, among numerous other available techniques.

Proof of Concept: Case Studies

In this section, we verify the effectiveness of the proposed architecture by demonstrating its capability in generating tangible cyber threat intelligence related to CPS environments. Furthermore, we consider three case studies that capture two real external CPS attacks and an internal CPS attack launched by the eminent Stuxnet malware.

By operating the customized CPS honeypot as described previously for almost one month, we were able to infer around 500 unique attackers generating thousands of diverse malicious activities. By executing IP geolocation, Fig. 5a illustrates the countries where these attacks originated from, while Fig. 5b shows the Internet service providers (ISPs) responsible for some of those attacks in a one-day specimen. Note that we are aware that the University of Michigan performs regular benign scanning attempts toward various Internet services, including CPS services, and thus its appearance on this list verifies and validates the setup of the CPS honeypot and its capability in inferring malicious attempts. It is also worthy to mention that the statistics behind Fig. 5 could be relatively skewed by the abundant use of spoofing attacks. Indeed, given that IP spoofing is still present on almost 20 percent of the Internet autonomous systems, the reported statistics should not be taken as an absolute representation but rather as a figurative and a relative view of the status of CPS security in the context of real cyber threat intelligence.

We also had a brief chance to observe and investigate the network traffic packets arriving at the CPS honeypot. Our analysis revealed a staggering 10,000 generic TCP and UDP scanning attempts and close to 2,000 TCP flooding denial of service attacks on various CPS communication protocols, including those targeting the open source DNP3 and Modbus CPS protocols. We also generate supplementary material related to such misdemeanors including geo-location information per source, organization, city, and region. Although we refrain from publishing this informa-

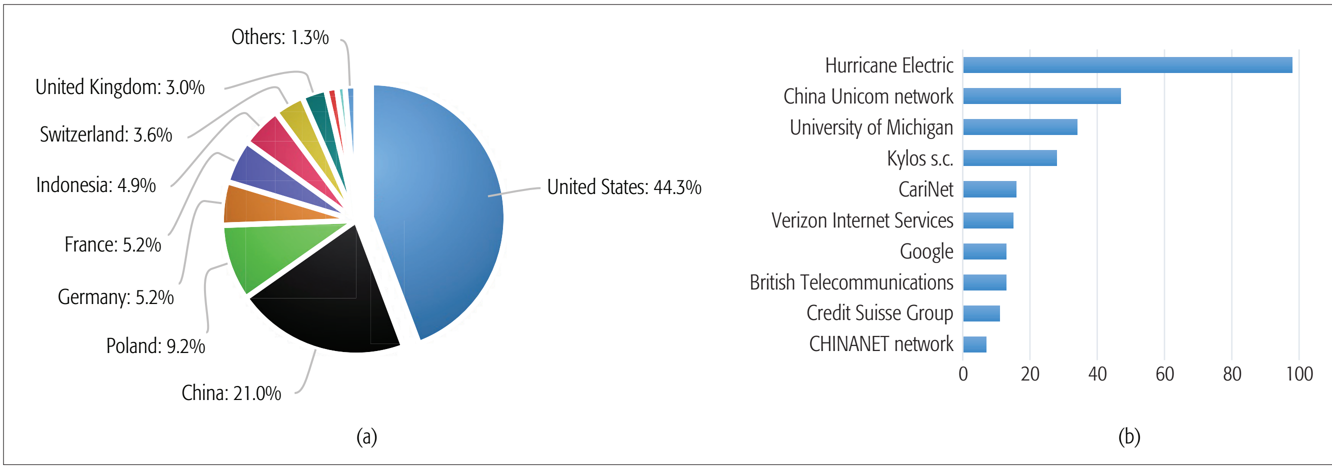


Figure 5. Real external CPS threats as inferred by the CPS Honeypot: a) distribution of all inferred threats per originating country; b) threats from top 10 Internet service providers during a 24-hour period.

tion due to sensitivity/legal issues, we can note that these attempts originated from 177 diverse operational providers, 124 distinct ISPs, and 71 different cities. Given that our analyzed period is relatively short, we concur that all the inferred attacks in terms of source diversity, target protocol diversity, frequency, and machinery are quite interesting and alarming.

We proceed by selecting and reporting on three real case studies that the proposed architecture was able to infer and remediate. The first case study captured an external CPS attacker that was inferred by employing the CPS honeypot. This scenario represented an attacker who attempted to gain elevation of privilege, and hence complete control over the plant, by striving to exploit the session manager of the CPS HMI. Part of this attack's inferred request is illustrated in Fig. 6a. Given that such an attack did not threaten the physical/control aspect of the CPS, a simplistic yet effective cyber strategy is applied to mitigate the attack by blocking any subsequent traffic from the attacker's IP address.

The second case study pinpointed another CPS external attack that was also perceived by leveraging the CPS honeypot. In this scenario, the attacker initially scanned the CPS plant using the SNMP to retrieve the map of all operational services. Once this has been achieved, the attacker read the Modbus variables and subsequently crafted some invalid input that is beyond a specified safe range. The latter attack appears to be an attempt to cause some sort of damage to the CPS equipment. Part of this attack's captured request is illustrated in Fig. 6b. In this attack scenario, a physical control remediation is obtained by simply discarding the injected malicious data, while a cyber countermeasure is enforced by filtering any further incoming packets from the attacker's IP address.

The third case study captured an instance of an internal CPS threat. In this attack scenario, detailed experimentation was conducted using a sample of the Stuxnet malware by closely following the proposed mechanism from earlier. Indeed, this malware employs replay attacks in an attempt to replicate previously-recorded sensor measurements to a system operator, while also injecting damaging inputs to the system. It is well known that this type of attack is stealthy to any passive

```
A. HTTP/1.1 GET request from ([redacted]', 3952): ('/index.html', ['[redacted]\r\n',
   'Connection: keep-alive\r\n', 'Cache-Control: max-age=0\r\n', 'Accept:
   text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8\r\n', 'Upgrade-
   Insecure-Requests: 1\r\n', 'User-Agent: Mozilla/5.0 (Windows NT 10.0; WOW64)
   AppleWebKit/537.36 (KHTML, like Gecko) Chrome/46.0.2490.71 Safari/537.36\r\n', 'Accept-
   Encoding: gzip, deflate\r\n')

B. New snmp session from[redacted] (3b845bb1-8111-4966-aac9-c0ddd0203b8e)
   SNMPv1 GetNext request from ('[redacted]', 28487): 1.3.6.1.2.1.1.2
   SNMPv1 response to ('[redacted]', 28487): 1.3.6.1.2.1.1.2.0 1.3.6.1.4.1.20408
   SNMPv1 GetNext request from ('[redacted]', 28487): 1.3.6.1.2.1.43
   SNMPv1 response to ('[redacted]', 28487): 1.3.6.1.2.1.43
   New Modbus connection from [redacted].38:55237. (2260e5fa-de68--4ff2-a03d-a79945ed46e2)
   Modbus traffic from [redacted].38: {'function_code': None, 'slave_id': 0, 'request':
   '133700000005002b0e0100', 'response': ''}
```

Figure 6. Snapshot of the requests generated by the external CPS attacks as inferred by the CPS honeypot.

physical detector because the resulting CP flows cannot be distinguished from benign CP flows characterizing an attack-free scenario. Nevertheless, in this case, the CP threat detector readily inferred the occurrence of an ongoing malicious activity in the physical realm since it captured the behavior and semantics of the malware as it interacted with the CPS. Furthermore, the proposed architecture was capable of attributing such activity to the exact sample of the Stuxnet malware, namely, Worm.Win32.Stuxnet.b . We also note the captured/modeled behavior of this malware, which included unauthorized reading and writing of the sensors' measurements as well as unauthorized writing toward CPS actuators' channels.

Indeed, the aforementioned case studies demonstrate the effectiveness of the proposed architecture in disclosing attackers' strategies and actions for both internal and external CPS attacks. More imperatively, the proposed approach also provides invaluable, tangible forensic evidence that can be employed for attribution in the cyber realm and resiliency in the physical realm.

Concluding Remarks and Future Research Directions

Following our vision to diminish the gap between highly theoretical solutions and practical approaches for securing CPS, this article uniquely proposed a federated approach for resilient CPS by exploiting real CTI. The core rationale behind

The core rationale behind the proposed architecture is to derive tangible CPS attack models from empirical measurements, which can be employed to infer and attribute real CPS attacks. An innovative CP threat detector amalgamated CP data flows from the physical realm with attack signatures from the cyber realm.

the proposed architecture is to derive tangible CPS attack models from empirical measurements, which can be employed to infer and attribute real CPS attacks. An innovative CP threat detector amalgamated CP data flows from the physical realm with attack signatures from the cyber realm to infer and score real attack scenarios, as well as provide evidence of attribution and a means to generate CPS resiliency mechanisms.

This effort presents a solid basis from which to expand into other directions in CPS security. Foremost, one thrust will be dedicated to automating the tasks associated with creating resiliency countermeasures and algorithms from the captured CTI. Another aim is to investigate and develop additional CP threat detector designs and types and evaluate their detection tradeoff and efficiency under various realistic attack scenarios. Finally, an open source utility is also being designed to incorporate the overall notions and ideas behind the proposed architecture. This latter deliverable will help facilitate the incorporation, rapid prototyping, and much-needed evaluation of this solution in real-world operational CPS environments.

References

[1] Xu Li *et al.*, "Smart Community: An Internet of Things Application," *IEEE Commun. Mag.*, vol. 49, no. 11, 2011, pp. 68–75.

[2] F. Kargl *et al.*, "Insights on the Security and Dependability of Industrial Control Systems," *IEEE Security & Privacy*, vol. 6, 2014, pp. 75–78.

[3] The Industrial Control Systems Cyber Emergency Response Team (ICS-CERT). https://ics-cert.us-cert.gov/.

[4] A.A. Cárdenas, S. Amin, and S. Sastry, "Research Challenges for the Security of Control Systems," *Proc. 3rd Conf. Hot Topics in Security*, Berkeley, CA, USA, 2008, pp. 61–66.

[5] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical Authentication of Control Systems: Designing Watermarked Control Inputs to Detect Counterfeit Sensor Outputs," *IEEE Control Systems*, vol. 35, no. 1, Feb. 2015, pp. 93–109.

[6] C. Neuman, "Challenges in Security for Cyber-Physical Systems," *DHS: S&T Wksp. Future Directions in Cyber-Physical Systems Security*, vol. 7, Citeseer, 2009.

[7] A. Teixeira *et al.*, "A Secure Control Framework for Resource-Limited Adversaries," *Automatica*, vol. 51, 2015, pp. 135–48.

[8] S. Weerakkody and B. Sinopoli, "Detecting Integrity Attacks on Control Systems Using a Moving Target Approach," *54th IEEE Conf. Decision and Control (CDC)*, 2015, pp. 5820–26.

[9] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure Estimation and Control for Cyber-Physical Systems Under Adversarial Attacks," *IEEE Trans. Automatic Control*, vol. 59, no. 6, June 2014, pp. 1454–67.

[10] M. Caselli, E. Zambon, and F. Kargl, "Sequence-Aware Intrusion Detection in Industrial Control Systems," *Proc. 1st ACM Wksp. Cyber-Physical System Security*, 2015, pp. 13–24.

[11] S. Zonouz, J. Rrushi, and S. McLaughlin, "Detecting Industrial Control Malware Using Automated PLC Code Analytics," *IEEE Security Privacy*, vol. 12, no. 6, Nov. 2014, pp. 40–47.

[12] S. Zonouz *et al.*, "SCPSE: Security-Oriented Cyber-Physical State Estimation for Power Grid Critical Infrastructures," *IEEE Trans. Smart Grid*, vol. 3, no. 4, Dec. 2012, pp. 1790–99.

[13] E. Bou-Harb, M. Debbabi, and C. Assi, "Cyber Scanning: A Comprehensive Survey," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 3, 2014, pp. 1496–519.

[14] S. V. N. Vishwanathan *et al.*, "Graph Kernels," J. Mach, *Learn, Res.*, vol. 11, Aug. 2010, pp. 1201–42.

[15] A. Teixeira *et al.*, "Min-Hash Fingerprints for Graph Kernels: A Trade-Off Among Accuracy, Efficiency, and Compression," *J. Information and Data Management*, vol. 3, no. 3, 2012, pp. 227.

Biographies

Elias Bou-Harb (ebouharb@fau.edu) is currently an assistant professor in the Computer Science Department at Florida Atlantic University, where he directs the Cyber Threat Intelligence Laboratory. Previously, he was a visiting research scientist at Carnegie Mellon University. He is also a research scientist at the National Cyber Forensic and Training Alliance (NCFTA) of Canada. He holds a Ph.D. degree in computer science from Concordia University, Montreal, Canada. His research and development activities and interests focus on the broad area of operational cyber security, including attack detection and characterization, Internet measurements, cyber security for critical infrastructure, and big data analytics.

Walter Lucia is an assistant professor at the Concordia Institute for Information Systems Engineering, Concordia University, Canada. He received an M.Sc. in automation engineering and a Ph.D. in systems and computer engineering from the University of Calabria, Italy. He was a visiting research scholar in the ECE Department at Northeastern University, USA, and a visiting postdoctoral researcher in the ECE Department at Carnegie Mellon University, USA. His current research interests include control of unmanned vehicles, switching systems, fault-tolerant control, model predictive control, and resilient control of cyber-physical systems.

Nicola Forti received his B.Sc. degree in mechanical engineering and a M.Sc. degree in automation engineering from the University of Florence, Florence, Italy, in 2009 and 2013, respectively. He is currently pursuing his Ph.D. degree in information engineering with the Automatic Control group at the University of Florence. In 2015 he visited the Department of Electrical and Computer Engineering at Carnegie Mellon University in Pittsburgh, PA, USA for one year. His main research interests include networked control systems, data fusion, secure estimation and control of cyber-physical systems, and multi-target tracking.

Sean Weerakkody received the B.S degree in electrical engineering and mathematics from the University of Maryland, College Park, USA, in 2012. He was awarded the National Defense Science and Engineering Graduate fellowship in 2014. He is currently currently pursuing the Ph.D. degree in electrical and computer engineering at Carnegie Mellon University, Pittsburgh, PA. His research interests include secure design and active detection in cyber-physical systems and estimation in sensor networks.

Nasir Ghani [SM] is a professor in the Department of Electrical Engineering at the University of South Florida (USF), and also a research liaison for the state-wide Florida Center for Cybersecurity (FC2). He has over 20 years of research and development experience in the academic and industry sectors. Earlier he was associate chair of the Electrical and Computer Engineering Department at the University of New Mexico, and prior to that, a faculty member at Tennessee Tech University. He has also spent several years working in industry (Nokia, IBM, and Motorola) and several hi-tech startups. Dr. Ghani has co-authored over 200 publications, and also has three highly-cited U.S. patent inventions. His research interests include high-speed networks, cybersecurity, cyberphysical systems, cloud computing, and disaster recovery. Currently he is involved in various projects on cybersecurity for the Internet of Things (IoT) and software defined networks. His work has been funded by the National Science Foundation (NSF), Department of Energy, Department of Education, Qatar Foundation, and Sprint-Nextel Corporation. He also received the NSF CAREER award in 2005. Dr. Ghani has served as an associate editor for *IEEE Communications Letters*, *IEEE/OSA Journal of Optical and Communications and Networking*, and *IEEE Systems*. He has also guest-edited special issues of *IEEE Network* and *IEEE Communications Magazine* and has co-chaired numerous symposia for IEEE GLOBECOM, IEEE ICC, and IEEE ICCCN. He also served as chair of the IEEE Technical Committee on High Speed Networks (TCHSN) from 2007-2010. He received his bachelor's degree from the University of Waterloo, his master's degree from McMaster University, and his Ph.D. degree from the University of Waterloo.

Bruno Sinopoli received the Dr. Eng. degree from the University of Padova in 1998, and his M.S. and Ph.D. in electrical engineering from the University of California at Berkeley, in 2003 and 2005, respectively. After a postdoctoral position at Stanford University, Dr. Sinopoli joined the faculty at Carnegie Mellon University, where he is a professor in the Department of Electrical and Computer Engineering with courtesy appointments in mechanical engineering and in the Robotics Institute and co-director of the Smart Infrastructure Institute, a research center aimed at advancing innovation in the modeling analysis and design of smart infrastructure. Dr. Sinopoli was awarded the 2006 Eli Jury Award for outstanding research achievement in the areas of systems, communications, control, and signal processing at U.C. Berkeley, the 2010 George Tallman Ladd Research Award from Carnegie Mellon University, and the NSF Career award in 2010. His research interests include the modeling, analysis, and design of secure by design cyber-physical systems with application to interdependent infrastructures, Internet of Things, and data-driven networking.

Requirements, Design Challenges, and Review of Routing and MAC Protocols for CR-Based Smart Grid Systems

Athar Ali Khan, Mubashir Husain Rehmani, and Martin Reisslein

The authors evaluate the requirements and key design challenges for routing and MAC protocols in the CR-based smart grid. The authors also provide a review of research carried out to date for routing and MAC protocols for the CR-based smart grid.

ABSTRACT

Cognitive radio technology can facilitate communication in smart grid applications through dynamic spectrum access. However, traditional routing and MAC protocols adopted for cognitive radio networks may not be beneficial in CR-based smart grid environments due to large data sizes and variable link quality among different functional blocks of smart grids. The interference and fading in wireless links necessitate efficient routing for reliable low-latency data delivery of smart grid applications. This low-latency data delivery must be achieved while protecting the legitimate primary users. Besides efficient routing, MAC layer protocols should be enhanced to achieve successful data delivery with simultaneous spectrum sensing and duty cycling for energy-efficient operation. In this article, we evaluate the requirements and key design challenges for routing and MAC protocols in the CR-based smart grid. We also provide a review of research carried out to date for routing and MAC protocols for the CR-based smart grid.

INTRODUCTION

Smart grids (SGs) are envisioned as future power grids to enhance the functionality of traditional power grids. The communication technologies in power grids suffer from connectivity problems due to dynamic topology changes, fading, and interference. A variety of communication technologies have been suggested to overcome these problems, and the cognitive radio network (CRN) is recognized as one promising solution. CRNs employ dynamic spectrum access (DSA) to search for available channels in both licensed and unlicensed bands. Hence, CRNs may not only counter the problems of traditional communication networks, but may also serve as a bidirectional communication paradigm between consumers and utilities in the smart grid [1].

CR communication in SGs must comply with the regulatory constraints for the various communication technologies. A detailed review of these regulatory constraints is outside the scope of this article, which focuses on routing and medium access control (MAC) protocols within the context of cognitive radio (CR)-based SGs. For one example of a communications technology with strict regulatory constraints, we point to IEEE 802.22 wireless regional area network (WRAN) communication, which is being considered for so-called smart utility networks (SUNs) that communicate over TV white space (TVWS) [2]. In the United States, CR communication by unlicensed SUN TV band devices has to comply with Federal Communications Commission (FCC) regulations. FCC regulations require TVWS devices to include a geolocation capability and the capability to access a database of protected radio services. Devices must check the geolocation database before transmission and must recheck the geolocation database periodically [3]. We also note that the growing interest in CR communication in SGs has spurred extensive standardization efforts. We refer the reader to [4] for an overview of these standardization efforts.

Smart grids have a multi-tiered architecture consisting of home area network (HAN), neighborhood area metwork (NAN), and wide area network (WAN), as shown in Fig. 1. The HAN encompasses the communication within a home, which is relayed via a HAN cognitive gateway to the NAN. The NAN interconnects the HANs in a neighborhood area with each other and with a NAN cognitive gateway. The NAN cognitive gateway relays the NAN communication to the WAN, which interconnects the NANs with the power utility facilities and control units. Besides this three-tiered architecture, hybrid architectures, such as the advanced metering infrastructure (AMI), are also present. The SG performance strongly depends on reliable, successful, and timely end-to-end message delivery among these architectural tiers (blocks). Multipath propagation, fading, interference, and noise phenomena vary greatly in both space and time, and so does the resulting link status. Multichannel communication and proper routing solutions can improve network capacity with interference-free transmissions over multiple channels.

TRAFFIC TYPES AND DELIVERY REQUIREMENTS

SG applications generate a wide range of traffic types. The traffic types have diversified requirements for quality of service (QoS), for example, in terms of reliability, delay, and throughput. The SG traffic types have been classified in the literature [4, 5] depending on SG application areas and architectural layers. One example classifica-

Digital Object Identifier: 10.1109/MCOM.2017.1500744

Athar Ali Khan and Mubashir Husain Rehmani are with the COMSATS Institute of Information Technology; Martin Reisslein is with Arizona State University.

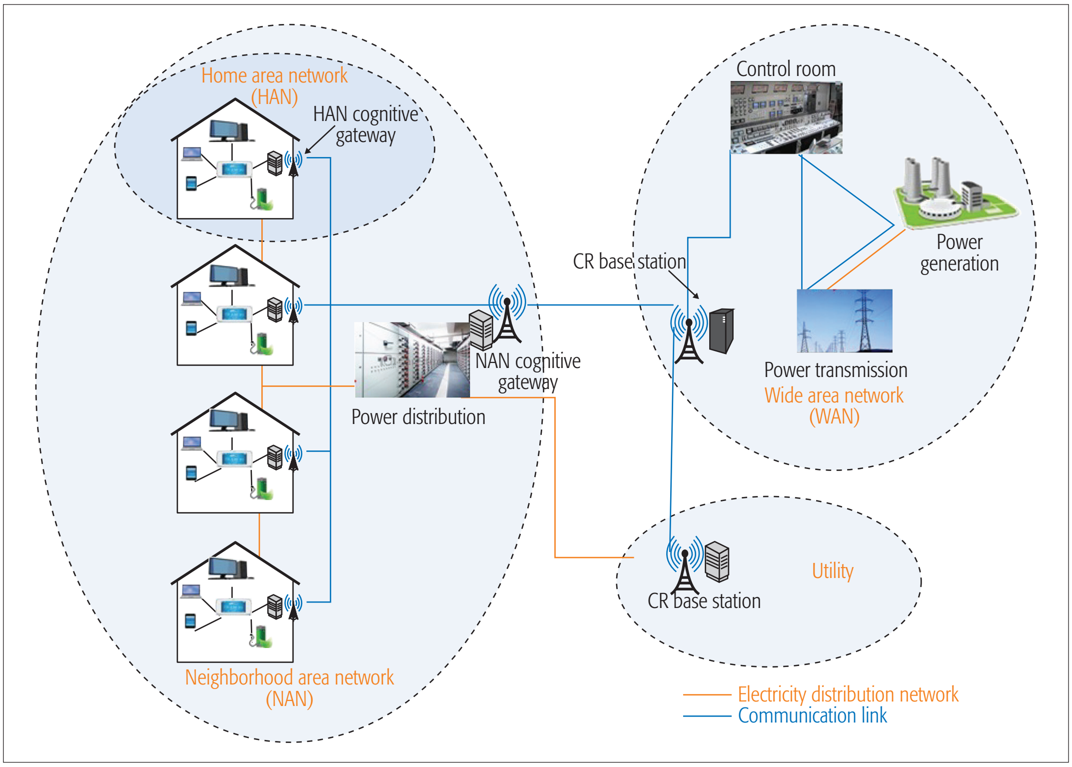


Figure 1. Conceptual CR-based SG architecture: A CRN is incorporated into the architecture to coordinate among three functional blocks (layers) of the SG: the HAN, NAN, and WAN.

CRNs have to search opportunistically for spectrum access in order to successfully transfer information. In particular, the physical layer should provide statistical information related to channel conditions to the upper layers.

tion defines three traffic types: multipoint-to-point (from devices to base station), point-to-multipoint (from base station to devices), and point-to-point (from device to device). Another classification in the literature defines data traffic according to priorities: highest priority vital messages are used for control, protection, and management; second highest priority messages convey system monitoring information; and third highest priority is for meter reading. SG data traffic types with payloads and delay in data delivery define network requirements. In Table 1, we summarize the different SG applications with typical data sizes as well as requirements for bandwidth, reliability, and latency.

Medium Access Control and Routing in Cognitive-Radio-Based Smart Grid

CRNs have to search opportunistically for spectrum access in order to successfully transfer information. In particular, the physical layer should provide statistical information related to channel conditions to the upper layers. The physical layer is responsible for providing opportunities to detect white space or primary user (PU) activity in time and frequency, while the network layer specifies their location dimension [6, 7]. The MAC layer should perform channel management functions, such as spectrum sensing, spectrum decision, spectrum sharing, and spectrum management. All these processes are power consuming; therefore they should be completed within a short time so as to maximize energy-saving sleep phases. In addition, it is very important that the CR MAC processes and transmission do not affect the legitimate PUs [1]. The design of novel MAC protocols for the CR-based SG requires the provisioning of low-overhead spectrum access. Moreover, the trade-off between spectral efficiency and energy efficiency should be balanced through optimal control of the duty cycle. The duty cycle is defined as the ratio of the duration of the listening time to the duration of the full listen-sleep period. CR nodes should be periodically turned off for as long as possible to minimize the duty cycle, thereby avoiding idle listening and reducing energy consumption. An adaptive duty-cycling mechanism may be a good strategy to keep the energy consumption low, in the sense that energy consumption does not necessarily affect the spectral efficiency. An efficient MAC protocol should not only strive for these simultaneous objectives, but also guarantee reliable operation in challenging smart grid wireless environments [8].

Minimum communication delays, which may be achieved through efficient multihop multicast routing algorithms, generally help to meet the objectives of SG applications. Multicast routing is especially promising for applications that need to reach multiple distributed grid components within a prescribed time period [9]. The routing protocols may be on-demand, table-driven, and QoS-aware [9], and should provide high packet delivery ratios, short end-to-end delays, and high energy efficiencies in different SG wireless environments [9]. The routing protocols attempt to deliver data after calculating a routing path with minimum cost. This minimum cost may correspond to either a minimum number of hops or a minimum total edge weight.

The overall satisfactory operation of the CR-based SG depends on the proper optimization of the MAC and routing protocols. These protocols should be designed to achieve necessary objectives while satisfying certain requirements. To date there has only been limited research on

SG application	SG architecture layer	Data size (bytes)	Bandwidth	Latency	Reliability (%)
Home automation	HAN	10–100	–	Seconds	>98
Building automation	HAN	>100	–	Seconds	>98
On-demand meter reading from meters to utility	NAN	100	10–100 kb/s/node, 500 kb/s for backhaul	<15 s	>98
Scheduled meter reading from meters to AMI	NAN	1600–2400	10–100 kb/s/node, 500 kb/s for backhaul	<4 h	>98
Bulk transfer of meter reading from AMI to utility	NAN	MBs (depending on the number of devices)	10–100 kb/s/node, 500 kb/s for backhaul	<1 h	>99.5
Pricing in terms of time of use (TOU) from utility to meters	NAN	100	–	<1 min	>98
Real-time pricing from utility to meters	NAN	100	–	<1 min	>98
Critical peak pricing from utility to meters	NAN	100	–	<1 min	>98
Demand response	NAN, WAN	100	14–100 kb/s per node/device	500 ms to several minutes	>99.5
Service switch operation	NAN	25	–	<1 min	>98
Distribution automation	NAN	25–1000	9.6–100 kb/s	<4 s to <5 s	>99.5
Outage and restoration management	NAN	25	–	<20 s	>98
Electric transportation	NAN	100–255	9.6–56 kb/s, 100 kb/s is a good target	<15 s	>98
Firmware updates	NAN	400–2000 k	–	<2 min to 7 days	>98
Customer information and messaging	NAN	50–200	–	<15 s	>99
Wide area situational awareness	WAN	>52	600–1500 kb/s	20–200 ms	>99.99

Table 1. SG applications with typical data sizes, as well as requirements for bandwidth, reliability, and latency.

such MAC and routing protocols. We summarize in this article these prior research studies in order to highlight the open design challenges. Our present work is different from previous overview articles such as [1, 4]. Reference [1] provided the motivations for employing CR-based communications in the SG. The general survey in [4] gave a broad overview of a wide range of aspects of the CR-based SG ranging from architectures, interference mitigation schemes, spectrum sensing mechanisms, as well as routing and MAC protocols to security, and also power and energy related schemes. In this article, we focus our attention on the routing and MAC protocols for the CR-based SG. More precisely, for both the routing and MAC protocols in the CR-based SG, we present in detail the requirements, comparisons of existing approaches, and challenges. In addition, we present in detail cross-layer protocols covering MAC and routing in the CR-based SG.

Outline of the Article

This article is organized as follows. The section on MAC protocols in the CR-based SG covers their requirements, surveys existing approaches, and outlines open challenges. We then cover the routing in the CR-based SG, detailing requirements, existing approaches, and main open challenges. Finally, we cover the cross-layer protocols presented in the literature. We summarize the existing work and outline possible future trends.

MAC Protocols in the CR-Based Smart Grid

The MAC layer in CRNs employs DSA to search for available spectrum without causing interference to legitimate primary users (PUs). CR users identify spectrum holes, that is, locally available channels that are not used by PUs. Alternatively, CR users can obtain up-to-date information about these spectrum holes from a geolocation database. CRs can change transmission/reception parameters according to spectrum availability on a range of channels [1].

The spectrum sensing process gathers information about the available channels and the presence of PUs. Spectrum sensing can require significant power in CR-based SG communication networks. Therefore, suitable approaches are necessary, such as minimizing hardware (e.g., using a single radio) and minimizing the sensing durations to an optimum level without compromising sensing accuracy. Besides achieving energy efficiency, the sensing process should account for multipath propagation, fading, and environmental noise. Moreover, the sensing process should avoid false alarms (i.e., should not detect PUs even if there

Requirements	Description
Accuracy in cognitive behavior	Accurate performance of the cognitive cycle (i.e., spectrum sensing, spectrum analysis, and spectrum decision) in terms of spectrum management functionalities is required [10]. Spectrum sensing thresholds should be set carefully, as a low threshold results in higher probability of detection at the expense of higher probability of false alarm. Spectrum sensing may cause harmful interference to PUs, may detect false presence of PUs, or may miss the detection of PUs [1]. Any wrong decision at the sensing stage further complicates other spectrum-related functions. In CR-based systems, PU protection is a very strict requirement, whereas CR communication performance may be compromised. Therefore, a high PU detection probability is a top priority for MAC protocol design, whereas a low false alarm probability is a secondary priority [8]. In [8, 11], the detection probability threshold is set to 0.9 and the probability of false alarm is set to 0.1.
Power optimization	Advanced solutions with power optimization capabilities should be designed. Sensing durations should be minimized to achieve power efficiency; however, sensing accuracy should remain high and data should be continuously delivered within limited time periods [1]. Sensing and data transmissions require power; therefore, energy-efficient mechanisms should be designed. Joint operation of event estimation, spectrum sensing, and channel identification can determine sleep schedules that achieve good CR data delivery performance while reducing energy consumption [6]. Experimental studies [8] found: Power drained in transmit mode = 66.2 mW, power drained in receive mode = 70.7 mW, and power drained in spectrum sensing = 65.8 mW. Large power savings can be achieved through sleep states whose power consumption levels are orders of magnitude lower than the listed power drain values for active states.
Efficiency	MAC protocol designs should consider channel interference and capacity. Data traffic from CRs may be sorted with respect to sensitivity and prioritized according to contention window sizes [10, 12]. The data should be forwarded to the node that provides the highest margin for delay budget [8]. During sensing, CR nodes do not forward data packets, thereby compromising network performance in terms of end-to-end throughput, latency, and packet loss ratio. Hence, there is a need to adopt suitable MAC protocols with optimum sensing durations [8].
Contention resolution	If a CR fails to access the channel in a time slot, it may attempt to access the channel in the next time slot. If there are many CRs trying to access the channel, an efficient contention resolution scheme is required; otherwise, packet loss increases. An increase of the timeout (backoff) duration may be a good option, which will increase the chances of channel access among CRs. Another possibility is to use receiver-based MAC protocols with suitable elective schemes in which receivers compete and the winner transmits instead of sender-based MAC protocols [8].
Single radio vs. multiple radio behavior	Single radio or low-complexity processors provide cost-effective solutions [8]. Contrary to this, multiple radios may result in long delays due to channel switching but may produce higher throughput [1]. The use of multiple radios may degrade the MAC protocol performance in case of dense CR node deployments [6].
Scalability	The SG consists of numerous users with a multitude of applications, utilizing many communication techniques, giving rise to scalability challenges. Scalability issues are often related to installation and maintenance costs of communication infrastructure. Wireless techniques reduce cost, but have limited coverage areas. To cover large areas, routers and access points have to be added, which typically increases the overall cost. The long range of IEEE 802.22 WRANs provides large coverage areas; however, availability of spectrum bands is an issue when moving between HANs, NANs, and WANs. A scalable SG system requires seamless connectivity with mobility and QoS support [1].

Table 2. Design requirements for MAC protocols in CR-based smart grid.

are none). False alarms can result in low spectrum utilization [1, 6].

In this section, we discuss MAC protocols for CR-based SG systems. We first discuss design requirements for MAC protocols in CR-based SG. We then analyze existing MAC approaches in the CR-based SG and present potential challenges for MAC protocols.

Requirements of MAC Protocols in CR-Based SG

Design requirements for an efficient MAC protocol in CR-based SG are summarized in Table 2.

Existing Approaches of MAC in CR-Based SG

In this section, we give an overview of the two main existing proposals, cognitive receiver-based MAC (CRB-MAC) and packet reservation multiple access (PRMA). Notable features of the existing approaches are summarized in Table 3. Please ignore the columns for carrier sense multiple access with collision avoidance (CSMA/CA) with distributed control algorithm (DCA) and suboptimal DCA for now; these two cross-layer approaches are covered later.

CRB-MAC: In the CRB-MAC protocol, nodes employ an optimal transmission time by starting a timer. The timer setting is subject to an interference constraint to improve overall network performance and to ensure PU protection [8]. Nodes can have a relatively short sensing time if they are in a region of low PU activity, thereby experiencing a low number of channel changes over time. CRB-MAC mitigates the performance degradation due to spectrum sensing by reducing the spectrum sensing time. To achieve this, the sensing time is initially set to a maximum value for a prescribed missed detection probability. Then the PU activity is followed, and based on the PU activity information, the sensing time may be decreased over time. In case of successive missed detection events, the node increases the sensing time [8].

CRB-MAC achieves energy efficiency and reliability through preamble sampling and opportunistic forwarding techniques. With preamble sampling, which is also referred to as asynchronous low-power listening, each node selects its sleep/wakeup schedule independent of the other nodes. Nodes sleep most of the time and sense the channels only briefly once during a so-called checking interval. Sending nodes prepend data packet transmissions with a preamble that has the same length as the checking interval to ensure that all receiving nodes sense the preamble and receive the data packet. Besides supporting sleep/wakeup modes without synchronization overheads for individual nodes, preamble sampling avoids idle listening, that is, time periods when secondary users (SUs) with CRs only listen to the channels and do not transmit data. Opportunistic forwarding benefits from the broadcast nature of wireless transmissions and employs multiple receivers. With opportunistic forwarding, a sender node transmits its data to all the neighbors in its communication range, without defining a particular node as a receiver. The transceiver can be tuned to any channel with multiple PU transmitters that have known locations and known maximum coverage ranges [8].

PRMA: In cognitive machine-to-machine (M2M) communications for the smart grid, a centralized MAC protocol may be tailored utilizing PRMA [11]. PRMA is a combination of slotted ALOHA, TDMA, and a reservation scheme. The protocol has a master-slave operation. Specifically, a powerful central network controller senses the spectrum for machine-type devices without spectrum sensing capabilities. This design aims to achieve low cost, low complexity, and low

Reference	CRB-MAC [8]	PRMA-based cognitive MAC [11]	CSMA/CA with DCA [10]	Suboptimal DCA [12]
Network type	CRSN	Cognitive M2M	CRSN	CRSN
Simulator used	Matlab	Matlab	NS-2	NS-2
PU transmitter stationary	Yes	–	–	–
Nodes stationary	Yes	–	Yes	–
Licensed/unlicensed operation	Licensed	Both	–	–
Single/multi-channel	Multi	Multi	Multi	Multi
Single/multihop	Multi	–	Multi	Multi
Energy consumption considered	Yes	Yes	No	No
Advantages	Few retransmissions; low delay; low energy consumption in good channels; high reliability; high packet delivery ratio	Diverse QoS support; low data rate optimization; periodic traffic patterns optimization; good scalability; low overhead dynamic spectrum access	Application-specific QoS requirements; low delay; good reliability; good throughput	QoS support; data scheduling; on-demand routing; low delay; good reliability; good throughput
Disadvantages	High energy consumption in poor channel conditions	Low throughput for low device density	Poor delay performance with increasing number of channels	No performance improvement with increasing number of channels; high contention on common control channel
Other features	Two state independent and identically distributed (i.i.d.) random process activity model; receiver-based; detection probability; low false alarm probability	Optimal reservation cycle; optimized throughput; two state i.i.d. random process traffic model; two state Markov chain for power demand	Joint optimization of routing; MAC and physical layer functions	Use of DSA to mitigate channel impairments; define multi-attribute priority classes; design distributed control algorithm for data delivery that maximizes network utility under QoS constraints
Performance metrics	Probability of channel switching; energy consumption; delay; reliability	Channel switching probs.; backoff; average access delay; throughput; duty cycle; interference ratio	Average delay; throughput; reliability	Reliability; packet latency; data rate
Objective	MAC protocol design for CRSNs in smart grid	Optimal frame structure for PU protection as well as high throughput and energy efficiency	Maximize weighted service of traffic flows belonging to different classes	Maximize weighted service of traffic flows belonging to different classes
Support for delay-sensitive apps.	Yes	Yes	Yes	Yes
Spectrum sensing technique	Energy	Energy	–	–
Architecture	HAN, NAN, WAN	HAN, NAN, WAN	HAN, NAN	HAN, NAN, WAN, AMI
Focused parameter	Energy	Energy	QoS	QoS
Number of PUs	4	–	6	4
PU (transmitter) activity model	Two state i.i.d. random process	Two state ii.i.d. random process	–	–

Table 3. Attributes of existing MAC protocols for CR-based smart grid; the DCA-based cross-layer approaches are reviewed in the section "Cross-Layer Protocols in CR-Based SG."

energy consumption. The underlying available cognitive channel has a number of fixed length time slots, each able to carry a single packet. A frame is formed by grouping a fixed number of time slots, in turn a fixed number of frames constitute a multi-frame. Uplink (UL) and downlink (DL) operate with time-division duplex on the same carrier. For high traffic levels in the UL, the ratio of DL to UL time slots is kept small, whereby only few time slots are reserved for DL communication and acknowledgments (ACKs). The DL time slots also carry broadcast status updates of UL time slots.

CHALLENGES OF MAC PROTOCOLS IN CR-BASED SG

The design requirements in Table 2 imply challenges faced by MAC protocols in the CR-based smart grid. However, there are additional challenges for efficient MAC protocols, which we outline next:

Dynamic Operation in Licensed and Unlicensed Bands: CRNs operate on both licensed and unlicensed bands. Channel switching probabilities in MAC protocols should be optimized to operate on these two bands. Moreover, channel switching to licensed bands may be affected by PUs, and operation in unlicensed bands may suffer from significant interference [8, 11]. In the existing studies, CRB-MAC operates in the licensed band, while PRMA can operate in both licensed and unlicensed bands.

Channel Access Delay: Average access delay (i.e., the average time a device has to wait before obtaining a reservation for the channel) is an important performance indicator of MAC protocols. Average access delay strongly depends on the presence of PUs, collisions among CRs, and the backoff schemes [11]. This average access delay has been examined in PRMA-based cognitive MAC [11].

Duty Cycle and Control Overhead: The optimization of the duty cycling has been an important challenge in all wireless networks, and MAC protocols in CR-based SGs may suffer from inefficient duty cycling. The availability of suitable duty cycles in energy-deficient devices remains a significant challenge [11]. Balancing the trade-off between energy efficiency and spectrum efficiency requires joint consideration of spectrum sensing and duty cycling. Moreover, MAC protocol reliability and effectiveness come at the cost of control overhead [6]. In dynamic radio environments, control overhead occurs in terms of channel switching overhead (i.e., when transmission changes from one channel to another channel) [11].

Error Correction: The continuously varying channel conditions may prevent packet recovery, impair packet forwarding, and cause congestion. Traditional fixed forward error correction (FEC) may not be sufficient for every channel. Therefore, the design of adaptive FEC schemes or hybrid automatic repeat request (HARQ) mechanisms remains a great challenge [6].

ROUTING IN CR-BASED SMART GRID

Routing layer protocols have to account for the link qualities of the wireless links, in terms of both individual links as well as the entire end-to-end path in dynamically changing environments.

REQUIREMENTS OF ROUTING PROTOCOLS IN CR-BASED SG

The design of an efficient routing protocol for the CR-based SG should address the requirements summarized in Table 4.

EXISTING APPROACHES OF ROUTING IN CR-BASED SG

Notable features of existing routing protocols are summarized in Table 5.

QoS/Energy-Based Approach for HAN: Adaptations of the Routing Protocol for Low power and lossy networks (RPL) for HANs were presented in [7]. RPL maintains network state information using one or more directed acyclic graphs (DAGs) in which all edges are oriented to avoid cycles. Basic RPL has four steps: expected transmission time (ETX) calculation, rank calculation, directed acyclic graph (DAG) formation and maintenance, and destination list (DL) update. The ETX is defined in [7] to correspond to the (physical layer) link quality between two nodes, and is commonly based on the signal strength (signal-to-noise ratio) of the received packet. The rank to each node in the DAG is computed on the basis of an objective function. To construct the DAG, the gateway broadcasts a control message called DAG information object (DIO) containing relevant network information. Each node updates its DL through a device announcement message, that is, source IP and the next hop node ID are recorded until the message reaches the coordinator. The features of this approach are realized through selective routing as battery powered devices do not participate in spectrum sensing.

An 802.15.4 radio in the ZigBee pro stack with high receiver sensitivity is considered to cover licensed bands. The dual-radio architecture is employed to independently update the channel backup list without quiet periods. After joining the network, a node is considered a non-spectrum sensing node if it listens to any DIO from spectrum sensor devices or the coordinator. If it fails to do so, it is a spectrum sensing node [7].

QoS-Based Approaches for HAN, NAN, WAN, and AMI: Solutions to routing problems in CR-enabled AMI networks and CR-enabled M2M networks were examined in [13, 14]. Cognitive and Opportunistic RPL (CORPL) protects PUs and satisfies the utility requirements of CRs, that is, requirements for end-to-end throughput, latency, and packet loss ratio under spectrum sensing. CORPL modifies the basic RPL, but retains its DAG-based approach. CORPL has two important steps: selection of a forwarder set and unique forwarder selection. Each network node selects multiple next-hop neighbors as its forwarder set. With unique forwarder selection, the best receiver of each packet forwards the packet. To select the best receiver, the protocol employs a simple overhearing-based coordination scheme based on acknowledgment (ACK) frames. To select a forwarder set, CORPL utilizes the parent structure of RPL. This structure requires at least one backup parent besides the default parent. To maintain the forwarder set, each node opportunistically selects the next hop neighbors. Nodes are dynamically prioritized by a cost function approach in the forwarder set, whereas a simple overhearing-based coordination scheme performs a unique forwarder selection [13, 14].

The routing class decides the cost function to

The continuously varying channel conditions may prevent packet recovery, impair packet forwarding, and cause congestion. Traditional fixed forward error correction may not be sufficient for every channel. Therefore, the design of adaptive FEC schemes or hybrid automatic repeat request (HARQ) mechanisms remains a great challenge.

Requirements	Description
Reliability	The SG consists of a large number of smart meters and access points interconnected with a mesh network. A reliable routing path should be established to achieve reliable demand response, demand supply, dynamic power pricing, and other benefits [1, 13].
Packet delivery ratio (PDR)	The PDR is affected by certain factors, such as packet size and network load. The PDR gives an indication of the protocol performance for given packet loss ratios, whereby high PDRs indicate a well-performing routing protocol [9, 13].
Multihop and multicast design	In case of link failure or node failure, multihop and multicast designs are suitable options for reliable and secure information transfer. The routing algorithms should be able to select the best forwarding paths and the best neighbors set for high-speed and easy-to-deploy wireless backbone systems [6, 9, 13]. An opportunistic "store-carry-and-forward" scheme may be a good alternative to a basic "store-and-forward" scheme [9].
PU protection	PU protection is the utmost requirement. Routing protocols should select paths for CR nodes with minimum interference to PUs. Optimal transmission times for CRs must be selected for PU protection [7, 13].
Throughput	Throughput is the measure of the average rate (bits per second) of payload data delivered to the ultimate destination over a long time horizon. For continuously sending data sources, a high throughput multicast tree may help ensure that all receivers receive all the data.
Quality of service	The latency requirements of certain normal SG applications, such as demand response management, range from about 500 ms to several minutes. Real-time applications (e.g., wide area situational awareness) may need high bandwidth resources (600–1500 kb/s) and low latency (20–200 ms). Low delay can be achieved with complex and updated infrastructures, thus requiring expensive deployment of communication networks, while larger delay may jeopardize SG system stability and reliability. Efficient QoS-specific routing algorithms are needed to support the normal and time-critical operations of the smart devices without compromising other metrics. Routing protocols should also use multicast trees with high transmission rates and small hop counts. Asymmetry in networks due to varying node behaviors should not violate QoS provisioning [7]
Energy consumption	Energy consumption in the network is the averaged and aggregated value of the energy consumed at each node. Cooperative routing protocols, such as diffusion-based cooperative routing protocols, can be developed to increase the energy efficiency of packet forwarding [6, 9].
Path determination	The most commonly used metric to determine the route from the source to a destination in multihop wireless networks is the hop count. However, the hop count cannot reflect the varying link quality in SGs. Hence, traditional multihop routing protocols of mobile ad hoc networks (MANETs) cannot perform satisfactorily by utilizing only the hop count routing metric; additionally, the link loss ratio and the interference among links of a path should be considered in SGs [9]. Routing protocols should also consider the topology changes due to joining/leaving nodes [7].

Table 4. Design requirements for routing protocols in CR-based smart grid.

prioritize the nodes in the forwarder set. There are two different routing classes in CORPL: class A supports PU receiver protection, whereas class B satisfies the end-to-end latency for high-priority delay-sensitive alarms. The routes selected for the secondary network should pass through regions of minimum coverage overlap with the PU transmission coverage. This reduces interference to PU receivers. Class B supports delay-sensitive alarms by selecting the next-hop that ensures the deadline. The packet will be dropped if the deadline has elapsed [13, 14].

Challenges of Routing Protocols in CR-Based SG

Smart grid communication faces a number of challenges, including secure connectivity among different parts of the system. Connectivity may be lost as nodes drop out either permanently or temporarily. Disasters and security breaches are other main causes of connectivity losses. Dynamic environmental conditions, such as multipath, fading, noise, attenuation, and varied channel availability, which may be exacerbated near high-power electrical grid installations, result in unstable and inconsistent link connectivity in wireless SG environments. These connectivity problems also arise in CR-based SG systems, making routing challenging [4]. This section covers the challenges faced by routing protocols for the CR-based SG.

Latency: Latency represents the end-to-end delay components from source to destination [9]. In addition to the transmission delays, a packet may experience processing delay and queuing delay in the nodes, and then propagation delay on the medium. In CR-based SG systems, queuing delay may arise due to waiting for channel access. Although smart meters and smart devices are generally static in the smart grid, the wireless links among these devices and meters may be unstable due to interference and fading effects. For large numbers of smart devices and dynamic link quality, long data delivery delays may be unavoidable. If long delays occur, SG applications that strongly rely on fast data-related actions, such as demand response, dynamic pricing, and wide area situational awareness, may act incorrectly.

Complexity: Routing protocols often suffer from dynamic wireless link changes. Successful operation of CR-based SG networks cannot be achieved with traditional simple routing protocols alone; additional support for channel awareness and interference is needed, increasing complexity and costs [7].

Operation under Spectrum Sensing State: Nodes periodically enter the spectrum sensing state to monitor the channel for PU activity. Nodes in the spectrum sensing state do not forward data packets. The operation of routing algorithms becomes challenging when nodes are in the spectrum sensing state. The spectrum sensing state may degrade the routing protocol performance, especially in SG systems with many nodes or large geographic areas [6, 13]. With large node numbers, many SUs may attempt to simultaneously access the spectrum, whereby each SU may have to establish a reliable link. In addition, SGs covering large geographic areas may frequently experience widely varying wireless channel conditions.

Trade-off between PU and CR Operation: The routing layer needs to protect PUs, and should provide QoS for CRs. The trade-off between these

Reference	[6]	CORPL [13]	CORPL [14]	RPL modifications [7]	CSMA/CA with DCA [10]	Suboptimal DCA [12]
Network type	CRSN	CRN	Cognitive M2M	CRSN	CRSN	CRSN
Simulator type	–	Matlab	Matlab	Matlab	NS-2	NS-2
PU transm. stationary	–	Yes	Yes	–	–	–
Nodes stationary	–	Yes	Yes	–	Yes	–
Licensed/unlicensed operation	–	Licensed	Licensed	Both	–	–
Single/multi channel	Multi	Multi	Multi	Multi	Multi	Multi
Single/multi-hop	Multi	Multi	Multi	Multi	Multi	Multi
Energy consumption considered	No	No	No	Yes	No	No
Advantages	Reliability support	Good PDR; min. PU collisions; improved performance in spectrum sensing state	Good PDR; min. PU collisions; improved performance in spectrum sensing state	Minimum network traffic through channel load balancing; optimization of the protocol stack	Application-specific QoS requirements; low delay; good reliability; good throughput	QoS support; data scheduling; on-demand routing; low delay; good reliability; good throughput
Disadvantages	No performance evaluation besides reliability	High DAG convergence time for low node density; duplicate packet forwarding	High DAG convergence time for low node density; duplicate packet forwarding	Complexity and cost involved in designing CR hardware and software; inter personal area network interference	Poor delay performance with increasing number of channels	No performance improvement with increasing number of channels; high contention on common control channel
Other features	Protocol design principles; study of applications areas and energy harvesting techniques	Two state i.i.d. random process activity model; opportunistic forwarding; minimum harmful interference to PUs	Two state i.i.d. random process activity model; opportunistic forwarding; minimum harmful interference to PUs	Adaptation of RPL to asymmetric networks	Joint optimization of routing; MAC and physical layer functions	DSA mitigates channel impairm.; multi-attribute priority classes; distr. control alg. for data delivery to max. network utility under QoS constraints
Performance metrics	Reliability	Reliability; delay, collision risk	Reliability; delay; collision risk	Total effective links; number of packets	Avg. delay; throughput; reliability	Reliability; packet latency; data rate
Objective	Study the potential of sensor networks for SG apps.	Enhance RPL for CR enabled AMI networks	Enhance RPL for cognitive M2M networks	Modifications of RPL for user requirements (joining procedure; asymmetry)	Maximize weighted service of traffic flows belonging to different classes	Maximize weighted service of traffic flows belonging to different classes
Support for delay sensitive applications	No	Yes	Yes	No	Yes	Yes
Spectrum sensing technique	–	Energy	Energy	Feature	–	–
Architecture	HAN, NAN, WAN	AMI	HAN, NAN, WAN	HAN	HAN, NAN	HAN, NAN, WAN, AMI
Focused parameter	QoS	QoS	QoS	QoS/energy	QoS	QoS
Number of PUs	–	9	9	–	6	4
PU (transmitter) activity model	–	Two state independent and identically distributed (i.i.d.) random process	Two state independent and identically distributed (i.i.d.) random process	–-	-	
Routing method	–	Multi root DAG	Multi root DAG	Multi root DAG	–	
Routing metric	–	ETX	ETX	ETX	–	Latency, data rate
Route maintenance	–	Proactive	Proactive	Proactive	–	Proactive

Table 5. Attributes of existing routing protocols for CR-based smart grid; the DCA-based cross-layer approaches are reviewed the section "Cross-Layer Protocols in CR-Based SG."

Although link scheduling is traditionally considered in conjunction with the link layer and MAC protocols, the joint consideration of routing and link scheduling can achieve significantly enhanced performance in CR-based SGs. Specifically, routing protocols can incorporate schedules with information about active links in each slot so as to minimize transmission conflicts.

two behaviors remains a significant challenge for routing protocols [13].

Link Scheduling: Although link scheduling is traditionally considered in conjunction with the link layer and MAC protocols, the joint consideration of routing and link scheduling can achieve significantly enhanced performance in CR-based SGs. Specifically, routing protocols can incorporate schedules with information about active links in each slot so as to minimize transmission conflicts. This approach is highly challenging as all schedules may have different delays, even for a single tree, making it difficult to select suitable link schedules.

Dynamic Route Discovery: Routing protocols start route discovery in case of a new event or addition/rejection of nodes. Dynamic applications in HANs and wide area SG monitoring pose challenges for successful and reliable route discovery [10].

Cross-Layer Protocols in CR-Based Smart Grid

In this section, we discuss cross-layer approaches for CR-based SG systems. The MAC and routing protocol aspects of the two main existing cross-layer approaches, which are based on a DCA, are summarized in Tables 3 and 5.

QoS-Based Approaches for HAN and NAN

Besides specifically focusing on QoS MAC protocol designs for the CR-based SG, cross-layer designs may be employed to meet the application-level QoS requirements. The DCA-based approaches in [10, 12] differentiate the traffic with heterogeneous QoS requirements into a set of priority service classes with different data rate, latency, and reliability levels. The weighted network utility maximization (WNUM) scheme is used to maximize the weighted sum of the flow service. A cross-layer heuristic solution is then employed to solve the utility optimization problem. The routing protocol interacts with the MAC and physical layers to select a suitable channel, and prioritizes transmissions by setting the MAC contention window size according to the class priority [10].

A CSMA/CA-based MAC protocol with DSA functionality is studied in [10]. An on-demand distributed routing protocol with lower route update frequency compared to ad hoc on-demand routing protocols is used to select a channel with sufficient capacity and constrained bit error rate. The operations of the routing, MAC, and physical layers are controlled through a routing frame period structure. Time is divided into routing frames to deliver data, whereby each routing frame has three periods: a fixed spectrum sensing period as well as variable control and data transmission periods. When the spectrum sensing period initializes, the routing agent switches the control to the physical layer and triggers spectrum sensing. After the sensing period, nodes switch to the common control channel in the control period and transmit their periodic hello beacons, contention frames, and broadcast messages. Route discovery is performed in the case of a new event or a new flow. Route discovery starts by broadcasting a route request message on the control channel during the control period of the routing frame [10].

QoS-Based Approach for HAN, NAN, WAN, and AMI

The cross-layer framework may be formulated as a Lyapunov drift optimization and a suboptimal DCA to support channel control, flow control, scheduling, and routing decisions [12]. The DCA selects the channel dynamically on the basis of the perceived signal interference and the resulting estimated channel capacity. It has been observed that the flows of higher priority classes are largely unaffected by an increase in the number of lower priority flows. At the same time, each class does not cause performance changes under limited number of channels with large PU footprints. In this strategy, the routing algorithm provides channel control by selecting the forwarding node and minimizes saturation to avoid interference. In addition, the routing algorithm initially attempts to provide a shortest routing path from source to destination, and utilizes load balancing to avoid congestion if there are many feasible paths.

Concluding Remarks and Future Trends

From the comprehensive review of existing approaches for MAC and routing protocols, we conclude that the design of routing and MAC protocols for CR-based SG networks is largely an unexplored area. Existing research has developed some initial novel MAC protocol designs for the CR-based SG, while routing protocol research has been limited to modifications of RPL.

More specifically, the novel CRB-MAC protocol has many attractive features but has high energy consumption in poor channel conditions since CRB-MAC requires reception by several next-hop receivers. Increasing the number of receivers to ensure a sufficiently large next-hop receiver set may not be a viable cost-effective solution. With the PRMA-based MAC approach, a device itself does not have spectrum sensing capabilities. Instead, a gateway senses the spectrum. This approach can reduce energy consumption, but delay-sensitive applications may suffer.

With timely reliable channel access and data transmissions, all SG applications can contribute to smooth SG operation. Delay-sensitive applications should receive priority; however, a fairness scheme should be maintained among all applications. Fairness can be ensured through backoff mechanisms in MAC protocols that are designed to support QoS requirements. Critical SG applications can prioritize channel access through a short backoff. However, sudden changes in dynamic CR-based SGs can trigger emergency alarms. If the applications with these alarms have low priority, they have to wait for a long backoff interval [8, 11]. A geolocation database, similar to the database used in IEEE 802.11af, could store the spectrum usage characteristics with number of channels and durations, and may be accessed on demand by SG applications. The database can shorten channel access delays and may be a good option for dynamic operation in licensed and unlicensed bands.

Existing routing studies on the CR-based SG have examined modifications to RPL that add complexity and cost for CR hardware and software. Furthermore, these modified routing protocols have not yet been widely studied in dynamic power systems where interference dominates the channel. As most of the routing protocols studied for CR-based SG systems so far have been based on modifications of RPL, a future research direction is to examine alternate routing strategies, such as the LOADng routing protocol, which is also designed for low-power, lossy networks [15]. Support for more general traffic patterns, flexible packet formats, and avoidance of control packet fragmentation make LOADng a promising alternate to RPL. Alternatively, for reliable data transfer in lossy channel conditions, sending multiple copies of the message concurrently over independent paths may be a good strategy. Retransmissions may be an alternative to account for packets that are dropped due to lost link connectivity. An important overarching future research direction is to develop and validate a common performance evaluation framework for MAC and routing protocols in CR-based SG systems.

Retransmissions may be an alternative to account for packets that are dropped due to lost link connectivity. An important overarching future research direction is to develop and validate a common performance evaluation framework for MAC and routing protocols in CR-based SG systems.

References

[1] V. C. Gungor and D. Sahin, "Cognitive Radio Networks for Smart Grid Applications: A Promising Technology to Overcome Spectrum Inefficiency," *IEEE Vehic. Tech. Mag.*, vol. 7, no. 2, 2012, pp. 41–46.

[2] S. Chin-Sean *et al.*, "Smart Utility Networks in Tv White Space," *IEEE Commun. Mag.*, vol. 49, no. 7, 2011, pp. 132–39.

[3] A. Yang, "Overview of FCC's New Rules for TV White Space Devices and Database Updates," *Proc. ITU Wksp.: Spectrum Management Issues on the Use of White Spaces by Cognitive Radio Systems*, 2014, http://www.itu.int/go/ITU-R/RWP1B-SMWSCRS-14, Accessed on Feb. 8, 2017.

[4] A. A. Khan, M. H. Rehmani, and M. Reisslein, "Cognitive Radio for Smart Grids: Survey of Architectures, Spectrum Sensing Mechanisms, and Networking Protocols," *IEEE Commun. Surveys & Tutorials*, vol. 18, no. 1, 2016, pp. 860–98.

[5] "Communication Requirements of Smart Grid Technologies," Dept. of Energy, tech. rep., Oct. 5 2010, http://energy.gov/gc/downloads/communications-requirements-smart-grid-technologies, accessed on Feb. 8, 2017.

[6] A. O. Bicen, O. B. Akan, and V. C. Gungor, "Spectrum-Aware and Cognitive Sensor Networks for Smart Grid Applications," *IEEE Commun. Mag.*, vol. 50, no. 5, 2012, pp. 158--65.

[7] A. Sreesha, S. Somal, and I.-T. Lu, "Cognitive Radio Based Wireless Sensor Network Architecture for Smart Grid Utility," *Proc. IEEE Long Island Systems, Applications and Tech. Conf.*, May 2011, pp. 1–7.

[8] A. Aijaz *et al.*, "CRB-MAC: A Receiver-Based MAC Protocol for Cognitive Radio Equipped Smart Grid Sensor Networks," *IEEE Sensors J.*, vol. 14, no. 12, 2014, pp. 4325–33.

[9] S. Temel, V. C. Gungor, and T. Kocak, "Routing Protocol Design Guidelines for Smart Grid Environments," *Computer Networks*, vol. 60, 2014, pp. 160–70.

[10] G. A. Shah, V. C. Gungor, and O. B. Akan, "A Cross-Layer Design for QoS Support in Cognitive Radio Sensor Networks for Smart Grid Applications," *Proc. IEEE ICC*, June 2012, pp. 1378–82.

[11] A. Aijaz and H. Aghvami, "PRMA Based Cognitive Machine-to-Machine Communication in Smart Grid Networks," *IEEE Trans. Vehic. Tech.*, vol. 64, no. 8, Aug. 2015, pp. 3608–23.

[12] G. A. Shah, V. C. Gungor, and O. B. Akan, "A Cross-Layer QoS-Aware Communication Framework in Cognitive Radio Sensor Networks for Smart Grid Applications," *IEEE Trans. Industrial Informatics*, vol. 9, no. 3, 2013, pp. 1477–85.

[13] A. Aijaz, H. Su, and H. A. Aghvami, "CORPL: A Routing Protocol for Cognitive Radio Enabled AMI Networks," *IEEE Trans. Smart Grid*, vol. 6, no. 1, July 2014, pp. 1–9.

[14] A. Aijaz, H. Su, and A. H. Aghvami, "Enhancing RPL for Cognitive Radio Enabled Machine-to-Machine Networks," *Proc. IEEE WCNC*, Apr. 2014, pp. 2090–95.

[15] J. Yi, T. Clausen, and Y. Igarashi, "Evaluation of Routing Protocol for Low Power and Lossy Networks: LOADng and RPL," *Proc. IEEE Conf. Wireless Sensors*, Dec. 2013, pp. 19–24.

Biographies

Athar Ali Khan is currently pursuing his Ph.D. degree in electrical engineering at COMSATS Institute of Information Technology, Wah Cantt, Pakistan. He received his M.S. degree in computer engineering from COMSATS Institute of Information Technology, Islamabad, Pakistan, in 2011. He received his B.Sc. in electrical (computer and communication) engineering with honors from CECOS University of IT and Emerging Sciences, Peshawar, Pakistan, in 2001. He is a reviewer of Elsevier's *Computers and Electrical Engineering* journal, the *Journal of Network and Computer Applications*, and *IEEE Access*.

Mubashir Husain Rehmani [M'14, SM'15] received his B.Eng. degree in computer systems engineering from Mehran University of Engineering and Technology, Jamshoro, Pakistan, in 2004, his M.S. degree from the University of Paris XI, France, in 2008, and his Ph.D. degree from the University Pierre and Marie Curie, Paris, in 2011. He is currently an assistant professor at COMSATS Institute of Information Technology. He received the certificate of appreciation, "Exemplary Editor of the IEEE Communications Surveys and Tutorials for the Year 2015" from the IEEE Communications Society.

Martin Reisslein [A'96, S'97, M'98, SM'03, F'14] (reisslein@asu.edu) is a professor in the School of Electrical, Computer, and Energy Engineering at Arizona State University, Tempe. He received his Ph.D. in systems engineering from the University of Pennsylvania in 1998. He currently serves as Associate Editor for *IEEE Transactions on Mobile Computing* and *IEEE Transactions on Education* as well as *Computer Networks* and *Optical Switching and Networking*.

Random Access Protocols for Massive MIMO

Elisabeth de Carvalho, Emil Björnson, Jesper H. Sørensen, Petar Popovski, and Erik G. Larsson

The authors address the need for random access by the devices to pilot sequences used for channel estimation, and shows that Massive MIMO is a main enabler to achieve fast access with high data rates, and delay-tolerant access with different data rate levels.

Abstract

5G wireless networks are expected to support new services with stringent requirements on data rates, latency and reliability. One novel feature is the ability to serve a dense crowd of devices, calling for radically new ways of accessing the network. This is the case in machine-type communications, but also in urban environments and hotspots. In those use cases, the high number of devices and the relatively short channel coherence interval do not allow per-device allocation of orthogonal pilot sequences. This article addresses the need for random access by the devices to pilot sequences used for channel estimation, and shows that Massive MIMO is a main enabler to achieve fast access with high data rates, and delay-tolerant access with different data rate levels. Three pilot access protocols along with data transmission protocols are described, fulfilling different requirements of 5G services.

Introduction

There is a growing consensus (3GPP, METIS, ITU-R) that 5G wireless networks will support three generic services:
- Enhanced Mobile BroadBand (eMBB), with very high data rates as the central feature.
- Massive Machine Type Communication (mMTC), with massive numbers of rather simple machine-type devices.
- Ultra-Reliable Low Latency Communications (URLLC), with very low latency and extremely high robustness.

Within each category there can be specific services with additional requirements, e.g. eMBB services that require low latency are referred to as *Tactile Internet applications*. While the sheer number of devices is central to the definition of mMTC, it also plays a significant role for eMBB in several challenging scenarios where a large crowd of users is served in a limited spatial region, such as a shopping mall, stadium, or open air festival [1]. In addition, macro-cells will remain important to provide coverage over larger spatial regions for mobile eMBB services, where dynamic crowds appear along congested streets [1]. In this article, we introduce the term cMBB (crowd MBB) to denote the distinct class within eMBB related to crowd scenarios, and describe efficient methods for devices to access the network in such scenarios. As described below, a key technique is decentralized assignment of pilot sequences, based on *random access*. Next, we describe the motivation for random access to pilots and the distinctive role played by Massive MIMO (multiple-input multiple-output).

Why Random Access to Pilots?

Channel state information (CSI) is necessary for coherent communication. CSI is particularly important at the base stations (BSs) in crowd scenarios, since legacy scheduling and power control algorithms are insufficient to manage many simultaneous connections. The antenna arrays at the BS are required to manage interference in the spatial domain through spatial multi-user beamforming, where each beam is tailored to the CSI of the corresponding device. The CSI acquisition through pilot sequences is challenging in use cases with large numbers of users, cMBB and mMTC. However, the limitations for CSI acquisition are different for these two services. Assuming a simple protocol with orthogonal pilot sequences, in cMBB the number/duration of the pilot sequences is limited by the channel coherence time due to mobility, while in mMTC the devices are generally quasi-static and the number of pilot sequences is rather limited by the uplink power budget. For mMTC devices in a dynamic environment, mobility may also become a limiting factor. Regardless of the reason, those restrictions put a fundamental limit on the number of pilots that can be shared by the devices, such that *the number of devices is much larger than the number of available orthogonal pilot sequences*.[1] This is the key motivation for devising scalable pilot assignment protocols for cMBB/mMTC.

The wireless traffic specifics are important for choosing the access method for the pilots. In mMTC, each device is sporadically active. In the initial access phase, an active device connects to the BS, identifies itself and establishes a coarse synchronization [2]. In principle, the BS can pre-allocate the pilot sequences to the devices, but this is impractical under intermittent user activity. The same holds for cMBB in non-streaming applications, where short periods of activity alternate with long periods of silence. High device density and intermittent traffic makes it infeasible to have a dedicated pilot allocation per device. Instead, we resort to random access to pilots (RAP), where a device that wishes to communicate selects randomly a pilot sequence from a predefined set. RAP may lead to *pilot collision*, which is essentially *pilot contamination*, and the access protocol should deal with it.

5G Services and Pilot Shortage

In Fig. 1, a schematic representation of the 5G services is provided. The services are positioned as a function of the number of devices (horizontal axis) and the number of pilot sequences available (vertical axis). They are described below.

[1] Any number of non-orthogonal pilot sequences can be generated to give each device a unique pilot sequence. Instead of being interfered by devices using the same pilot, each device will be partially interfered by a much larger set of devices. This interference will be substantial, potentially larger than in the case of orthogonal pilots, thus the use of non-orthogonal pilots does solve the problem. However, it transfers some of the issues from the MAC-layer to PHY-layer interference mitigation, which is not the subject of this work.

Digital Object Identifier: 10.1109/MCOM.2017.1600477CM

Elisabeth de Carvalho, Jesper H. Sørensen, and Petar Popovski are with Aalborg University; Emil Björnson and Erik G. Larsson are with Linköping University.

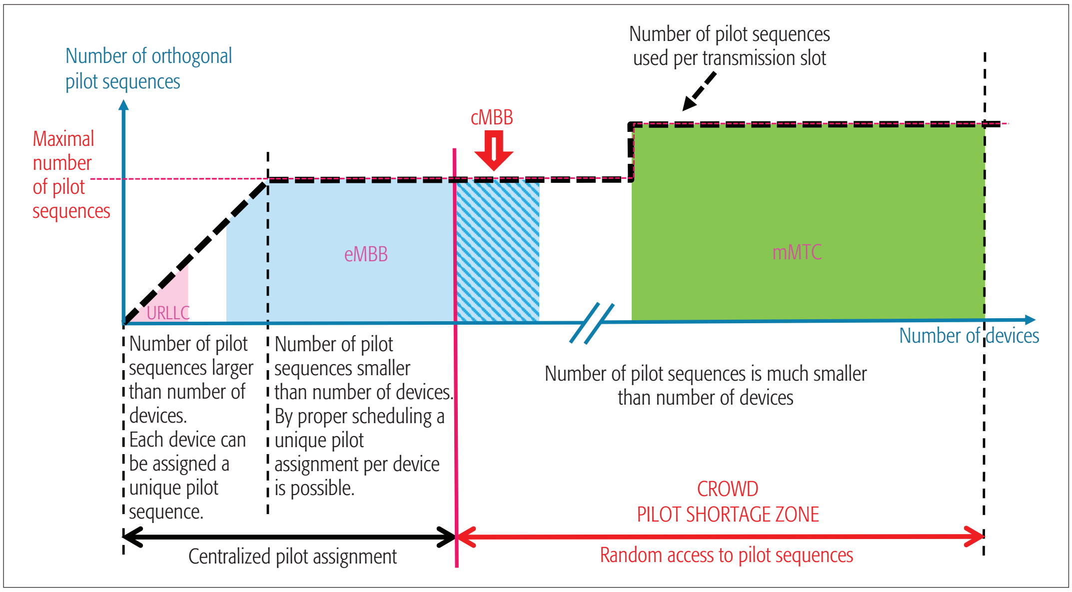


Figure 1. Schematic representation of 5G services as a function of the number of devices (horizontal axis) and pilot sequences (vertical axis). When the number of devices is not much larger than the number of orthogonal pilot sequences, centralized pilot assignment is possible. In a crowd scenario, the number of devices is much larger than the number of orthogonal pilot sequences. In this pilot shortage zone and with intermittent traffic, random access to pilot sequences is a viable solution. In the figure, the maximal number of pilot sequences is arbitrary.

The antenna arrays at the BS are required to manage interference in the spatial domain through spatial multi-user beamforming, where each beam is tailored to the CSI of the corresponding device. The CSI acquisition through pilot sequences is challenging in the use cases with large numbers of users, cMBB and mMTC. However, the limitations for CSI acquisition are different for these two services.

URLLC: Due to high reliability requirements, a small number of pilots should be exclusively allocated for URLLC. The channel hardening due to large array beamforming mitigates the small-scale fading. To harness this effect, the channel estimation should be precise and frequent, such that when critical low-latency data arrives, it can be sent with very high reliability.

eMBB: The central feature is the high data rate. The number of devices is moderate and can be larger than the number of pilot sequences. Coordinated pilot assignment remains possible without a large overhead. An efficient inter-cell pilot reuse plan can suppress pilot contamination.

cMBB: The data rate remains as a central feature, but the number of devices is much larger than the number of pilots. The traffic is intermittent, making the coordinated pilot allocation impractical. An option for pilot reuse is to allocate a small set of orthogonal pilots to each cell and use large reuse distances. The assignment in each cell is then done by RAP.

mMTC: The number of devices is in the order of 10000 or more per BS. Each device is only sporadically active, such that it is necessary to use RAP to connect.

The focus of this article is on the random access mechanisms for scenarios in the pilot shortage zone, as in cMBB and mMTC. There are three major differences between cMBB and mMTC in terms of pilot access:

1. The number of devices, which is orders of magnitude larger in mMTC.
2. The devices in cMBB are located in a small area, while the mMTC are located over a wide area.
3. The characteristics of the downlink traffic.

In cMBB, traffic volume in the downlink dominates over the uplink, while it is the opposite for mMTC. In this article, we assume that the set of pilots allocated to cMBB is orthogonal to the set of pilots dedicated for mMTC. However, how to multiplex pilots across different traffic types is an open question that warrants further research.

Massive MIMO as an Enabler of 5G Services and Random Pilot Access

Massive MIMO, a key ingredient of 5G, can provide very high spectral efficiency (measured in bit/s/Hz/cell) in sub-6 GHz bands [3, 4]. The key idea is to use many antennas at the BSs, which simultaneously serve many devices through spatial multiplexing. Three physical phenomena are important in this regime. First, an *array gain* amplifies the signal by focusing it spatially. Second, *channel hardening* appears, which effectively eliminates the small-scale fading; that is, each device sees an almost deterministic (scalar) channel. Third, the *high spatial resolution* improves the ability to separate devices spatially and facilitate spatial multiplexing to many devices simultaneously. Effectively, each device gets an exclusive, focused data beam and does not suffer from either small-scale fading or interference. The interference between devices that utilize the same pilot is, however, hard to suppress, resulting in pilot contamination [3].

Random access in Massive MIMO benefits from the three aforementioned phenomena [5, 6]. The array gain improves the signal-to-noise ratio (SNR), and hence detects weak devices. Channel hardening facilitates the application of algorithms that exploit asymptotic channel properties. The multiplexing capability offers the possibility of *spatially* resolving collisions, an entirely new opportunity as compared to legacy systems [7].

The introduction of Massive MIMO on the physical layer requires a paradigm shift at the MAC layer. Several essential assumptions that underpin resource allocation algorithms in 3G or 4G become questionable. For example, device

We say that a collision occurs if more than one device transmit the same pilot sequence simultaneously. If multiple devices are involved in a collision, but the channel of one or more of them can be estimated and their identity can be determined, then a capture occurs.

scheduling becomes unnecessary due to the fact that every device in Massive MIMO can utilize the full bandwidth and through a nearly deterministic channel, while being spatially separated from the other devices. This calls for a new MAC design for eMBB/cMBB in order to unleash the full potential of Massive MIMO. For mMTC, a MAC-layer redesign is required since these services pose entirely new problems, which cannot be handled with legacy networks or small-scale MIMO technology.

In the following, we give an overview of the mechanisms to access a limited set of pilots in cMBB and mMTC, including new approaches. Pilot selection is performed at the device, and CSI is acquired at the BS based on the uplink pilots. A time-division duplexing (TDD) system is targeted where channel reciprocity is used to obtain downlink CSI estimates directly from the uplink CSI estimates.

Two Classes of Transmission Methods Based on Random Access

The access methods are described using a terminology that is derived from conventional random access. When a device wants to access the wireless network, it randomly selects an *access sequence*, which is a pilot sequence that can be used for channel estimation, but also to request a pilot sequence for a subsequent collision-free transmission. We say that a *collision* occurs if more than one device transmit the same pilot sequence simultaneously. If multiple devices are involved in a collision, but the channel of one or more of them can be estimated and their identity can be determined, then a *capture* occurs. Specifically, if the capture of a device occurs due to its favorable spatial position with respect to the other devices relative to the BS array, then we refer to it as a *spatial capture*. If the BS or device is capable of detecting a collision, then it can start a *collision resolution* process by explicitly sending messages that govern the future (randomized) action of the accessing UEs.

The schemes described here are based on slotted transmission, where a slot is limited by the channel coherence time and bandwidth. We assume that the initial timing mismatches in the cell are within the cyclic prefix (CP), such that the orthogonal access sequences remain orthogonal when received at the BS.[2] The following two classes of approaches are distinguished.

Random access to pilots (RAP). Here random access is performed for the sole purpose of being granted a pilot sequence that can be used in a collision-free transmission. A special set of non-dedicated access sequences are used and collisions happen in the pilot domain. When a pilot access is collision-free, the corresponding device can be identified, admitted to the network, and assigned a cell-unique pilot. It is henceforth allowed to transmit and receive data, without intra-cell pilot contamination. Collision-free access is enabled by a mechanism that is iterative in general, implying multiple transmission phases between the BS and the devices. RAP finds its primary application in two cases:
- The data size is sufficiently large to justify the overhead of random access.
- The traffic is delay-tolerant (a norm in mMTC), but the data volume per mMTC device should be larger than a threshold (not determined here) to justify the access overhead.

Random access to pilots and data transmission (RAPiD). These are based on random access to the pilot sequences, followed by uplink data transmission. Collisions happen in the pilot domain, while collision-induced interference happens in the data domain. The RAPiD transmission schemes rely on multiple-slot transmission with pilot hopping across the slots. Pilot hopping brings diversity, since the data from a given device will be affected by different contamination events across the transmission slots. A mechanism is needed to identify the transmitting devices. Two transmission schemes are described later. The scheme later is for delay-tolerant applications with uplink dominant traffic, while the scheme later is for delay-stringent applications.

Random Access to Pilots (RAP)

Figure 2a provides a generic depiction of the RAP protocol for three methods described in this section. The first method takes two transmission phases, while the other two take four transmission phases. The transmission phases are as follows:

Phase 1: Uplink access sequence transmission. Each active device selects an access sequence uniformly randomly from a predefined set. In addition, a message allowing collision detection at the BS might be transmitted.

Phase 2: Message broadcast from BS to colliding devices. For the centralized collision resolution method, this message indicates that a collision was detected. For the distributed collision resolution method, a precoded pilot signal is transmitted, enabling distributed collision resolution within the group of colliding devices.

Phase 3: Uplink pilot retransmission. For the centralized resolution method, a new random access sequence is transmitted, possibly from a different set of pilot sequences. For the centralized resolution method, a distributively selected device retransmits the access sequence used in Phase 1.

Phase 4: Network admission message to non-colliding devices and possibly downlink data.

Spatial Capture

When there is a collision in Phase 1, the BS can use the received signal to separate the devices, utilizing the high spatial resolution of massive MIMO. In [8], spatial capture is enabled by timing mismatches between the devices in an OFDM system. The access channel consists of N consecutive subcarriers over which orthogonal access sequences are transmitted in parallel. In the frequency domain, a timing offset gives each device a unique signature that spreads over the N subcarriers of the access channel, similarly to the spatial signature of a transmitting source impinging on a uniform linear antenna array of size N. The sample covariance matrix is computed by averaging over the antennas and used to determine the number of colliding devices and the corresponding timing offsets. These are then used to estimate the respective channels, which the BS uses in Phase 4 to grant access. The estimation only works if the colliding devices have resolvable timing offsets and their number is less than N.

Subspace-based methods based on the covariance matrix of the received signal at the BS can also be employed for spatial capture. In a model

[2] This assumption is valid for a cell radius of 750 m with an LTE-type normal CP, while a cell radius of 2.5 km is supported by the extended CP.

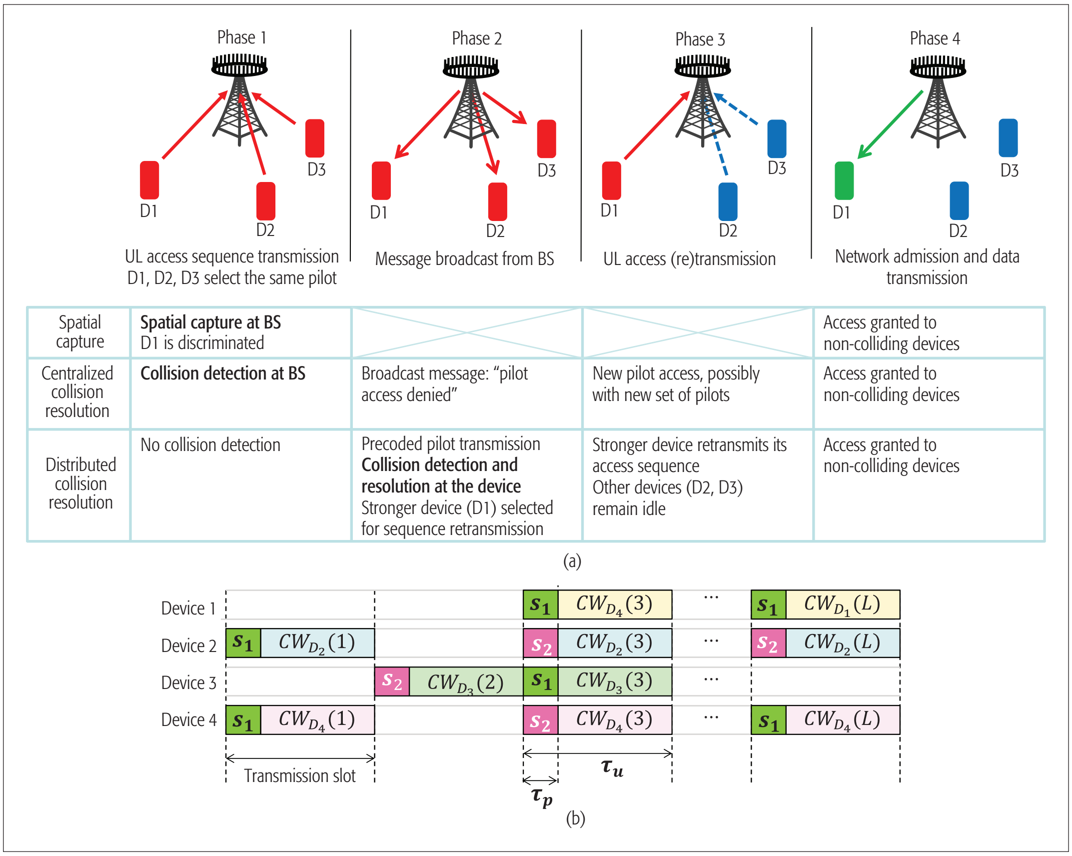


Figure 2. Two classes of random access protocols: a) random access to pilots (RAP); b) random access to pilots and data transmission (RAPiD): simplified example with four devices and two pilot sequences $\{s_1; s_2\}$. $CW_{D_i}(j)$ denotes the portion of the codeword transmitted by device i during transmission slot j.

with pilot transmission, the channels of the colliding devices cannot be distinguished based on the pilot signals. Hence, subspace-based methods assume data mode transmission where the transmitted data from each device is sufficiently different to allow channel separability. As blind methods leave an ambiguity in the multi-device channel estimation, additional training is required. These methods are sensitive to the estimation quality of the sample covariance matrix as well as the number of devices. In [9], the difference of received power between devices in a cell and contaminating devices in neighboring cells is exploited to separate the associated subspaces. This distinction cannot be done in our framework, as the colliding devices belong to the same cell and have comparable received power levels.

Centralized Collision Resolution

Centralized collision resolution assumes the existence of a mechanism at the BS to detect whether a collision occurred or not. The collision detection mechanism in [10] can be adapted to our framework. Pilot access in [10] does not rely on random access but rather on an aggressive pilot reuse plan that causes pilot collisions in a distributed network. Collision detection at the BS is enabled by coded pilot sequences. The coded pilot sequence has two parts:

1. Non-zero symbols used for channel estimation (useful part).
2. A number of l null symbols used for collision detection that are placed at random positions in the sequence.

Each device is identified via a unique on-off pattern. If, at reception, the number of silent symbols is smaller than l, it means that a collision occurred. A new pilot access is then performed in Phase 3 and the probability of a new collision event is reduced. If the number of silent symbols is equal to l, then no collision occurred. At last, note that subspace-based methods from earlier also allow for collision detection.

Distributed Collision Resolution

Here the access collisions can be detected at each device. The BS uses the received signal from Phase 1 for estimating the sum of the channels that the access sequence propagated over. This estimate is used to send a precoded response in Phase 2, which becomes multicasted toward the colliding devices. In case of no collision, one device receives a precoded signal in Phase 2 that is M times stronger compared to the reception when the BS sends in broadcast mode. The device can measure the array gain reliably due to the channel hardening. If, instead, two devices collide, with path-gains β_1 and β_2, they can measure the individual array gains

$$\frac{\beta_1}{\beta_1+\beta_2}M,$$

and

$$\frac{\beta_2}{\beta_1+\beta_2}M,$$

respectively, which sum up to M. The first device can distributively detect the collision by checking if

In a model with pilot transmission, the channels of the colliding users cannot be distinguished based on the pilot signals. Hence, subspace-based methods assume data mode transmission where the transmitted data from each device is sufficiently different to allow channel separability.

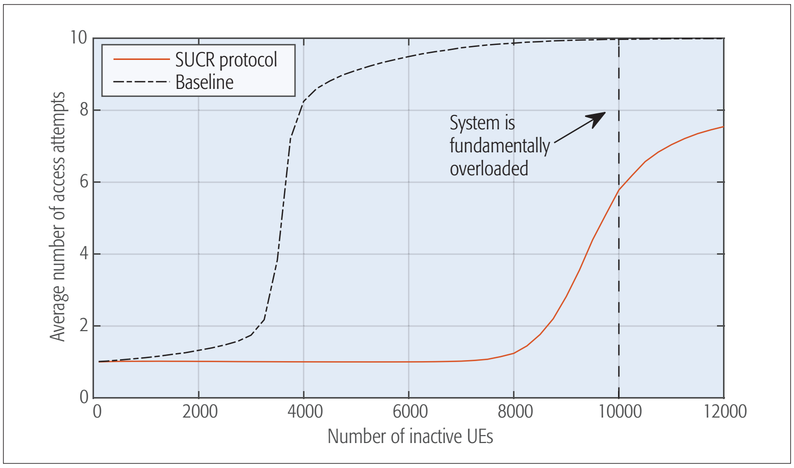


Figure 3. In the method "distributed collision resolution," access collisions can be resolved in a distributed manner, even under very high load.

$$\frac{\beta_1}{\beta_1+\beta_2}M < M$$

and similar for the second device. This procedure can be applied with any number of colliding devices.

The estimated individual array gains can be used for distributed collision resolution, by setting a criterion on when a device may repeat its pilot in Phase 3. For example, the ratio

$$\frac{\beta_1}{\beta_1+\beta_2}$$

informs Device 1 how strong its path-gain is compared to the contenders. The strongest-user collision resolution (SUCRe) decision criterion was proposed in [11], where only the device with the strongest path-gain repeats the pilot in Phase 3. In a two-user collision, Device 1 can be sure to be strongest if

$$\frac{\beta_1}{\beta_1+\beta_2} > 0.5.$$

The SUCRe protocol exploits the natural variations in path-gains that occur due to different propagation distances and shadowing, in contrast to LTE, which attempts to mitigate these variations by power control.

Numerical results have shown that the SUCRe protocol can distributively resolve 90 percent of the collisions. When there is a collision-free pilot transmission in Phase 3, the BS grants access to the corresponding device in Phase 4. The remaining collisions can be handled by repeating the protocol after a random waiting time and/or applying collision resolution and spatial capture at the BS.

The benefits of this method are illustrated in Fig. 3 for a crowd scenario with $K \in [100, 12000]$ devices, $M = 100$ antennas, and $\tau_p = 10$ access sequences. The devices are uniformly distributed in a hexagonal cell, and each one decides to access the network with 0.1 percent probability, which corresponds to one access per minute if the protocol is repeated every 60 ms. The average cell-edge SNR is 0 dB. If a device is not admitted immediately, it makes a new attempt at the next access occasion with probability 0.5. After 10 failed attempts, the access is considered denied. Both numbers can be optimized for a given scenario.

Figure 3 shows the average number of attempts per device, as a function of K. The SUCRe protocol handles up to $K = 8000$ devices without noticeable delays. Notice that with $K = 10000$ there is on average $K \cdot 0.001/\tau_p = 1$ device per access sequence, meaning that the network is fundamentally overloaded. Nevertheless, the average access delays are small because an astonishing 90 percent of the devices are still admitted to the system. This behavior remains also for $K > 10000$. Figure 3 also shows a baseline protocol where pilot collisions are not resolved at the devices, but by making new attempts after random time instants. This protocol requires many more access attempts and breaks down at around $K = 3000$. With $K = 10000$ only 1.5 percent of the devices are ever admitted. Note that we do not compare the SUCRe protocol with the spatial capture scheme from [9], since they are targeted at two fundamentally different scenarios, i.e. small timing mismatches and large timing mismatches, respectively.

Random Access to Pilots and Data Transmission (RAPiD)

Two schemes pertaining to the RAPiD protocol are described on Fig. 2b, and both rely on multiple-slot transmission and pilot hopping. In the first scheme, codewords are transmitted over multiple transmission slots so that contamination is averaged out. In the second scheme, a given device retransmits the same data content in each transmission slot, according to the activation probabilistic model; this is done until the pilot transmission is collision-free.

E-RAPiD: Averaging the Contamination

The basic idea of this protocol is to randomize the effect of pilot contamination over multiple transmission slots. Figure 2b depicts a simplified example, where τ_u is the duration of a transmission slot. We assume a block fading model, with independent realization in each slot and for each device. The protocol relies on three main features:

Pilot hopping: In each transmission slot, each active device randomly selects one sequence from the set of orthogonal pilot sequences. Note that the CSI needs to be estimated at the BS in each transmission slot.

Ergodic data transmission: For each device, the codeword is divided into multiple parts, sent after the pilot sequence.

Device discrimination: From a single transmission slot it is impossible to discriminate between colliding devices. However, the whole series of pilots selected by a device across all transmission slots provides a unique identifier that is used for decoding.

For an asymptotically large number of transmission slots, one codeword is affected by an asymptotically large number of channel realizations and interference events. Interference includes the interference caused by pilot contamination: for a given device, in each of its active transmission slots, contamination comes from a different random set of devices, so that asymptotically the device is affected by all possible sets. Depending on the type of receiver, interference might also come from devices that use different pilots. The ergodic properties of this transmission process achieve a reliable data rate. The transmission protocol is called Ergodic Random Access to Pilot and Data transmission (E-RAPiD).

The performance of E-RAPiD is characterized by using a lower bound on the uplink sum rate that accounts for the probabilistic device activity [6]. The bound $\mathcal{R}$ depends the total number of BS antennas M and the number of pilot sequences, τ_p: the larger those quantities, the more devices can be multiplexed. $\mathcal{R}$ is also a function of the device activation probability, p_a. To maximize the sum rate, one can optimize p_a and τ_p. Heuristics leads to the following approximate solution:

- One third of the transmission slot is used for pilots.
- The average number of active devices is equal to
$$x\sqrt{M\tau_u},$$
where τ_u is the slot duration and the scalar x depends on the distribution of device channel path-gain. The larger the variance, the larger x and hence active devices.

This solution leads to a scaling of the sum rate as
$$\sqrt{M\tau_u}.$$

Figure 4 shows bound $\mathcal{R}$, optimized w.r.t p_a and τ_p for a scenario with K = 800 devices and as a function of the transmission slot duration τ_u. Maximum ratio combining is applied to the BS. The channel path-gains vary uniformly at random around a fixed value $\overline{\beta}$ with a maximal gap of $0.25\overline{\beta}$. The performance is limited by interference and hence is quite insensitive to the SNR. For τ_u = 300, the average number of active devices is around 60 for M = 100 and around 140 for M = 400, so that the average rate per active device is equal to 0.5 bit/s/Hz for both M = 100 and M = 400.

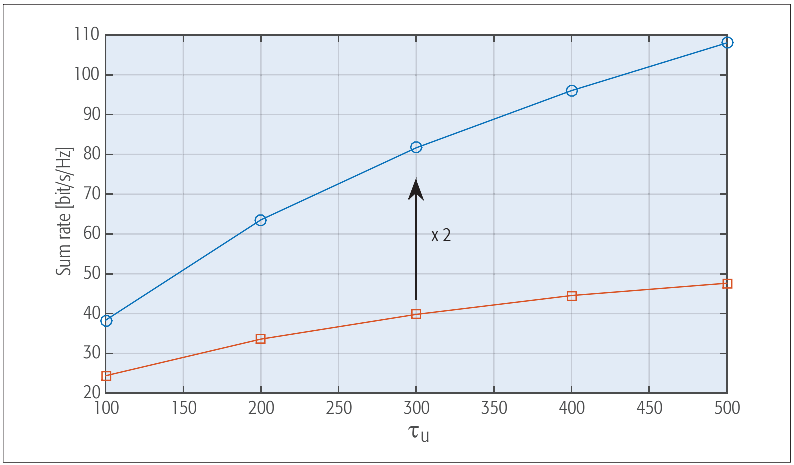


Figure 4. The lower bound $\mathcal{R}$ on the sum rate for K = 800, for M = 100 and M = 400 antennas as a function of the transmission slot duration.

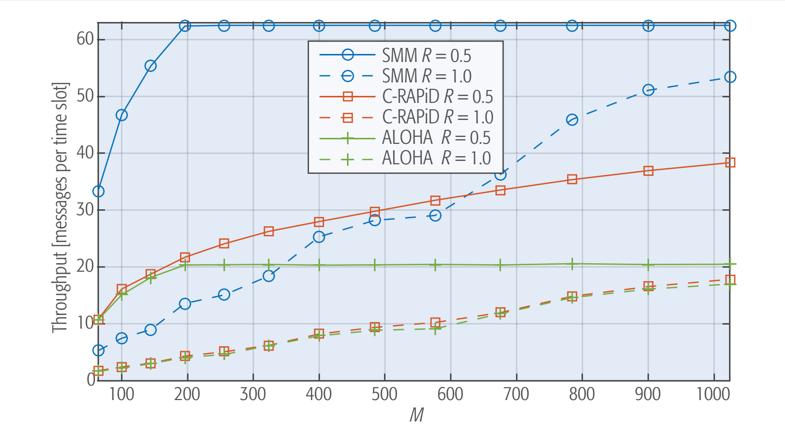


Figure 5. Comparison of throughputs at R = 0.5 and R = 1 and optimized values of τ_p, p_a and the frame size.

C-RAPiD: Intra-cell Interference Cancellation

The third access protocol targets mMTC scenarios, such as the E-RAPiD protocol. It also uses pilot hopping for joint pilot and data transmission. However, instead of spreading the codewords across multiple transmission slots, the codewords are replicated in each transmission slot within a predefined duration, called a frame. The same technique is applied in works on coded random access [12, 13]. The duration of a frame, Δ, is in general lower than ergodic transmissions require, hence this protocol targets more delay-sensitive applications in the mMTC category. The main idea is to use the successful decoding of a data transmission in one slot to cancel the interference brought by replicas in other slots. While sending multiple replicas increases the intra-cell pilot contamination, it also provides multiple opportunities to successfully decode the codeword. This is the key trade-off in the Coded Random Access to Pilot and Data transmission (C-RAPiD) protocol, such that the careful selection of p_a is very important.

The C-RAPiD protocol consists of three steps [14]:

RAPiD access: All K devices participate in a RAPiD access procedure, as described earlier, for Δ transmission slots. In each slot, each device is active with probability p_a. An active device selects randomly one of the τ_p pilots and transmits the pilot and a replica of its data message.

Maximum ratio combining (MRC): The received signals are processed using MRC based on the contaminated estimates achieved from the pilot transmissions. MRC transforms the signals from linear combinations with small-scale fading coefficients to linear combinations with large-scale fading coefficients (the Euclidean norm of the channel coefficients). Note that this transformation is only possible due to the channel hardening and beam decorrelation brought by Massive MIMO.

Successive interference cancellation (SIC): The linear combinations achieved through MRC represent a system of equations, solved through SIC. Initially, the BS locates immediately decodable data, which is practically done through error checking codes. Embedded in the data is the random activity and pilot choices of the device, which allows the BS to locate all the replicas of the same packet. This enables the BS to cancel the interference caused by these replicas. Potentially, this enables the decoding of additional data messages, whereby the iterative SIC procedure continues.

We compare C-RAPiD with two baseline schemes: scheduled Massive MIMO (SMM) and ALOHA. SMM relies on fully scheduled transmissions and thus serves as an upper bound. ALOHA is the classical approach where devices randomly select a pilot sequence and a time slot, and only collision-free transmissions contribute to the throughput.

We apply the channel model from earlier with power control, such that all devices have an SNR of 10 dB. A channel code is applied with rate R and QPSK modulation. For all protocols, τ_p, p_a and Δ have been numerically optimized

Embedded in the data is the random activity and pilot choices of the device, which allows the BS to locate all the replicas of the same packet. This enables the BS to cancel the interference caused by these replicas. Potentially, this enables the decoding of additional data messages, whereby the iterative SIC procedure continues.

for maximum throughput. As expected, the performance of all protocols increases with M (see Fig. 5). However, for $R = 0.5$, the performance of ALOHA saturates at roughly $M = 200$, whereas C-RAPiD continues to increase. ALOHA can only benefit from the increased SINR until the point that collision-free signals are decoded with high probability. C-RAPiD is able to benefit further due to the SIC mechanism. C-RAPiD achieves 45 percent of the throughput of SMM with $M = 400$ and $R = 0.5$, while ALOHA achieves 33 percent. At $M = 1024$, the performance of C-RAPiD is increased to 61 percent.

Conclusions and Perspectives

Massive MIMO is currently one of the most compelling technologies for 5G wireless networks. The operation in TDD and the resulting uplink-downlink reciprocity renders the system entirely scalable with respect to the number of BS antennas, leaving channel coherence (device mobility) as the only remaining, fundamental limiting factor. Fast-moving devices result in short coherence and room for fewer orthogonal pilots in each cell.

In this article, we have addressed crowd-eMBB and mMTC scenarios in which there exist many more devices in a cell than there are unique orthogonal pilots, and where devices periodically or sporadically want to access the network, without prior coordination with the BS. Specifically, we saw how the abundance of spatial degrees of freedom, and the presence of channel hardening, in Massive MIMO, facilitates efficient resolution to resolve colliding transmissions, even in a case where the colliding packets use the same pilots. This led to the central conclusions of the article:

- Massive MIMO is a fundamental enabler for crowd-eMBB scenarios, sensor networks, IoT, and M2M communications.
- The creation of an efficient standard for wireless networks based on Massive MIMO technology will require a complete re-design of the multiple-access layer.
- The spatial domain provides new resources for collision resolution that are unused in legacy systems.

Acknowledgments

This work was performed partly in the framework of the Danish Council for Independent Research (DFF133500273), the Danish National Advanced Technology Foundation via the VIRTUOSO project, the Horizon 2020 project FANTASTIC-5G (ICT-671660), the EU FP7 project MAMMOET (ICT-619086), ELLIIT, and CENIIT. The authors would like to acknowledge the contributions of the colleagues in FANTASTIC-5G and MAMMOET.

References

[1] M. Fallgren *et al.*, Deliverable D1.1: Scenarios, Requirements and KPIs for 5G Mobile and Wireless System, ICT-317669-METIS, 2013, https://www.metis2020.com/.
[2] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*, 1st Ed. Academic Press, 2011.
[3] T. L. Marzetta, "Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, 2010, pp. 3590–600.
[4] T. L. Marzetta *et al.*, *Fundamentals of Massive MIMO*, Cambridge University Press, 2016.
[5] E. Björnson *et al.*, "Random Access Protocol for Massive MIMO: Strongest-User Collision Resolution (SUCR)," *Proc. IEEE ICC*, 2016.
[6] E. de Carvalho *et al.*, "Random Access for Massive MIMO Systems with Intra-Cell Pilot Contamination," *Proc. IEEE ICASSP*, 2016.
[7] 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, 2013, pp. 86–93.
[8] L. Sanguinetti *et al.*, "Random Access in Uplink Massive MIMO Systems: How to Exploit Asynchronicity and Excess Antennas," *IEEE Globecom*, 2016.
[9] R. Muuller, L. Cottatellucci, and M. Vehkaperá, "Blind Pilot Decontamination," *IEEE J. Sel. Topics Signal Proc.*, vol. 8, no. 5, 2014, pp. 773–86.
[10] O. Y. Bursalioglu *et al.*, "RRH based Massive {MIMO} with "On the Fly" Pilot Contamination Control," *Proc. IEEE ICC*, 2016.
[11] E. Björnson *et al.*, "A Random Access Protocol for Pilot Allocation in Crowded Massive MIMO Systems," *IEEE Trans. Wireless Commun.*, 2016, submitted, Available: http://arxiv.org/pdf/1604.04248.
[12] E. Paolini *et al.*, "Coded Random Access: How Coding Theory Helps to Build Random Access Protocols," *IEEE Commun. Mag.*, vol. 53, no. 6, June 2015, pp. 144–50.
[13] G. Liva, "Graph-based Analysis and Optimization of Contention Resolution Diversity Slotted Aloha," *IEEE Trans. Commun.*, vol. 59, no. 2, Feb. 2011, pp. 477–87.
[14] J. H. Sørensen *et al.*, "Coded Pilot Access: A Random Access Solution for Massive MIMO Systems," *IEEE Trans. Wireless Commun.*, 2016, submitted, http://arxiv.org/pdf/1605.05862.

Biographies

Elisabeth de Carvalho (edc@es.aau.dk) received a Ph.D. in electrical engineering from Telecom ParisTech, France. After her Ph.D. she was a post-doctoral fellow at Stanford University, USA, and then worked in industry in the field of DSL and wireless LAN. Since 2005, she has been is an associate professor at Aalborg University, where she has led several research projects in wireless communications. Her main expertise is in signal processing for MIMO communications with a recent focus on massive MIMO, including channel measurements, channel modeling, beamforming and protocol aspects. She is a coauthor of the text book *A Practical Guide to the MIMO Radio Channel*.

Emil Björnson (emil.bjornson@liu.se) received the M.S. degree from Lund University, Sweden, in 2007 and the Ph.D. degree in 2011 from KTH Royal Institute of Technology, Sweden. From 2012 to 2014, he was a joint postdoc at Supelec, France, and KTH. He is currently an associate professor at Linköping University, Sweden. He received the 2014 Outstanding Young Researcher Award from IEEE ComSoc EMEA, the 2016 Best PhD Award from EURASIP, and five best paper awards.

Jesper H. Sörensen [M'10] (jhs@es.aau.dk) received the B.Sc. in electrical engineering in 2007, the M.Sc. (cum laude) with the maximum grade average in 2009, and the Ph.D. in wireless communications in 2012, all from Aalborg University, Denmark. In the second half of 2011 he worked as an intern at Mitsubishi Electric Research Laboratories. From 2012 to 2016 he was a postdoctoral researcher at Aalborg University, where he is currently an assistant professor. His work is in the areas of channel coding, multimedia communication, and protocol design.

Petar Popovski [F] (petarp@es.aau.dk) is a professor at Aalborg University. He received his Dipl.- Ing. (1997)/Magister Ing. (2000) in communication engineering from Sts. Cyril and Methodius University, Skopje, Macedonia, and his Ph.D. from Aalborg University (2004). In 2015 he received an ERC Consolidator Grant and in 2016 he received the IEEE Fred W. Ellersick award. He currently serves as an area editor for *IEEE Transactions on Wireless Communications*. His research interests are in wireless communications/networks and communication theory.

Erik G. Larsson [F] (erik.g.larsson@liu.se) is professor at Linköping University, Sweden. He has co-authored many papers and patents, and two Cambridge University Press textbooks: *Fundamentals of Massive MIMO* (2016) and *Space-Time Block Coding for Wireless Communications* (2003). Recent service includes membership on the IEEE Signal Processing Society (SPS) Awards Board (2017--2019), chairing of the IEEE SPS SPCOM Technical Committee (2015–2016), chairing of the steering committee for *IEEE Wireless Communications Letters (2014–2015)*, and organization of the Asilomar SSC conference (general chair in 2015, technical chair in 2012). He has received the *IEEE Signal Processing Magazine* Best Column Award twice, in 2012 and 2014, and the IEEE ComSoc Stephen O. Rice Prize in Communications Theory in 2015.


ieee.tv
Featured Videos
Videos by Series
Event Showcase
Most Popular
NEW
New Videos
Join IEEE
IEEE
About IEEE
About This App
tv
About IEEE.tv
Member Login
Share
IEEE

ADVERTISING SALES OFFICES

Closing date for space reservation: 15th of the month prior to date of issue

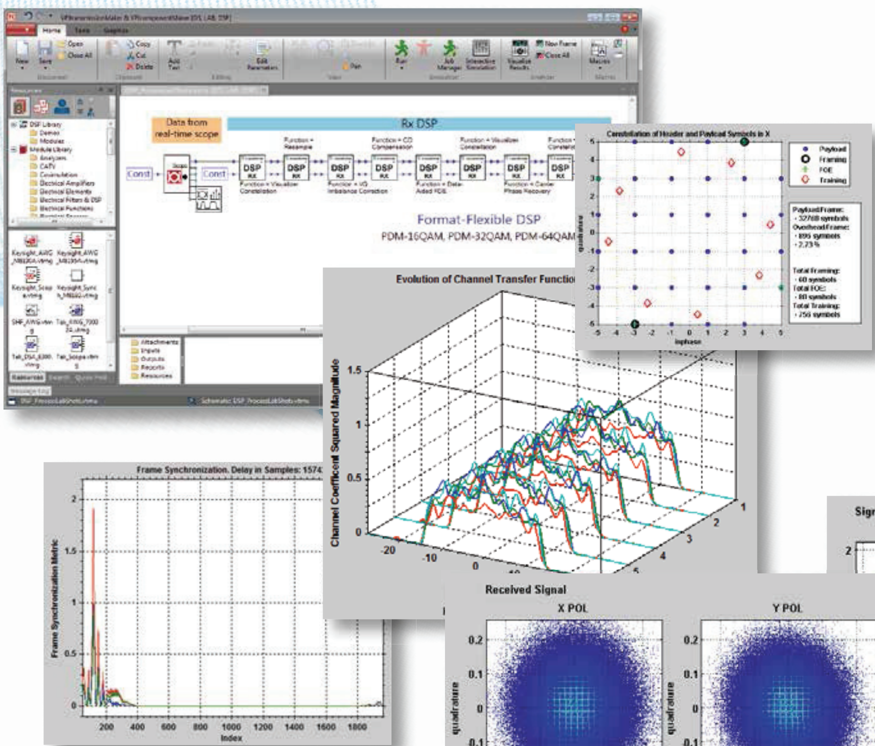
COMPANY .... PAGE

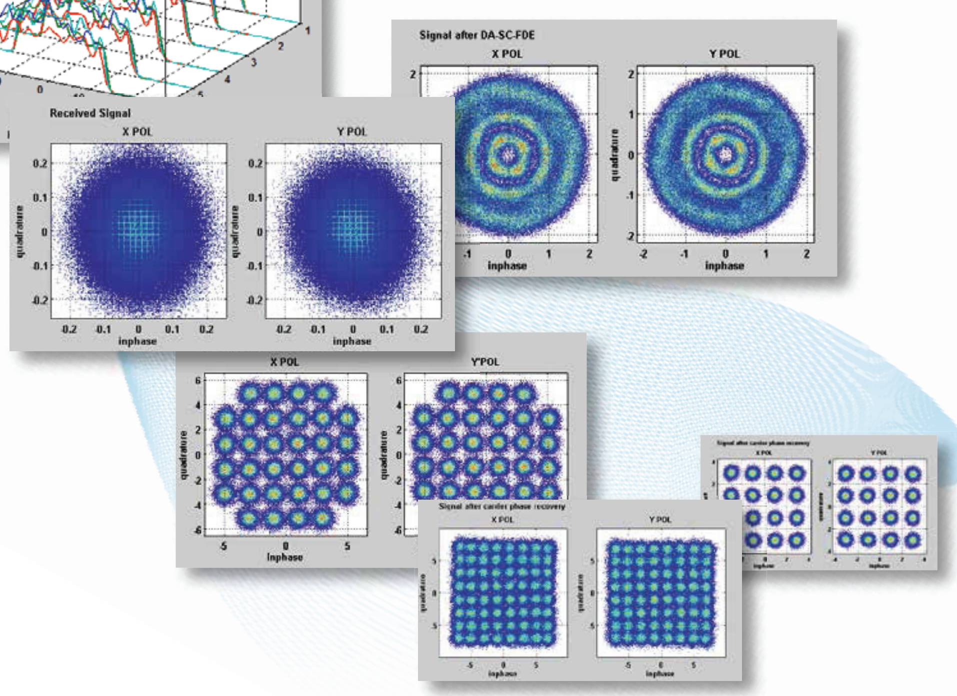
Fraunhofer

Heinrich Hertz Institute

Ready-to-use DSP-Library for optical system simulations and experiments

The DSP-Library for coherent optical systems is available as pluggable toolkit for VPItransmissionMaker™ Optical Systems and VPIlabExpert™. It provides an extensive collection of lab-proven DSP algorithms designed to speed up your development of 400G and 1T applications.





- Benchmark DSP algorithms
- Nonlinear component modeling and pre-distortion filter synthesis
- Design next-generation transceivers
- Address trade-off between DSP complexity and its performance
- Compare modulation formats
- System performance analysis
- Define component requirements

In Cooperation with:

VPIphotonics DESIGN AUTOMATION

Further Information:
http://www.vpiphotonics.com/Tools/DSPLibrary/