May 2015, Vol. 53, No. 5

IEEE Communications Magazine

www.comsoc.org

- *Emerging Applications, Services and Engineering for Cellular Cognitive Systems*
- *Full Duplex Communications*
- *Communications Education and Training*
- *Satellite Communications and Networking*
- *Green Communications and Computing Networks*

IEEE
IEEE COMMUNICATIONS SOCIETY
A Publication of the IEEE Communications Society

May 2015, Vol. 53, No. 5

IEEE Communications MAGAZINE

www.comsoc.org

- *Emerging Applications, Services and Engineering for Cellular Cognitive Systems*
- *Full Duplex Communications*
- *Communications Education and Training*
- *Satellite Communications and Networking*
- *Green Communications and Computing Networks*

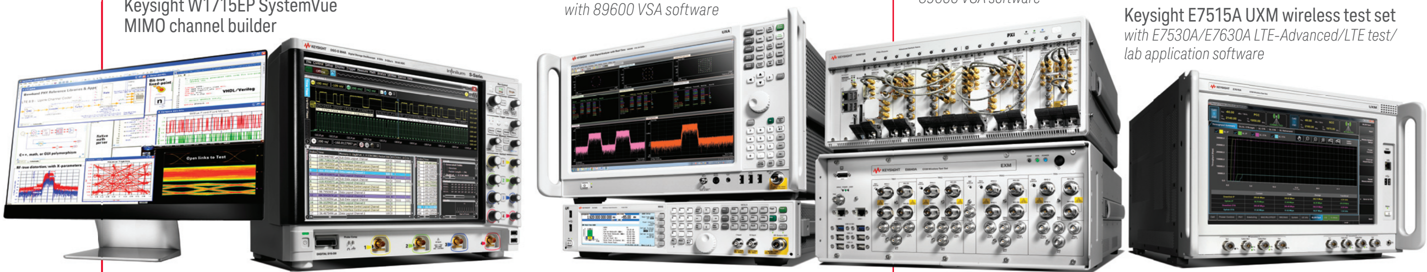
IEEE
IEEE COMMUNICATIONS SOCIETY
A Publication of the IEEE Communications Society

Keysight W1715EP SystemVue MIMO channel builder

Keysight N9040B UXA signal analyzer
with 89600 VSA software

Keysight MIMO PXI test solution
with N7624/25B Signal Studio software for LTE-Advanced/LTE FDD/TDD and 89600 VSA software

Keysight E7515A UXM wireless test set
with E7530A/E7630A LTE-Advanced/LTE test/lab application software



Keysight Infiniium S-Series high-definition oscilloscope
with N8807A MIPI DigRF v4 (M-PHY) protocol decode software

Keysight N5182B MXG X-Series RF vector signal generator
with N7624/25B Signal Studio software for LTE-Advanced/LTE FDD/TDD

Keysight E6640B EXM wireless test set
with V9080/82B LTE FDD/TDD measurement applications and N7624/25B Signal Studio software for LTE-Advanced/LTE FDD/TDD

HARDWARE + SOFTWARE

The more complex your LTE-A design, the more you need help from test and measurement experts. Keysight is the only company that offers benchtop, modular and software solutions for every step of the LTE-A design process. From R&D to manufacturing, we can give you the expertise, instruments and applications you need to succeed.

- *Complete LTE-Advanced design and test lifecycle*
- *Identical software algorithms across platforms*
- *300+ software applications for the entire wireless lifecycle*

PEOPLE

We know what it takes for your designs to meet LTE-A standards. After all, Keysight engineers have played major roles in LTE-A and other wireless standards bodies, including 3GPP. Our engineers even co-authored the first book about LTE-A design and test. We also have hundreds of applications engineers. You'll find them all over the world, and their expertise is yours for the asking.

- *Representation on every key wireless standards organization globally*
- *Hundreds of applications engineers in 100 countries around the world*
- *Thousands of patents issued in Keysight's history*

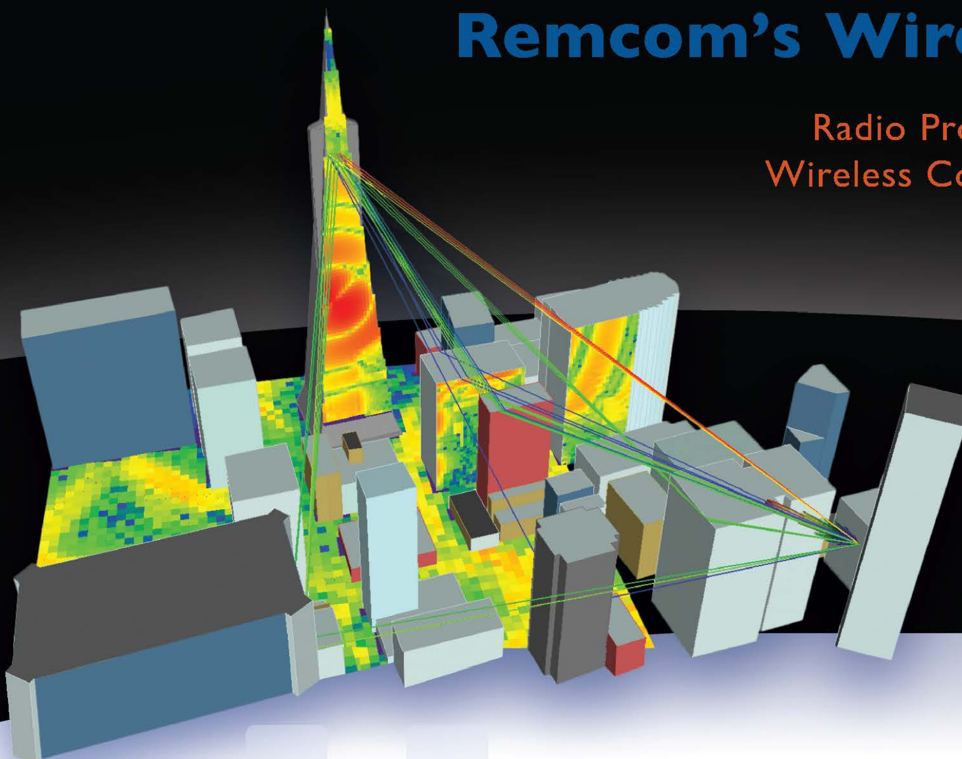
KEYSIGHT TECHNOLOGIES

Unlocking Measurement Insights

IEEE Communications MAGAZINE

MAY 2015, Vol. 53, No. 5

www.comsoc.org/commag



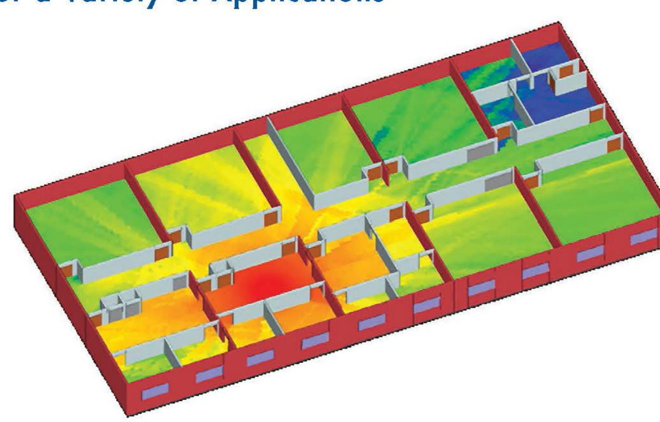
Remcom's Wireless InSite®

Radio Propagation Software for Wireless Communication Planning

Wireless InSite is a suite of ray-tracing models for analyzing EM propagation and communication channel characteristics in complex urban, indoor, rural and mixed path environments.

Wireless EM Propagation Capabilities for a Variety of Applications

- Indoor WiFi
- Moving vehicle or aircraft
- LTE and WiMax throughput analysis
- Tower placement for urban coverage
- Ad-hoc and temporary networks
- Base station coverage analysis
- Microcell coverage



Now integrated with the Geospatial Data Abstraction Library (GDAL).
See all the latest enhancements at **www.remcom.com/wireless-insite-features**

REMCOM®

+1.888.7.REMCOM (US/CAN) | +1.814.861.1299
www.remcom.com

Visit Us at IMS 2015
Booth #1722

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

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

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

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

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

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

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

Satellite Communications and Networking

Guest Editors: Claudio Sacchi, Kul Bhasin, Naoto Kadowaki, and Fred Vong

Communications Education and Training: Industry Certification and University Accreditation

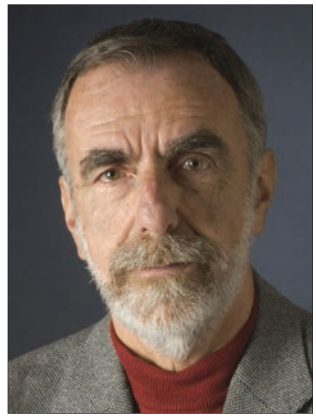
Guest Editors: David G. Michelson, Wen Tong, and Larry Reeves

Green Communications and Computing Networks

Series Editors: Jinsong Wu, John Thompson, Honggang Zhang, and Daniel C. Kilper

EDUCATION AND TRAINING: THE THIRD PILLAR OF COMSOC

This month's President's Page is devoted to the initiatives and plans of the IEEE Communications Society (ComSoc) in the area of education and training. Enhancing the IEEE societies' activities in the area of education and training is one of the strategic goals of IEEE, and it has been recognized as such also by ComSoc's strategic and planning committee (see the January 2015 President's Page). Education and training can be developed as the third pillar of ComSoc's major technical activities and a revenue source to complement the two existing pillars: publications and conferences.



SERGIO BENEDETTO

ComSoc's plans in this area will be illustrated by Khaled ben Letaief, Vice President of ComSoc and Chair of the strategic and planning Sub-Committee on Education Programs, Content and Services, together with Michele Zorzi, ComSoc's Director of Education and Training.

KHALED BEN LETAIEF

MICHELE ZORZI

Dr. Letaief received the BS degree with distinction, and MS and Ph.D. degrees in electrical engineering from Purdue University at West Lafayette, Indiana, USA. From 1990 to 1993 he was a faculty member at the University of Melbourne, Australia. He has been with the Hong Kong University of Science & Technology since 1993, where he has held numerous administrative positions, including the Head of the Electronic and Computer Engineering Department, Director of the Center for Wireless IC Design, Director of the Huawei Innovation Laboratory, and Director of the Hong Kong Telecom Institute of Information Technology.

He is currently Chair Professor and Dean of the HKUST School of Engineering. He is also an internationally recognized leader in wireless communications and networks. He has served as a consultant for different organizations, including Huawei, ASTRI, ZTE, Nortel, PricewaterhouseCoopers, and Motorola. He is the founding Editor-in-Chief of the *IEEE Transactions on Wireless Communications* and has served on the editorial board of other prestigious journals. He has also been involved in organizing a number of flagship international conferences and events.

He is the recipient of many other distinguished awards and honors, including the 2007 IEEE Communications Society Joseph LoCicero Publications Exemplary Award, 2009 IEEE Marconi Prize Award in Wireless Communications, 2010 Purdue University Outstanding Electrical and Computer Engineer Award, 2011 IEEE Communications Society Harold Sobol Award, 2011 IEEE Wireless Communications Technical Committee Recognition Award, and 12 IEEE Best Paper Awards.

Dr. Letaief is recognized as a long time volunteer with dedicated service to professional societies, and in particular IEEE, where he has served in many leadership positions. These include Treasurer of the IEEE Communications Society, Vice-President for Conferences of the IEEE Communications Society, Chair of the IEEE Committee on Wireless Communications, and elected member of the IEEE Product Services and Publications Board.

Dr. Letaief is a Fellow of IEEE and a Fellow of HKIE. He is currently serving as the IEEE Communications Society Vice-President for Technical Activities, member of the IEEE Fellow Evalution Committee, and member of IEEE TAB Periodicals Committee.

Michele Zorzi received his Laurea and Ph.D. degrees in electrical engineering from the University of Padova in 1990 and 1994, respectively. During the academic year 1992–1993 he was on leave at UCSD. After being affiliated with the Dipartimento di Elettronica e Informazione, Politecnico di Milano, Italy, the Center for Wireless Communications at UCSD, and the University of Ferrara, in November 2003 he joined the faculty of the Information Engineering Department of the University of Padova, where he is a professor. His present research interests include performance evaluation in mobile communications systems, random access in mobile radio networks, ad hoc and sensor networks, energy constrained communications protocols, and underwater communications and networking. He was Editor-in-Chief of *IEEE Wireless Communications* from 2003 to 2005, Editor-in-Chief of *IEEE Transactions on Communications* from 2008 to 2011, and Guest Editor for several special issues in *IEEE Personal Communications*, *IEEE Wireless Communications*, *IEEE Network*, and IEEE JSAC. He served as a Member-at-Large of the Board of Governors of the IEEE Communications Society from 2009 to 2011, and is currently its Director of Education.

ENVISIONING THE THIRD PILLAR

As the world's leading organization for communications professionals, ComSoc is a community comprised of a diverse group of researchers and professionals from academia and industry with a common interest in advancing communications science, engineering, and technology.

To continue revolutionizing, ComSoc must evolve and continue to be at the heart of communications technology development in the world. Two of our key and most valuable products until now have undoubtedly been our highly regarded publications and conferences. In the future, we intend to build a third key pillar on Education and Training, which will eventually represent the third leg in ComSoc's leadership, revenues policy, and future development. In this regard, a Strategic Planning Sub-Committee was formed in 2014 as part of ComSoc's Strategic Committee led by Byeong Gi Lee, with the task of making recommendations regarding ComSoc's Education and Training program to the Board of Governors.

The Sub-Committee on Education Programs, Content and Services is chaired by Khaled Letaief. The Sub-Committee held multiple meetings and also had various discussions with the ComSoc's Strategic Planning Committee as well as with the Director of Education and Training, Michele Zorzi. Accordingly, it concluded that ComSoc should offer a world class Training and Professional Education program, addressed primarily to ComSoc's members (both current and prospective), which provides high quality instruction, at a reasonable cost and with easy access, to address the career needs of working professionals in communications and related fields. The proposed program, which takes as its starting point the very successful training program currently available (see http://www.comsoc.org/training), will be developed under the name of ComSoc Continuing Professional Education and Training (CPET) with the following objectives:

1. Extend our existing training and education programs and platforms.
2. Promote the creation of additional new ComSoc courses to build credibility as a preferred provider of knowledge in the areas of information and communications technologies.
3. Share the expense and production challenges by partnering, when possible, with IEEE and Sister Societies in the development and maintenance of the training and education portfolio and with universities to create and deliver courses.
4. Offer both online and face-to-face (or combined online/physical) courses with clear educational and financial objectives.
5. Consider involvement in online education and partnership with consortia like Coursera, or with other societies in an IEEE-level MOOC program.
6. Consider partnership with universities through an agreement that may leave the university the ownership of the material in their own country and provide IEEE with worldwide rights (except the originating country).
7. Engage our Technical Committees to serve as the "content" providers.
8. Engage our Chapters in our professional education program both in terms of content creation/adaptation and updating, and in linking with the local audience, soliciting participation, and hosting face-to-face teaching on Chapter premises.
9. Find a good mechanism to sustain content updating through crowdsourcing (moderated). This will also allow fine tuning to different geographical areas and better market focus.
10. Establish an expanded presence with credible new courses even before the plan is fully worked out.

IMPLEMENTATION AND WAY FORWARD

To implement the above objectives, it was decided that the Vice-President for Technical Activities should take charge of this endeavor with assignment to the Education and Training Board for the actual execution. To that end, much has already been implemented and a variety of initiatives are under development. Below we will describe a few of these.

ComSoc's Education and Training Portal: As a follow-up to the Strategic Planning Committee recommendations, it was decided to create an education portal. The objective is to provide an easy, intuitive, personalized, and user-customizable web-interface for facilitating access to information and services related to education and training offerings. In particular, the goal is to set up a system that can consolidate all of ComSoc's educational offerings as well as provide versatile functions to organize collections of different and multiple sources of information and service resources for dissemination to many users according to their specific privileges, needs, and interests. Such portal should also include cross-links and cross-marketing with communications related education/training offerings from outside of ComSoc, e.g., the Computer Society and IEEE Educational Activities.

As a first step, we are implementing a pilot project to make education content available in the area of software defined radios (including white papers, slides, code, teaching material, etc.), as a follow-up of the Feature Topic on Hands-on Education published in the May 2014 issue of the *IEEE Communications Magazine*. We expect this platform to be operational soon, so as to provide a concrete example of what can be provided and to stimulate feedback that will help us design and evolve the portal.

Online Training and Courses: ComSoc already has a good set of course offerings, mostly focused on wireless communications engineering in the context of our WCET program (see http://www.comsoc.org/training for more information). To further build ComSoc's credibility as the preferred provider of ICT knowledge, we need to experiment with a new set of courses that cover a broader set of topics across the many areas in which ComSoc is active, and include different delivery modes than the ones we currently have. Of particular interest is the development of online courses that should be attractive to our members and meet the needs of the community at large, and especially of professional engineers. A proposal to the IEEE FDC Incubator Fund by Steve Weinstein and Khaled Letaief for the development of such courses was approved in 2014. Accordingly and as a result of that, the following new two-day courses were developed:

- A course on 5G technologies was developed by Ted Rappaport from NYU and taught in December 2014. The course was quite successful and the feedback received was very positive.
- Ashutosh Dutta of AT&T Labs has developed a course on Mobility Protocols and is scheduled to teach it in June 2015.
- Amarnath Gupta of UC San Diego has developed a course on Big Data and is scheduled to teach it in July 2015.

In addition to the above courses, the Education and Training Board is currently reviewing the current ComSoc offering and is coming up with proposals for the development of at least two new courses in 2015. In addition, the Education and Training Board will be creating a sub-committee to coordinate the development, review, and monitoring of the new courses as well as of ComSoc's online and education courses. There may also be a market opportunity for high-quality education in some countries (e.g., in Asia, Africa and Middle East), and ComSoc can play a role in meeting such a need. We are currently discussing some possible pilot initiatives to test whether this may become a concrete possibility.

IEEE ComSoc Summer School: There is no doubt that students are the future of our Society. It is therefore of fundamental importance that we provide special membership development opportunities for them. To do so, we have launched a new summer school whose objectives are to

- Provide high-quality courses on selected topics in our field.
- Engage local chapters in membership development and educational activities.
- Link distinguished lecturers to relevant membership development activities.
- Potentially develop high quality tutorial materials that can be disseminated online to the larger ICT community.

The first summer school will be held on 6-9 July, 2015 in Trento, Italy, and is being organized by Fabrizio Granelli. Of the 100 students who applied, we expect to be able to accommodate about 50 to 60 and to provide some financial support to half of them. In addition, we are working with ComSoc's staff about the possibility of recording the lectures and making them available through our portal, thereby expanding the impact of this important initiative. For further details, please see http://www.comsoc.org/summer-school.

Technical Sessions on Education and Training at ICC and GLOBECOM: We have been proposing special sessions on Education and Training at the two ComSoc's flagship conferences, ICC and GLOBECOM. Such sessions were offered within the Industry Forum program, and were designed to provide a venue for the discussion of education and training resources that can be made available to students and professionals. Following a number of such sessions in the past few editions of these conferences, we are now looking at evolving this initiative into something that is more deeply integrated in the conference's main program. A sub-committee has been created with liaisons to conference development, in order to grow this organically into something more valuable for our members.

***IEEE Communications Magazine* series on Education and Training:** Since May 2014, we have been publishing a series on Education and Training in the *IEEE Communications Magazine*. This series is meant to showcase relevant contributions in the area of education, dealing with topics of importance in this area as well as with current experiences and lessons learned. After the first two issues published in May 2014 ("Hands-on Software Defined Radio") and December 2014 ("Expanding the Student Experience"), this issue includes a topic on "Student Competitions," and another is being advertised for the December 2015 issue ("Ethics Training and Standards"), which will feature IEEE President-Elect Barry Schoop as one of the Guest Editors, along with the Series Editor, Dave Michelson. Future topics being considered for 2016 are "Industry Certification and University Accreditation Programs," and "Education on Standards and Regulations."

ABET Accreditation of Telecommunications Engineering: ComSoc's CPET program will go in line with the ABET's accreditation policy. ComSoc has been supporting ABET by helping them update their accreditation criteria. After more than six years of efforts, Tarek El-Bawab, with the help of several people along the way, succeeded on November 1, 2014 in getting the ABET's approval of the new accreditation criteria for "Electrical, Computer, Communications, Telecommunication(s), and Similarly Named Engineering Programs" (see http://www.abet.org/eac-criteria-2015-2016/). Tarek will be editing a textbook series for Springer, which will publish teaching material for Telecommunications Engineering curricula. This is a great success story for ComSoc, and the Education and Training Board is working on moving it forward while keeping the ComSoc BoG and IEEE fully aware of the developments.

Education Subjects and Materials: For ComSoc's CPET program to be fully developed and remain competitive, it is necessary to secure a concrete production system that will keep the education subjects current on the leading edge and that will generate new materials to support those subjects. Such a systematic approach will be instrumental in providing high-quality, state-of-the art education content to our members and customers. In order to make it possible, ComSoc, with the responsibility of the VP-Technical Activities, should devise new mechanisms to create new education subjects and materials. One very promising example of possible education content are the contributions of Technical Committees, which encompass the overall fields of studies in communications in theory and applications.

In this article we have described some of ComSoc's plans on Education and Training, and some of the recent activities that have been undertaken so far. These are exciting times for Education and Training activities, with great opportunities for ComSoc to confirm its technical and professional leadership in its areas of interest. We do hope you share our enthusiasm about these important achievements and the bright future in front of us. If you have any comments and/or you are interested in any volunteering positions, please feel free to contact Khaled Letaief, VP-Technical Activities, at eekhaled@ust.hk, or Michele Zorzi, Director of Education and Training, at zorzi@dei.unipd.it.

Unlocking The Insights That Inspire Your Next "Eureka!" Moment–Sooner

Keysight's leading-edge measurement tools and sophisticated, future-friendly software cover all phases of the 5G development cycle.

Mark Wallace
Vice President and General Manager
Keysight Technologies, Inc.

For more than 75 years we have been helping you unlock measurement insights, first as the electronic-measurement businesses of Hewlett-Packard and Agilent Technologies, and now, as Keysight Technologies.

From Day 1, there have been two sides to the story. One is the work we do, creating leading-edge instrumentation and software. The other is the work you do: designing, developing, debugging, troubleshooting, manufacturing, testing, installing and maintaining components, devices and systems.

Those seemingly unrelated activities are actually connected by something profound: the "A-ha!" that comes with a moment of insight. When those happen for us, the results are innovations that help you reach new breakthroughs.

Enabling the right idea at the right time

This is our legacy. Keysight is a company built on a history of firsts, dating back to the days when Bill Hewlett and Dave Packard worked in the garage at 367 Addison Avenue in Palo Alto, California. Our firsts began with U.S. patent number 2,268,872 for a "variable-frequency oscillation generator." Appropriately, the heart of Bill's design was a light bulb, which is often used to symbolize a new idea.

Our future depends on your success, and our vision is simple: by helping engineers find the right idea at the right time, we enable them to bring next-generation technologies to their customers—faster.

Offering expertise you can leverage

This is happening in aerospace and defense applications where increasingly realistic signal simulations are accelerating the development of advanced systems that protect those who go in harm's way. It's happening in research labs where our tools help turn scientific discovery into the discovery of new sciences.

It's taking place with 400G Ethernet and the enabling PAM-4 technology, where our end-to-end solution ranges from simulation of new designs to characterization of hardware inputs, outputs and connectors. And in wireless communications we're providing leading-edge measurement tools and sophisticated, future-friendly software that cover all phases of the 5G development cycle.

Within these application areas, there are often more standards than a single engineer can keep up with. That's why so many of our people are involved in standards bodies around the world. We're helping shape those standards while creating the tools needed to meet the toughest performance goals.

Through our global presence, we also have measurement experts near you: our application engineers have the skills and experience to help you unite the hardware and software solutions that meet your unique requirements.

Helping inspire your next breakthrough

To help Keysight customers continue to open new doors, we're concentrating our effort and experience on what comes next in test and measurement. Our unique combination of hardware, software and people will help enable your next "A-ha!" moment, whether you're working on mobile devices, cloud computing, semiconductors, renewable energy, or the latest glimmer in your imagination. Keysight is here to help you see what others can't, and then make it reality—sooner.

KEYSIGHT TECHNOLOGIES

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 29 May 2015. We encourage your careful consideration as you cast your vote for the future success of the Society. The election ends 24 July 2015.

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 2015.

If you do not receive a ballot email from ieee-comsocvote@ieee.org on 29 May 2015 or a paper ballot by 30 June 2015, but you feel your membership was valid before 1 May 2015, 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.
Vijay Bhargava
Past President & Chair
Nominations & Elections

CANDIDATES FOR VICE PRESIDENT

VP – TECHNICAL ACTIVITIES

LUIGI FRATTA

CANDIDATE'S STATEMENT

The success of the IEEE Communication Society depends on its members to get involved and to help. I look forward to leading ComSoc Technical Activities by taking advantage of my diverse background and extensive experience in the communications field.

If elected as Vice-President for Technical Activities, I will serve our Society with a goal to sustain and enlarge its role as a major resource to support our members' needs throughout the world. I will cooperate with the President and the other Vice Presidents to help ComSoc become an even more effective and dynamic organization and provide a higher value to all its members. My activity will rely on the work of all the existing technical committees and the new emerging technical subcommittees to effectively promote scientific research and technological development.

Continuing in the present direction of the evolution of ComSoc, the major goals during my term will be to:

- Ensure that the Technical Committees keep abreast of the latest technology development and evolution.
- Pay attention to the evolving technology improvements and identify 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 the Education and Training Program in new initiatives to increase continuing education and services that are beneficial for our members.

BIOGRAPHY

Luigi Fratta graduated in 1966 in EE from Politecnico di Milano, where he was a professor until 2012. While a professor at Politecnico di Milano he held several positions as a visiting professor and scientist in several universities: UCLA, University of Hawaii, University of Canterbury, New Zealand, Imperial College, UK, IBM T.J. Watson Research Center, Yorktown Heights, IBM San Josè Research Laboratory, Bell Communication Research, Morristown, and NEC Network Research Lab, Japan. He has been consulting with major telecom companies: Siemens, Italtel, Alcatel, and Vodafone.

He has served as Technical Program Chair for INFOCOM '92 and a number of other conferences, including the IEEE LAN MAN Workshop, PIRMC'96, MMT'99, and NETWORKING 2009. He is member of the Steering Committee of: MEDHOCNET, ITC, and IEEE INFOCOM. He has served on the editorial board of several journals, including *Computer Networks*, *Wireless Networks*, and *Photonic Network Communications*, and he was co-guest editor for special issues in IEEE JSAC in 1991 and the *Journal of Communications and Networking* in 2000. He is the author of 150+ refereed papers and holds five patents. From 1982 to 2012 he led several national and European funded research projects. He is an IEEE Fellow (1998) and an IEEE Life Fellow (2009).

HIKMET SARI

CANDIDATE'S STATEMENT

Technical 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 in several leadership positions, and I would be honored to serve our society and membership as Vice President – Technical activities for 2016 – 2017. 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 the 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 a 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 currently a professor and head of the Telecommunications Department at CentraleSupelec, 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 the ENST, Paris, and the 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 Award in 2003, the Harold Sobol Award in 2012, and election to the European Academy and to the Science Academy of Turkey in 2012.

Dr. Sari 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, editor of the *IEEE Transactions on Communications*, associate editor of *IEEE Communications Letters*, and guest editor of IEEE JSAC. Presently he is also serving as executive co-chair of ICC 2016 and executive chair of ICC 2017.

VP – MEMBER RELATIONS

STEFANO BREGNI

CANDIDATE'S STATEMENT

ComSoc is our global home, where we network with top experts and publish our best research. In my 23 years as an enthusiastic volunteer, I have contributed with facts supported by solid accomplishments to serve our community. In GITC, I worked to define the current ICC/GLOBECOM standard paper review process. As conference chair, I ensured that paper review always followed such strict transparent procedures. As a Distinguished Lecturer, in seven years I visited 14 countries and 29 Sections/Chapters worldwide, always preferring areas where students have smaller opportunities to attend global conferences, especially in Latin America and Asia.

As ComSoc Vice-President, Member-at-Large, and Director–Education, I again contributed with facts proved by concrete achievements to facilitate participation from all countries, also addressing economic barriers. I coordinated the successful proposal to bring ICC2016 to Kuala Lumpur, Malaysia. I am on the LATINCOM Steering Committee. I worked to set up new online educational programs.

In my first term as VP-MR, I identified five strategic directions: globalization, academia, industry, women, and students. I led various successful initiatives, e.g. the ComSoc Student Competition and the new *Global Communications Newsletter*. The Women in Communications Engineering Committee has been revamped. We are working on a Child Care Program for parents attending ICC/GLOBECOM.

I need your vote to continue these activities and launch new ones:

- Lower expenses for attending conferences by favoring more affordable locations.
- Facilitate participation in ComSoc from all continents addressing economic barriers.
- Support and involve young students.
- Reorient our proposition for industry.
- Promote the role of women in communications engineering and true equal opportunities.
- Promote free/low-cost online education services for members and reinforce the Distinguished Lecturer Program.

BIOGRAPHY

Stefano Bregni is associate professor of telecommunications at Politecnico di Milano, Italy. He graduated in electronics engineering in 1990. After nine years in industry, he joined Politecnico in 1999. Stefano was an IEEE Distinguished Lecturer for seven years (2003-2009). In ComSoc, he served as: Vice President–Member Relations (2014-15); Board-of-Governors Member-at-Large (two terms: 2010-12, 2013); Director of Education (2008-11); Transmission, Access and Optical Systems TC Chair (2008-09; Secretary/Vice Chair (2002-07); GLOBECOM/ICC Technical Content (GITC) Committee Member-at-Large (2006-09). He received the 2014 ComSoc Harold Sobol Award for Exemplary Service to Meetings & Conferences. He is ICC2016 Technical Program Co-Chair. He has been GLOBECOM2012 TP Vice-Chair, LATINCOM2011 TP Co-Chair, GLOBECOM2009 Vice-Chair for Symposia, Symposium Co-Chair in nine other ICC/GLOBECOMs. He has been the Editor of the *Global Communications Newsletter* since 2007. He has contributed to ETSI/ITU-T synchronization standards. He is the author of 80+ papers and of the book *Synchronization of Digital Telecommunications Networks* (Wiley, 2002).

ELENA NEIRA

CANDIDATE'S STATEMENT

As Vice President of Member Relations, I will bring my passion and energy to understand and respond to the needs of ComSoc members. My plans call for engagement with ComSoc Chapters and with the ComSoc community-at-large in training and career development programs, in event organization, in publications aspects, and in technical initiatives. I will drive diversity and inclusion with activities that are relevant to professional engineers both in industry and academia. More importantly, I will lead strategic efforts to tackle a major challenge of ComSoc and its members in the 21st century: to retain our relevant and prominent role adapting to a new world order where the majority of value, innovation, and growth in communication technologies is being created outside traditional areas.

Presently I am Director of Online Content and sit on the Board of Governors. In this role I have repeatedly demonstrated my focus on membership issues. I conducted a bottom-up redesign of online efforts, which included introducing new media formats, modernization of content, new services, and improvements to editorial boards, to appeal to a broad ComSoc member base, to engage industry and academic partners, and to attract new members. I am also leading member engagement and development via social networks such as Facebook, Instagram, LinkedIn, Tumblr, Twitter, and YouTube. I will continue to expand these and create new programs and services to meet the needs of the newer and emerging communities while ensuring continuity to the established constituencies.

BIOGRAPHY

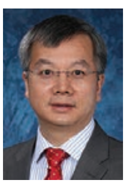
I am an IEEE Senior Member and ComSoc member. I have served as WCET core team member (2008-present), Social Media Vice-Chair (2010-2011), BoG appointed officer (2013-present), *Communications Technology News* and *eNews* editor-in-chief (2014-present), and ComSoc Beats Host

(2014). My career in the mobile communications industry spans more than 15 years and includes leadership, management, and executive roles in R&D, business management, product design, standardization, and operations in TMT (technology, media, and telecommunications) companies including large corporations (Verizon, Texas Instruments, Ericsson) and startups (SNR Labs, Arien Inc). I hold a dual master's degree from the Massachusetts Institute of Technology, an engineer's degree from Universidad Politécnica de Madrid, and a certified mediator title in accordance with Massachusetts General Laws, c. 233, sec. 23C. As current Director of Online Content, my bio can also be found at http://www.comsoc.org/about/bog/ officers/AppointedOfficers.

VP – CONFERENCES

CHENGSHAN XIAO

CANDIDATE'S STATEMENT



IEEE ComSoc sponsors more than 60 conferences with a total of 18,000 submissions and 7,000 published papers. These conferences involve 5,000 volunteers and attract 12,000 attendees. These statistics reveal how important our conferences are as one of the premier platforms for our Society to maintain its leading role in communication and information technologies and for our members to strengthen their professional careers and technical contributions. To best serve those missions, our conferences and publications must address the major challenges to increase attractiveness to industry and academia, maintain technical excellence, strengthen globalization, and improve operational efficiency.

I have had the privilege to serve ComSoc in a variety of technical programs. If elected as Vice-President for Conferences, I will rely on my diverse technical background and global working experience to:

- Shorten the conference submission-to-publication period and maintain the review quality by enhanced training of conference organizers and TPC members.
- Expand our conference portfolios by broadening conference participation from academia, industry, and government and by keeping the conferences affordable.
- Enhance openness and fairness of conferences by improving the transparency of Steering Committee appointments, TPC Chair selections, and paper award decisions.
- Strengthen the diversity of volunteers serving in conferences by engaging people from different geographical and occupational backgrounds in the decision-making process.
- Increase student travel grants to facilitate conference participation of students.

BIOGRAPHY

Chengshan Xiao received the B.S. degree from UESTC, Chengdu, China, the M.S. degree from Tsinghua University, Beijing, China, and the Ph.D. degree from the University of Sydney, Australia. He was a senior engineer with Nortel Networks, Ottawa, Canada, and he is now a professor at Missouri University of Science and Technology, USA. He is an IEEE Fellow, the recipient of the 2014 IEEE ComSoc Joseph LoCicero Award for exemplary service to publications, and a recipient of 2014 Humboldt Research Award. He has published over 80 technical journal papers and holds three US patents. Two of his invented algorithms have been implemented in wireless base station products.

Dr. Xiao is currently the Director of Conference Publications of the IEEE Communications Society (ComSoc), and the Chair of the Steering Committee of *IEEE Transactions on Wireless Communications*. He has served as an elected member of the IEEE ComSoc Board of Governors, the editor-in-chief of *IEEE Transactions on Wireless Communications*, the Technical Program Committee (TPC) Chair for IEEE ICC 2010, the Founding Chair of Technical Committee on Wireless Communications, and a member of ComSoc Fellow Evaluation Committee. He has also served in many other positions for ComSoc conferences, technical committees, and journals.

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 key factors for the success of ComSoc conferences. As a long time ComSoc volunteer, I want to have the opportunity to sustain the best ComSoc conferences and improve other ComSoc conferences so that each of them can best serve the ComSoc community. If elected, I will strive to

- Improve fairness in the paper review process.
- Improve the author experience and reduce the cost of conference publications.
- Develop new revenue structures to keep conference registration fees affordable.
- Increase industry papers presented and increase industry attendance to make for more balanced conferences.
- 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 & 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 premier conferences such as 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.

Larry 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 multiple Symposium Chairs for ICC/GLOBECOM. He served as a General Co-Chair of CNS'2014. He is the 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 *IEEE Transactions on Wireless Communications*. He currently serves as Vice Chair of the IEEE INFOCOM Steering Committee.

VP – PUBLICATIONS

NELSON FONSECA

CANDIDATE'S STATEMENT

ComSoc's publication assets make ComSoc the leading publisher in the area of communications. Nevertheless, maintaining the value of its publications and coping with the pace of technological changes present various challenges. ComSoc's publications should be the most rewarding place for our members to submit their valuable work. For that, fair review processes, sound editorial decisions, and timely publication should be assured. If elected, I will strive to:

- Maintain the high quality and integrity of current publications.
- Work to initiate and foster publications in emerging fields.
- Continue efforts to include volunteers on editorial boards, especially young people and those from industry, as well as promote regional and gender balance.
- Create specific venues for the needs of those in industry.
- Strengthen and expand initiatives such as free on-line tutorials and Best Readings.
- Capitalize on digital dissemination of ComSoc publications, adding new features and social network sharing.

To accomplish this proposed program, I intend to capitalize on my ComSoc experience of genuine service to members. During my term as editor-in-chief of *IEEE Communications Surveys & Tutorials*, the number of submissions quadrupled, making it the journal with the second highest impact factor of all IEEE journals. As EiC, I introduced the current HTML format to ComSoc e-News and greatly increased the global participation in the *Global Communications Newsletter*.

BIOGRAPHY

Nelson Fonseca received his Ph.D. degree in computer engineering from The University of Southern California in 1994. He is a full professor at the Institute of Computing at the University of Campinas, Brazil. He has published 350+ papers and supervised 60+ graduate students. He is a ComSoc Distinguished Lecturer. Currently, he serves as Director for Conference Development of the IEEE Communications Society. He has served as ComSoc's Vice President of Member Relations, Director of the Latin America Region, and Director of On-Line Services. He is the recipient of the 2012 ComSoc Joseph LoCicero Award for Exemplary Service to Publications, the Medal of the Chancelor of the University of Pisa (2007), and the Elsevier *Computer Network Journal* Editor of Year 2001 award. He is a past EiC of *IEEE Communications Surveys & Tutorials*. He is a senior editor for *IEEE Communications Surveys & Tutorials* and a senior editor for *IEEE Communications Magazine*, a member of the editorial board of *Computer Networks*, *Peer-to-Peer Networking and Applications*, *Journal of Internet Services and Applications*, and the *International Journal of Communication Systems*. He created the ComSoc Student Competition Program, IEEE LATINCOM, and the series of Multimedia Communications Symposia in GLOBECOM/ICC. He has been a technical chair for over 10 IEEE conferences.

KHALED B. LETAIEF

CANDIDATE'S STATEMENT

Publications are our life blood, and providing the highest quality information represents the most valuable service IEEE offers to the community at large. But we are facing major challenges such as open access, timeliness, and the need to engage practitioners. If elected, it will be my honor to serve while sustaining ComSoc's role as a major resource supporting the needs of all members throughout the world. I will work with the editors-in-chief to achieve this by:

- Enhancing our products and services to better address member needs.
- Positioning ComSoc to target new areas while putting greater emphasis on making online publications, tutorials, and short courses in emerging technologies widely available to members.
- Strengthening globalization activities and further opening the door to recruiting young volunteers in publications.
- Growing our value to academics while intensifying and encouraging industrial participation.

I have been involved with publications for over 25 years. I also had the privilege to lead IEEE journals as editor-in-chief as well as serving in other capacities. I believe that my extensive IEEE service and leadership experience have put me in a unique position to successfully bring these initiatives into fruition while providing exceptional services to members as we move forward for a strong and healthy IEEE.

BIOGRAPHY

Dr. Letaief received the B.S. with Distinction, M.S. and Ph.D. degrees from Purdue University, USA. He is currently dean of engineering at HKUST. He is the recipient of many awards, including the Michael Medal for Distinguished Teaching, 2007 IEEE Joseph LoCicero Publications Exemplary Award, 2009 IEEE Marconi Prize Award in Wireless Communications, 2010 Purdue University Outstanding Electrical and Computer Engineer Award, 2011 IEEE Harold Sobol Award, 2011 IEEE Wireless Communications Technical Committee Recognition Award, and 12 IEEE Best Paper Awards.

Dr. Letaief has served as a consultant for various organizations including Motorola, Huawei, ASTRI, ZTE, Nortel, and PricewaterhouseCoopers. He is the founding editor-in-chief of *IEEE Transactions on Wireless Communications*, and has served on the editorial boards of other prestigious journals including *IEEE Journal on Selected Areas in Communications – Wireless Series* (as editor-in-chief).

He has chaired many of IEEE's leading conferences and served in many IEEE leadership positions, including ComSoc Vice-President for Technical Activities, ComSoc Treasurer, ComSoc Vice-President for Conferences, and member of IEEE Product Services and Publications Board. He has also served on numerous IEEE committees (e.g. TAB Periodicals Committee, Recertification, Ontology, Technical Activity Council, Publications/Conference Boards, Asia-Pacific Board, Fellow Committee, and Finance Committee). He is IEEE Fellow, HKIE Fellow, and an ISI Highly Cited Researcher.

VP – STANDARDS ACTIVITIES

ROBERT S. FISH

CANDIDATE'S STATEMENT

I am honored to be nominated for a second term as ComSoc VP of Standards Activities. As VP-Standards, I will again emphasize a full cycle approach to standards, including pre-standards emerging technology, market-driven open standardization, and post-standards education with full opportunities for both academic and industry members to participate.

Under my leadership, ComSoc has significantly expanded its portfolio, adding projects in Access Networks, Consumer Networking, SDN/NFV, and IoT to our existing projects such as Dynamic Spectrum and Power Line Communications.

Standards activities should reflect the full breadth of ComSoc's technical scope.By pioneering a Rapid Reaction Standardization methodology, we proactively jumpstarted standardization in Software Defined Networks, Internet of Things, and 5G, which attracted wide participation and significant financial support from IEEE. This support allowed us to subsidize travel costs for academic researchers.

A Communications Standards supplement to *IEEE Communications Magazine* launched in December 2014 and is the vehicle for incubating a new magazine of the same name. We also created a new Conference on Standards in Communications and Networking which will take place in October 2015. Next I will consider starting a Communications Standards journal to serve scholarly research in standards and standards-related disciplines. Looking ahead, further new initiatives will benefit from my service as a BoG member of the IEEE Standards Association and the coordination of ComSoc's activities with IEEE new initiative and future direction funding.

BIOGRAPHY

Robert S. Fish received his Ph.D. from Stanford University. Dr. Fish is President of NETovations, LLC. From 2007 to 2010 he was Chief Product Officer and Senior VP at Mformation, Inc. From 1997 to 2007 Rob was Vice President and Managing Director of Panasonic US R&D laboratories. Prior to this, Rob was Executive Director, Multimedia Communications Research at Bellcore after starting his career at Bell Laboratories. Dr. Fish has over 30 publications and 17 patents. During his career, Dr. Fish and his organizations have initiated and managed standards development activities in IEEE, ISO/IEC, 3GPP, OMA, IETF, ATSC, CableLabs, OSGi, and SDRF.

Rob is VP–Standards Activities of ComSoc. He co-edited a series in *IEEE Communications Magazine* on IEEE Standards in Communications and Networking. He has been a MAL of the ComSoc BoG and Chair of GIMS. Rob is a member of the Board of Governors of the IEEE Standards Association, and a founding member of the IEEE-SA Corporate Advisory Group. For his leadership and contributions to the Multimedia Communications Technical Committee, Rob was the recipient of MMTC's Distinguished Service Award.

KEVIN W. LU

CANDIDATE'S STATEMENT

ComSoc members have contributed significantly to technology advances such as 5G, Machine-to-Machine Communications, Network Functions Virtualization, Software-Defined Networking, and the Internet of Things. At the same time, ComSoc Standards Activities continue to facilitate the progress from research to standards by members' dedication and contribution to research groups, study groups, working groups, and ballot groups on communications standards. Recurring standards workshops and publications have provided members with opportunities to collaborate and stay informed.

I am committed to further broadening opportunities for engaging members in relevant, timely, and useful standards. In addition to face-to-face meetings, we can conduct online communications among researchers, standards developers, and practitioners in all phases of Standards Activities:

- Needs identification, assessment, and selection.
- Liaison to relevant alliances, consortia, and other standards development organizations.
- Standards development, implementation, and maintenance.
- International standards harmonization.
- Standards testing and deployment support.
- Standards training modules and outreach program.

Timeliness and adaptability of Standards Activities are critical to the standards' relevance and usefulness since markets and technologies are rapidly evolving. We can facilitate collaborations among members to further expedite contribution, peer review, discussion, revision, and publication. Most importantly, I want to know your ideas for our Standards Activities, and emerging technologies that you care about.

BIOGRAPHY

Dr. Kevin W. Lu is an adjunct professor of electrical & computer engineering at Stevens Institute of Technology, teaching a graduate course on the Internet of Things. He has served as chair (2012–2013) and advisor (2014–2015) of the ComSoc Standards Development Board, and a member (2013–2015) of the IEEE Standards Association (IEEE-SA) Standards Board's New Standards Committee. He is a member of the IEEE SCC42 Standards Coordinating Committee on Transportation, and the IEEE-SA contact for the Global Standards Collaboration Task Force on Emergency Communications.

Kevin was a chief scientist and executive director at Ericsson/Telcordia Applied Research until 2012, then a senior principal scientist at Broadcom, where he contributed to 3GPP Radio Access Network Working Groups RAN1 and RAN4 until 2013. He was chair (2007–2010) of the TIA TR-48 Engineering Committee on Vehicular Telematics, and authored "All-in-One: Making Connected Vehicles Possible" in the February 2012 issue of *ISO Focus+*. He contributed to the 2011 ATIS Machine-to-Machine Focus Group and the 2011–2014 Strategic Plan for the USDOT Intelligent Transportation System Standards Program. Kevin received a B.S. in control engineering from National Chiao Tung University, and M.S. and D.Sc. degrees in systems science and mathematics from Washington University in St. Louis.

CANDIDATES FOR MEMBERS-AT-LARGE

AMERICAS – NA/LA REGIONS (1-7 AND 9)

ALI ABEDI

CANDIDATE'S STATEMENT

Making global change with local actions: I have demonstrated my commitment to IEEE and ComSoc by serving IEEE for over 16 years on both technical (ComSoc NA Region Board, JCN Associate Editor) and MGA (CLE, GUOS) committees. Most recently, I have served COMSOC Conferences such as PIMRC (PHY Track Chair, 2014), GLOBECOM (Publications Chair, 2014), and WiSEE (TPC Chair, 2014). If elected, I plan to utilize my experience (General Chair for 2010 IEEE Northeast Industry Day, 2010 IEEE Fly By Wireless Conference, and 2013 IEEE WiSEE) to bring industry, academia, and government agencies together and take ComSoc to the next level. I will put particular emphasis on identifying local industry needs and position ComSoc to play a leading role by proposing new ways to communicate these workforce needs to academic institutions. Students and industry professionals around the world are at the core of my activities to achieve this goal.

BIOGRAPHY

Ali Abedi received his BSEE ('96) and MSEE ('98) from Sharif University of Technology, and his Ph.D ('04) from the University of Waterloo. He joined the University of Maine in

2005, where he is currently a professor of electrical & computer engineering, director of the Wireless Sensor Networks Laboratory, and director of the Center for Undergraduate Research at the office of the Vice President for Research. He held visiting scholar appointments at the University of Maryland (2012) and NIST (2012-3), and served as a lecturer at the Air Force University (1998-2000), University of Waterloo (2003-4), and Queen's University (2004-5). Dr. Abedi's research in the wireless communications area is focused on analytical performance evaluation of high performance channel codes, and applications of distributed coding in sensor networks used in space and biomedical applications. He is the co-founder of two startup companies and has published over 80 conference proceedings papers and journal papers as well as three books and two book chapters.

JOSE DAVID CELY

CANDIDATE'S STATEMENT

As a global leader in technology, the Communications Society has important challenges related to new technologies and professional development. If I am elected I will focus on the following goals:

- Strengthen the participation of new members as volunteers, promoting training about ComSoc's organization, programs, and benefits.
- Promote ComSoc globalization by attracting professionals from countries with potential growth.
- Improve the tools and benefits available for young professionals.
- Promote new tools for engaging students as ComSoc members.

My energy and knowledge of the Communications Society and IEEE are a strong guaranty to achieve these goals.

BIOGRAPHY

Jose David Cely graduated as an electronics engineer from the Universidad Distrital Francisco Jose de Caldas, Bogota, Colombia. He has served in several universities as a professor in Colombia; currently he is an assistant professor at the Universidad Distrital Francisco Jose de Caldas. As an IEEE volunteer, he has served in both appointed and elected positions at the Chapter, Section, Region, Society, TAB, and MGA levels. He has been involved in the activities of the ComSoc Colombia chapter since its beginning. Under his leadership the Colombia ComSoc Chapter earned the Latin America Chapter Achievement Award in 2003, the Chapter of the Year Award in 2006, and the Latin America Chapter Achievement Award in 2008 and 2010.

He has participated on the committees of several organized IEEE conferences in Latin America, not only ComSoc but other IEEE societies as well. He has participated actively as part of the organizing committees for LATINCOM 2009, ANDESCON 2010, LATINCOM 2010, LASCAS 2011, LARC 2011, ISGT LA 2011, 2014 T&D LA, and he served as General Chair of 2014 LATINCOM. He was Director of IEEE Communications Society in Latin America Region. In 2010 he was awarded as a volunteer by IEEE Member & Geographic Activities (MGA) with their Achievement Award "for engaging and developing members by organizing conferences within the IEEE Colombia Section."

LEONARD J. CIMINI, JR.

CANDIDATE'S STATEMENT

ComSoc members have played a critical role in the telecommunications revolution; future innovations will be more disruptive. To continue to be influential, ComSoc must transform the products and services we provide, and where, when, and how we provide them. I believe my primary role as a Member-at-Large is to represent, and advocate for, the interests of ComSoc members during this transformation. If elected, I will:

- Work to make ComSoc products more easily accessible, and more affordable, to a broader audience.
- Encourage and facilitate participation in volunteer activities by personally mentoring junior members.
- Strengthen ComSoc's relevance by creating more interdisciplinary technical areas and building a bridge across the growing divide between academic and industrial needs.

I believe that my broad experience in ComSoc, combined with my background in industry and academia, puts me in a unique position to effectively represent the ComSoc membership.

BIOGRAPHY

Len Cimini received his Ph.D. from the University of Pennsylvania in 1982, and worked at AT&T, first in Bell Labs and then AT&T Labs, for 20 years. Since 2002 he has been a professor at the University of Delaware. For more than 25 years he has been very active in all facets of ComSoc, including governance, publications, conferences, and technical activities, and among other positions, has been a Member-at-Large, VP–Publications, VP–Technical Activities, and editor-in-chief of the *IEEE J-SAC: Wireless Communications Series*. He is currently Director of Journals. Len was elected an IEEE Fellow in 2000 for contributions to the theory and practice of high-speed wireless communications. For this pioneering work, he was given the 2007 James R. Evans Avant Garde Award and the 2010 Innovators Award from the New Jersey Inventors Hall of Fame. He has received several ComSoc awards, including the Donald W. McLellan Meritorious Service Award.

TOM HOU

CANDIDATE'S STATEMENT

I joined the IEEE 26 years ago as a student member and later became an active member and volunteer in ComSoc. ComSoc has played a central role in my career advancement. As a member, I hope our colleagues will consider ComSoc as the most valuable professional community in their careers. Over the years I have contributed greatly to ComSoc's important journal editorial boards and conference organization committees. As a Member-at-Large, I will work to support the new initiatives of ComSoc's president and vice presidents and move the Society forward. In particular, I will make efforts to:

- Enhance the value and prestige of ComSoc membership to our members.
- Support ComSoc Board of Governor's initiatives to reach out to industry and produce greater impact on a global scale.
- Nurture young members and students and prepare them as future leaders of ComSoc.

It is my privilege to serve our members and help move ComSoc to a higher level of stature, value, and participation. Thank you for your confidence and vote.

BIOGRAPHY

Tom Hou is Bradley Distinguished Professor of Electrical & Computer Engineering at Virginia Tech, USA. He received his Ph.D. degree in electrical engineering from New York University Polytechnic School of Engineering in 1998. From 1997 to 2002 he was a researcher at Fujitsu Labs in California. Prof. Hou was named an IEEE Fellow for contributions to modeling and optimization of wireless networks. His research was recognized by five best paper awards from IEEE and two paper awards from ACM. He holds five U.S. patents. He served as an area editor of *IEEE Transaction on Wireless Communications* (Wireless Networking area), and an editor of *IEEE Transactions on Mobile Computing*, *IEEE Journal on Selected Areas in Communications – Cognitive Radio Series*, and *IEEE Wireless Communications*. Currently he is an editor of *IEEE/ACM Transactions on Networking*. He is Chair of IEEE INFOCOM's Steering Committee and a ComSoc Distinguished Lecturer. For more information, please visit http://www.tom4comsoc.org.

T. RUSSELL HSING

CANDIDATE'S STATEMENT

I am committed to the IEEE Communications Society's mission of serving humanity through the ComSoc Golden Triangle cornerstones of Globalization, Young Leaders, and Industry. If elected Member-at-Large, I will focus my efforts on creating excellence for the IEEE Communications Society through the following proposed actions:

- Increased promotion of ComSoc collaborations with emerging countries globally.
- Establish more initiatives to stimulate interactions among IEEE ComSoc through academia, industry, and government sectors.
- Create more interactive and online educational services for members.
- Create and maintain global balance for Communications Society officers, leaders, and activities across all geographic regions.
- Develop strategy and executable action plans to continue recruiting more members around the world.

If elected, I will work closely with technical leaders and the IEEE Board of Governors to set visions, strategies, and policies that will make ComSoc a more effective and financially healthy organization, a Society that can respond to rapidly changing dynamics.

BIOGRAPHY

T. Russell Hsing is an IEEE Life Fellow and Fellow of the British Computer Society. He accumulated his rich R&D experience of 35 years as technical staff and research director through affiliations with Xerox, GTE Labs, and Bellcore/Telcordia/Ericsson. He pioneered the commercialization of emerging technologies and services through joint business ventures. He is now a professor and also advisor of the Next Generation Mobile Network (NGMN) Alliance and EDGE Laboratory at Princeton University. He has been an active IEEE Communications Society volunteer for many years. He was a member (2006–2008) and then Chair (2010–2011) of the Fellow Evaluation Committee, as well as a member of the Awards Committee (2010–2012). He was Founding Chair (2010–2012) of ComSoc's Sub-Technical Committee on Vehicular Networks and Telematics Applications. Within the IEEE, he was a member (2008–2010), Chair (2010–2011), and Past Chair (2012) of the IEEE Kiyo Tomiyasu Award Committee, and then the IEEE Eric Sumner TFA Committee (2010–2012). He has been a member of the IEEE Fellow Committee since 2012 and the Strategic Planning Committee in 2013. He is now the Vice Chair of the IEEE TFA Committee.

RULEI TING

CANDIDATE'S STATEMENT

Educated both in China and the U.S., I have conducted research, engineering, and business efforts both in the Asia Pacific region and the U.S. I value my personal experiences in a fast developing technical community as well as in the state-of-the-art research and engineering community. I have been devoting my volunteering efforts to have both sides of the world benefit from each other. Working with the ComSoc Board, I strive to reduce costs while increasing services to our members, including professional education, publications, standards, student exchange via travel, conference participation, chapter activity support, member services, and much more. I have demonstrated passion and results in the telecom profession and by volunteering for society. I hope that my background and experience will continue to be an asset to the IEEE Communications Society.

BIOGRAPHY

Rulei Ting earned his B.S. from Shanghai Jiao Tong University, China, and a Ph.D. from CUNY, NY. Dr. Ting earned an executive master's in technology management at Wharton and Penn Engineering at the University of Pennsylvania. Rulei has been with AT&T & Bell Labs (NJ) for 20+ years, with responsibilities from Distinguished MTS 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 served as Senior Director in telecommunication equipment start-ups. Rulei was awarded the AT&T Bell Labs President Award; IEEE Millennium Award; IEEE Region1 Award, and the President's Volunteer Service Gold Award of USA.

Over the past 15 years Rulei 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 has 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. Rulei has been a positive addition to several ComSoc Board meetings in recent years.

ASIA PACIFIC REGION (10)

VIVEK S. DESHPANDE

CANDIDATE'S STATEMENT

First as a student member and then a professional member, I have been a member of the IEEE Communication Society for many years. Now I am an IEEE Senior Member, and am currently volunteering as Chair for the Communication Society, Pune Chapter. If elected as a Member at Large, I will try to direct more of my expertise to the members within Region 10. The Distinguished Lecture programs need strengthening. Also, more IEEE Communication Society conferences and publications need to be initiated so that the researcher community in Region 10 will have the best platform for their knowledge exhibition.

BIOGRAPHY

I am working as advisor, editor, and reviewer for many international conferences and journals, and chairperson as well as session chair for many international conferences. I have delivered many lectures, tutorials, and technical talks at IEEE International Conferences. I have published more than 75 research papers and have 12 patents to my name. I have explored multiple approaches for implementation of the algorithms and protocols, which proved a good indicator of my research skills. I have also published two books with a renowned publisher in India, and my next book is in queue. The chapter from one of the books from CRC press is already published. Apart from my technical expertise, I exude excellent interpersonal skills and am well versed with analytical skills.

Currently I am working as an associate professor at the MIT College of Engineering, Pune, India. I am also heading the Wireless Network Lab in my organization. In my 25 years of experience, I have 15 years of teaching experience and 10 years of industrial experience. My academic experience includes teaching undergraduate students and postgraduate students. I work extensively in research in the area of wireless sensor networks. I have been instrumental in developing different networking layer protocols with maintaining quality of service parameters in wireless networks.

NEELESH B. MEHTA

CANDIDATE'S STATEMENT

I shall strive to redouble ComSoc's efforts to fulfill its growing Asia-Pacific member base and reach out to it. This involves globalizing its education and training programs, conferences, and improving interaction with local industry. My second goal, affecting all members, is to endeavor to keep membership and conference registration fees reasonable. My third goal is to ensure that ComSoc publications remain beacons of excellence.

Over the past two decades, as a student and then a scientist in the U.S., as a participant in wireless standardization, and as an academician in India, I experienced the significant role played by ComSoc in shaping the future of information and communication technologies. I shall utilize this perspective and my considerable experience as an active ComSoc volunteer and leader to ensure that ComSoc remains a key driver behind these technologies and a key influence on engineers who work on them.

BIOGRAPHY

Neelesh Mehta is an associate professor at the Indian Institute of Science Bangalore, India. Prior to joining IISc he was a scientist in the U.S. from 2001 to 2007 at AT&T Research Laboratories (NJ), Broadcom Corp. (NJ), and Mitsubishi Electric Research Laboratories (MA). He received his M.S. and Ph.D. degrees in electrical engineering from the California Institute of Technology, USA, in 1997 and 2001, respectively, and his bachelor of technology degree from the Indian Institute of Technology (IIT), Madras, India in 1996.

He works in the area of wireless communications. He has co-authored 55+ IEEE transactions papers, 75+ conference papers, and is a co-inventor of 25+ issued US patents. He is an executive editor of the *IEEE Transactions on Wireless Communications*, and an editor for *IEEE Transactions on Communications* and *IEEE Wireless Communications Letters*. He is a Fellow of the Indian National Academy of Engineering and the National Academy of Sciences India. He served as Director of Conference Publications from 2012 to 2013, and as a Member-at-Large since 2014. He also serves on the Education & Training Board and the Marketing & Industry Relations Board.

QIAN ZHANG

CANDIDATE'S STATEMENT

For the past 16 years I have had the privilege of working in both industry (Microsoft Research Asia) and academia (Hong Kong University of Science and Technology). With such experience, I understand and appreciate the needs of both. If elected as a Member-at-Large, I will fully commit myself to the realistic globalization of ComSoc and focus on promoting new Society activities such as interdisciplinary and emerging topics conferences, addressing new telecommunications technologies, and supporting sustainable development, particularly in Asia-Pacific regions. I will also put emphasis on providing encouragement and support to female researchers and get them involved more in ComSoc technical and organizational activities. In short, using my past volunteer experience, I can effectively serve more members by supporting their diverse needs.

BIOGRAPHY

Qian Zhang is currently a professor in the Department of Computer Science and Engineering, the Hong Kong University of Science and Technology (HKUST), Hong Kong. She is also serving as the co-director of the Huawei-HKUST Innovation Laboratory. At Microsoft Research Asia, Beijing, from July 1999 to August 2005, she was the research manager of the Wireless and Networking Group. Qian earned her Ph.D. from Wuhan University in computer science in 1999. She has served ComSoc and IEEE in many roles, including Steering Committee Member of the IEEE ICME Conference, associate editor for several IEEE transactions and magazines, Chair of the Multimedia Communication TC (2008-2010), Chairs of the Chapter Coordination Committee (2010-2012) and Technical Activity Committee (2008-2010) of ComSoc's Asia Pacific Board. She was the TPC Co-Chair of IEEE INFOCOM 2011 and TPC Vice-Chair of IEEE GLOBECOM 2010. She has also served several times as the TPC Area Chair of INFOCOM and Symposium Chair of IEEE GLOBECOM and IEEE ICC.

EUROPE, MIDDLE EAST, AFRICA REGION (8)

KLAUS KOHRT

CANDIDATE'S STATEMENT

For well over 15 years I have been active in the IEEE Communications Society. From the very beginning I developed a strong interest and took an active role in volunteer work by supporting conference organizations both from the steering committee perspective and hands-on in local organizing committees. If elected as a Member-at-Large, I will focus on strengthening industry involvement and having a representation in all aspects of technical and organizational activities by promoting the excellence and leading role of ComSoc, especially for the practitioner. Furthermore, it is my intention to help develop activities such as conferences and discussion panels on interdisciplinary and emerging topics.

BIOGRAPHY

Klaus Kohrt received his masters and Ph.D. degrees in mathematics and computer science from Christian-Albrechts-University (CAU) in Kiel, Germany in 1980 and 1984, respectively. He then joined the telecommunications division of Siemens AG and for many years helped to develop and promote their vision and product strategy. In 2003 he took over government and industry relationship management for Siemens mobile. He served on a number of industry committees, and in January 2004 was elected Vice Chair of the UMTS-Forum.

Currently, he is self-employed as an independent advisor and evaluator for various national and European funded research projects. His primary areas of expertise are in mobile and wireless communication technology and future Internet architecture, in particular with respect to business viability of new technologies.

He has given many keynotes and spoken at numerous international conferences. For many years he contributed to the GICC/GIMS-committee, where he was in charge of the Patronage Working Group, which, under his leadership, developed the Patronage Handbook. He was a member of the organizing committee as Patronage, Publicity or Panels Chair of major ComSoc conferences (ICC, GLOBECOM, WCNC, DySPAN) and has been active in the German sister society VDE as committee member and working group chair for more than 20 years.

CHIARA PETRIOLI

CANDIDATE'S STATEMENT

My 20 years in academia and my recent experience leading a startup company have been driven by the belief that the future of our world and economy will be based on communications and IT. I believe IEEE ComSoc will play a leading role in shaping this future. As an MAL, I will help IEEE ComSoc enhance its services and activities:

- By providing personalized training material to its members, preparing them for the technical challenges in the fast evolving world of communications.
- By further extending the topics covered by ComSoc conferences and journals, maintaining excellence, reducing costs of participation/subscription, and offering new publication venues to novel research areas and multidisciplinary fields.
- By initiating and leading standardization activities on emerging applications and technologies.

I will also devote my energy to expand ComSoc membership and activities across all European regions, involve students in local chapters, increase membership among women, and organize conferences and other events on topics of importance to European researchers and developers.

BIOGRAPHY

Chiara Petrioli received her Ph.D. in computer engineering from Rome University "La Sapienza" (1998). Chiara was a Fulbright scholar (Boston University) and a postdoc (Politecnico di Milano) before joining "La Sapienza" where she is a full professor. She is director of the Sensor Networks and Embedded System laboratory, and the Cyber Physical Systems laboratory. She is co-founder of the spinoff WSENSE s.r.l. Her research focuses on the design and evaluation of mobile and sensing systems. She has contributed over 100 papers with 3400+ citations. Chiara has extensively contributed to IEEE ComSoc activities. She has served on the steering committees of *IEEE Transactions on Mobile Computing* (TMC) and the IEEE SECON conference, and has been an associate editor of IEEE TMC and *IEEE Transactions on Vehicular Technology*. She has served on the organizing and technical program committees of dozens of IEEE events, including TPC co-chair of IEEE SECON 2009 and IEEE INFOCOM 2016.

ROBERT SCHOBER

CANDIDATE'S STATEMENT

I have gathered valuable leadership experience through various professional and volunteer activities over the past several years. If elected Member-at-Large, I will use this experience to foster ComSoc's excellence. Therefore, I will focus on the following three main goals during my term. My first goal will be to further improve the quality and value of ComSoc's journal and conference publications by increasing industry participation, making the review processes more transparent, and promoting high quality papers. My second goal is to provide a larger forum for emerging technologies such as molecular communications, smart grid communication, and social networking in our community, publications, and conferences. My third goal is to increase the benefits for and involvement of students and young professionals through more volunteering opportunities, summer/winter school programs, and free tutorials at conferences.

BIOGRAPHY

Robert Schober received the Dipl.-Ing. and Ph.D. degrees from the Friedrich-Alexander-University of Erlangen-Nürnberg (FAU) in 1997 and 2000, respectively. From 2002 to 2012 he was a professor and Canada Research Chair at the University of British Columbia (UBC), Canada. Since January 2012 he is an Alexander von Humboldt Professor and the Chair of the Institute of Digital Communication at FAU, Erlangen, Germany. He currently serves as editor-in-chief of *IEEE Transactions on Communications* and as Chair of the Steering Committee of the new *IEEE Transactions on Molecular, Biological & Multiscale Communications*. He has served as TPC and Tutorial Co-Chair at various IEEE conferences. He is the Vice Chair of the German IEEE ComSoc Chapter. He has received several awards for his work, including a 2002 DFG Heinz-Maier-Leibnitz Award, the 2004 Vodafone Innovations Award, the 2008 UBC Charles McDowell Award for Excellence in Research, and a 2012 NSERC E.W.R. Steacie Fellowship. He is a Fellow of the IEEE, the Canadian Academy of Engineering, and the Engineering Institute of Canada.

SOCIETY MEMBERS NAMED TO IEEE FELLOW GRADE

Election to the grade of IEEE Fellow is one of the highest honors that can be bestowed upon our members by the Institute in recognition of their technical, educational, and leadership achievements. Only a select few IEEE members earn this prestigious honor.

Congratulations to the following Communications Society members for their election to the grade of Fellow of the IEEE. They now join company with a truly distinguished roster of colleagues.

MATTHEW ANDREWS

For contributions to network design and wireless resource allocation.

JEAN ARMSTRONG

For contributions to the theory and application of orthogonal frequency division multiplexing in wireless and optical communications.

GERHARD BAUCH

For contributions to iterative processing in multiple-input multiple-output systems.

DANIEL BLISS

For contributions to adaptive sensor systems in radar and communications.

AZZEDINE BOUKERCHE

For contributions to communication protocols for distributed mobile computing and wireless sensor networks.

RAJKUMAR BUYYA

For contributions to cloud computing.

JIANNONG CAO

For contributions to distributed computing in mobile wireless networks.

JOSEPH CAVALLARO

For contributions to VLSI architectures and algorithms for signal processing and wireless communications.

BIAO CHEN

For contributions to decentralized signal processing in sensor networks and interference management of wireless networks.

XIUZHEN CHENG

For contributions to localization and detection in sensor networks.

CHEN-NEE CHUAH

For contributions to MIMO communications and network management.

MOOI CHOO CHUAH

For contributions to wireless network system and protocol design.

IAIN COLLINGS

For contributions to multiple user and multiple antenna wireless communication systems.

JOHN DALLESASSE

For contributions to oxidation of III-V semiconductors for photonic device manufacturing.

SAJAL DAS

For contributions to parallel and distributed computing.

NUNO BORGES DE CARVALHO

For contributions on characterization and design of nonlinear RF circuits.

Mrouane Debbah

For contributions to the theory and application of signal processing in wireless networks.

Joe Decuir

For contributions to computer graphics and video games.

Frank Effenberger

For contributions to passive optical networking standards and technology.

Pingzhi Fan

For contributions to signal design for wireless communications.

Michael Fitz

For contributions to the theory and practice of multiple antenna radio.

Xiqi Gao

For contributions to broadband wireless communications and multirate signal processing.

Amitabha Ghosh

For leadership in cellular communication system standardization.

Monisha Ghosh

For contributions to cognitive radio and signal processing for communication systems.

Manimaran Govindarasu

For contributions to security of power grids.

Josep Guerrero

For contributions to distributed power systems and microgrids.

K.V.S. Hari

For contributions to high-resolution signal parameter estimation.

Zhihai He

For contributions to video communication and visual sensing technologies.

Ekram Hossain

For contributions to spectrum management and resource allocation in cognitive and cellular radio networks.

Howard Huang

For contributions to multiple antenna techniques in wireless cellular networks.

Mohan Kalkunte

For contributions to ethernet switching architectures and merchant-switching silicon.

Youngky Kim

For leadership in mobile communication systems.

Deepa Kundur

For contributions to signal processing techniques for multi-media and cyber security.

Thomas Lee

For contributions to the design of CMOS radio-frequency integrated circuits.

Baochun Li

For contributions to application-layer network protocols and network coding.

Jianhua Lu

For contributions to the theory and engineering applications of wireless transmission technologies.

Bhaskar Ramamurthi

For development of wireless technology in India.

XiangYang Li

For contributions to performance analysis and resource allocation in wireless networks.

Detlev Marpe

For contributions to video coding research and standardization.

David Richardson

For contributions to optical fiber technology.

Yunhao Liu

For contributions to wireless sensor networks and systems.

Gianluca Mazzini

For contributions to chaos-based electronic and telecommunication systems design.

Markus Rupp

For contributions to adaptive filters and communication technologies.

Zicheng Liu

For contributions to visual processing for multimedia interaction.

Stefan Mozar

For development of safety solutions for electronic equipment.

Peter Smith

For contributions to statistical modeling and analysis of wireless communication systems.

Wenjing Lou

For contributions to information and network security.

Radha Poovendran

For contributions to security in cyber-physical systems.

Suresh Subramaniam

For contributions to optical network architectures, algorithms, and performance modeling.

David Love

For contributions to feedback-adaptive wireless communication systems.

Robert Qiu

For contributions to ultra-wideband wireless communications.

Dan Keun Sung

For contributions to network resource management.

HSIAO-CHUN WU

For contributions to digital video broadcasting and wireless systems.

LIUQING YANG

For contributions to theory and practice of ultra-wideband communications.

AYLIN YENER

For contributions to wireless communication theory and wireless information security.

BÜLENT YENER

For contributions to network design optimization and security.

WEI ZHANG

For contributions to cognitive radio communications.

HAITAO ZHENG

For contributions to dynamic spectrum access and cognitive radio networks.

No photograph was submitted for the following:

NIHAR JINDAL

For contributions to multiuser multi-antenna communications.

KRISHNA NARAYANAN

For contributions to coding for wireless communications and data storage.

MICHAEL SHEBANOW

For contributions to superscalar out-of-order processors.

WONYONG SUNG

For contributions to real-time signal processing systems.

YIN ZHANG

For contributions to computer network measurement and management.

COMMERCIAL DATA MINING: PROCESSING, ANALYSIS AND MODELING FOR PREDICTIVE ANALYTICS PROJECTS

BY DAVID NETTLETON

MORGAN KAUFMANN–ELSEVIER, 2014, ISBN 978-0-124-16602-8, SOFTCOVER, 304 PAGES

REVIEWER: LUCJAN JANOWSKI

Data mining has become a must-know topic for any scientist, and still it becomes increasingly important for the companies. Such a popular topic has already been described in many books, therefore a question may arise: Why should I read this new and rather thin volume? A fast look inside will convince a potential reader about the lack of equations, a fact surprising for a book describing a mathematical tool, which data mining indeed is. However, lack of equations and thin size appears to be one of many advantages of this book. It is a detailed description of a data mining project. From the first question, “Is it really worth it to start a data mining project?”, which is, unfortunately, asked so rarely in practice, to the last, “How to feed the data mining results to the organization decision-making and operative procedures?”, a data mining project is described with all its important steps.

The book is divided into 19 chapters and contains appendices with some examples. The chapters are grouped into different data mining project steps (chapters 2 to 10), and specific data mining domains (chapters 11 to 19). Chapter 2 describes business objectives definition, the element of data mining projects that is often forgotten or lost in the process of data analysis. Any book related to data mining describes data preparation at least briefly, even if it is typically the most time consuming process. Here, this topic is divided into five chapters. First, Chapter 3 describes potential sources of information, including the related classification, various source specific aspects, and data accuracy. In Chapter 4, data representation is discussed by a short description of basic data types, normalization, and outliers detection. The difference between categorical and numerical variables is discussed. The chapter elaborates on data representation. Chapter 5 presents an important difference between having data and trusting data, dealing with the fact that only after getting to know which data pieces are relevant and reliable, the selection of variables is possible. This topic is described thoroughly in Chapter 6. The next three chapters are focused on the most classical data mining problems. Chapter 7 overviews data sampling and partitioning. Chapter 8 describes data analysis, while Chapter 9 presents data modeling. A very practical problem of using the analysis obtained by a data mining project is described in chapter 10. Chapters 11 to 17 present specific data mining tools and use cases that help a reader to understand the differences between data mining projects. Chapter 18 raises the important problem of data privacy that often limits a range of applicable techniques that can be used by a particular data mining project. The appendices make it possible to follow selected data mining projects in detail.

This work should be at least browsed by a broad range of potential readers, and many classes of them can be envisaged. The volume will be interesting for a specialist working on data mining algorithms and will enable them to understand in a better way how such an algorithms will be used. On the other hand, a company executive director, who makes decisions based on the outcome of a data mining project, and wishes to understand better what stands behind the presented data, will also benefit from the reading. In general, Nettleton’s book is a mandatory volume for anyone who runs data mining projects, since all the steps and most important details that should not be forgotten are described here. Due to the author’s focus on the topic, finding the most important information is easy. The added examples make it easy to understand the presented concept in practical use cases. I strongly recommend this book for anyone even slightly involved with data mining projects.

MACHINE-TO-MACHINE COMMUNICATIONS: ARCHITECTURES, TECHNOLOGY, STANDARDS, AND APPLICATIONS

EDITED BY VOJISLAV B. MISIC AND JELENA MISIC

CRC PRESS, 2014, ISBN 978-1-4665-6123-6, HARDCOVER, 332 PAGES

REVIEWER: EDDY BAJIC

An explosion of communication technologies and of connected objects have invaded our everyday life, as before they have submerged industries and companies in the 90s with the spreading of automatic identification technologies, robotics, and industrial communication networks. Currently, at the dawn of the 4th Industrial Revolution, we entered the era of machine-to-machine (M2M) communications that is tenfold by the profusion of objects connected to the Internet in a wide open cyber-physical world. At a time when more things rather than people are connecting to the Internet, forecasts say that 25 billion devices will be connected by 2015, and 50 billion by 2020. The question is thus arising: How will all those devices communicate and exchange information and for what benefits and applications? That is what this book is about, presenting how M2M communications offer new opportunities for modern machines to cooperate with and without human interaction, up to the implementation of Internet of Things (IoT) concepts. Being part of a global cyber physical system, M2M communications offers relevant solutions to address the challenges of innovative intelligent control, monitoring, and reporting for manufacturing systems and servicing applications in suitable domains, such as smart grid, e-health, or transportation management.

The book offers a coherent collection of chapters developing the scientific study and analysis of cutting-edge topics in M2M networking, all written by scientists and researchers from several universities in different countries, also accompanied by experts from telecom companies. The book is organized in three parts, with 10 well-balanced chapters, providing a wide cross section of many topics related to M2M communications, such as architecture, standards, technologies, and applications.

The first chapter presents a description of modern M2M architectures and communication standards from ETSI and 3GPP with a scan of relevant case studies. Chapter 2 presents a specific focus on communication modeling, presenting the elementary principles and formulations of traffic factors. Chapter 3 suitably shows how the M2M and IoT concepts are beset with important new challenges, including big data analysis, reliability, privacy, and security issues, which are critical for large scale applications such as manufacturing automation and smart cities on a wider scope. The second part, embracing Chapters 4 through 7, provides several detailed scientific studies and presents the research breakthrough on methods and technologies allowing development of reliable and energy-efficient M2M communications. In particular, an interesting focus sharply oriented toward a scientist audience is put on modeling of throughput performance of wireless standard 802.15.4. This standard is currently implemented on widespread protocols for IoT device communication such as Zigbee and 6LoWPAN, though never cited in the text. Remarkably, Chapter 7 treats energy-efficiency in M2M communications by reporting how access control, routing, and quality of service impact energy consumption, a critical issue for deployment of a large number of autonomous communicating devices. The last part presents three chapters (8 through 10) focused on specific issues on relevant application cases for smart grid and mobile crowd-sensing. This chapter, being more technological, will be received with pleasure by engineering readers, who will find presentations of available M2M technologies that implement M2M concepts to build the current and future cyber-physical world of communication solutions.

This book is built on a scientific structure with each chapter giving a relevant reference list of standards and research papers. Therefore, it can be recommended principally to scientists, researchers, and students. It will also serve as an introduction to the baseline concepts and technologies of M2M. Mainly, it provides analysis of M2M protocols, architectures, and standards embracing the development of the cyber physical world surrounding us by billions of communicating devices, the world that we are very close to living in already.

Conference Calendar

Updated on the Communications Society’s Web Site
www.comsoc.org/conferences

2015

MAY

IEEE CQR 2015 — IEEE Int’l. Workshop Technical Committee on Communications Quality and Reliability, 10–15 May
Charleston, SC.
http://www.ieee-cqr.org/

IEEE CTW 2015 — IEEE Communications Theory Workshop, 10–13 May
Dana Point, CA
http://www.ieee-ctw.org/

ONDM 2015 — 2015 Int’l. Conference on Optical Network Design and Modeling, 11–14 May
Pisa, Italy.
http://ondm2015.sssup.it/

IFIP/IEEE IM 2015 — Int’l. Symposium on Integrated Network Management, 11–15 May
Ottawa, Canada.
http://im2015.ieee-im.org/

IEEE BlackSeaCom 2015 — IEEE Int’l. Black Sea Conference on Communications and Networking, 18–21 May
Constanta, Romania
http://www.ieee-blackseacom.org/2015/ index.html

IEEE 5G — 1st Int’l. 5G Summit, 26 May
Princeton, NJ
http://www.5gsummit.org/

UBI-HEALTHTECH 2015 — 2nd Int’l. Symposium on Future Information and Communication Technologies for Ubiquitous HealthCare, 8–12 May
Beijing, China.
http://www.ubi-health.org/

JUNE

IEEE ICC 2015 — 2015 IEEE Int’l Conference on Communications, 8–12 June
London, U.K.
http://icc2015.ieee-icc.org/

IEEE IWQOS 2015 — IEEE/ACM Int’l. Symposium on Quality and Service, 15–16 June
Portland, OR.
http://www.ieee-iwqos.org/

–Communications Society portfolio events appear in bold colored print.

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

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

GCN GLOBAL COMMUNICATIONS NEWSLETTER

May 2015
ISSN 2374-1082

DISTINGUISHED LECTURE TOUR

Advancement of Technology and Use in Managing Events and Enriching Attendee Experience: Preparing the GLOBECOM 2014 IT Infrastructure

DLT/DSP Event, Dec. 2014, New Orleans

By Fawzi Behmann, Austin Chapter Chair, USA

After a successful IEEE GLOBECOM 2014 conference held in Austin (December 8-12, 2014), I was invited by Richard Miller, the New Orleans COMSOC chapter chair and also the Social Program Chair at GLOBECOM 2014, to go to New Orleans and share my experience with the New Orleans Chapter and section, as current Chair of the IEEE COMSOC chapter in Austin and Local Arrangement & Marketing Chair for GLOBECOM 2014.

The event at New Orleans was held December 22 in one of the local restaurants. The event started at 6 pm with one hour for networking followed by dinner and then the presentation. There were 17 attendees (five non-IEEE members). Some of the IEEE members came from a long distance to be part of the event, such as R5 Director Francis Grosz. Many of the members are technology savvy and asked many interesting questions. This made it the talk interactive and interesting.

I had the opportunity to deliver a presentation on the topic "Advancement of Technology and Use in Managing Events and Enriching Attendee Experience."

The presentation presented an overview of technology advancement, and focused on several use cases, including conference IT Infrastructure and the conference mobile app. Examples were provided that covered events such as GLOBECOM 2014.

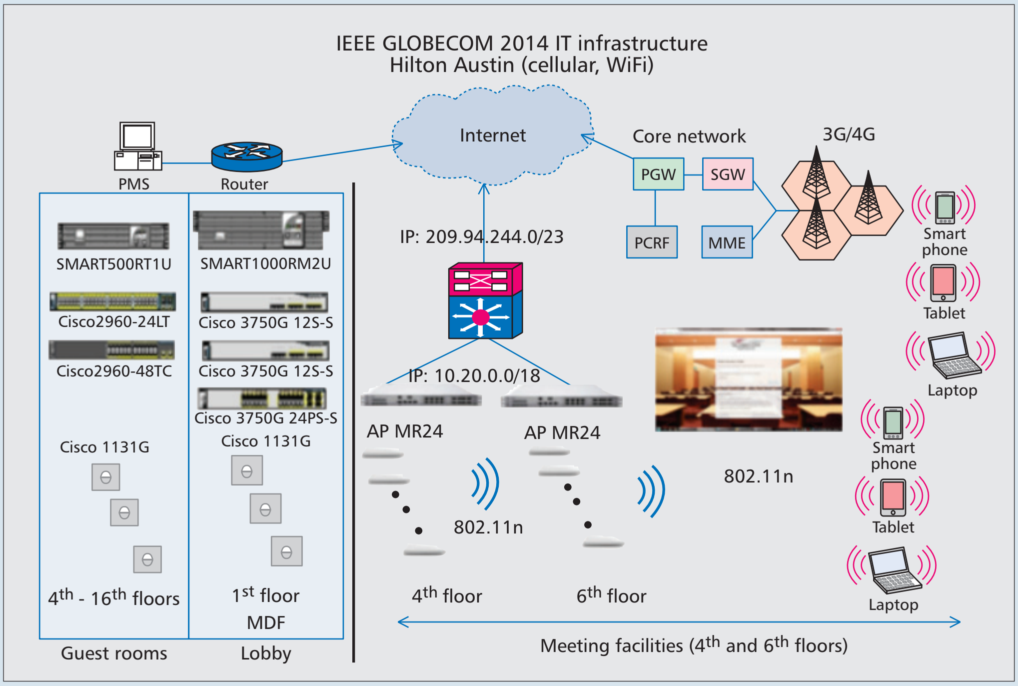
The core of the presentation focused on the preparation and complexity related to the GLOBECOM 2014 IT Infrastructure. The key objective was to provide a sustainable WiFi/cellular signal and coverage at all meeting facilities, common areas, and guest rooms.

A committee of 26 members (with representation from AT&T, Hilton, and IEEE) was formed and led by Fawzi Behmann. The committee members worked diligently and professionally for close to two years with more than 10 meetings conducted. The milestones of the project were:

- Conduct on-site feasibility assessment (Q3 2013).
- Test the upgraded infrastructure against SXSW event (Q1 2014).
- Address other remaining items such as one-level authentication, security branding, and reliability (Oct. 2014).

This is the first time the GLOBECOM conference was supported by a state of the art WiFi/cellular IT infrastructure spanning the conference meeting areas, common areas, lobby, and guest rooms for a sustainable signal and coverage. Over 3,100 client devices used the WiFi/cellular network with intensive usage (180 Mbps bandwidth), which is equivalent to 7,500 client devices used during SxSW, a major event held in March 2014. This comparison shows that GLOBECOM 2014 attendees are more savvy and heavy users of the network. The infrastructure easily handled the intensive concurrent download of conference proceedings.

After the talk, I received positive feedback by many, who commented that the subject was interesting and the delivery made the content easily understandable. To many attendees, especially those retired, the topic brought them up-to-date on the advancement of technology and use cases. Others, were able to ask several technical questions about the network that was put in place and the cloud based management tools for monitoring the performance of the network in real time identifying congestion and responding to abnormal situations.



IEEE GLOBECOM 2014 IT infrastructure, Hilton Austin (cellular WiFi).

CONFERENCE REPORT

Highlights from the 6th International Workshop on Reliable Networks Design and Modeling (RNDM 2014)

By Jacek Rak, Poland; James P.G. Sterbenz, US/UK; Gangxiang Shen, China; Dimitri Papadimitriou, Belgium; Krzysztof Walkowiak, Poland; and Boris Bellalta, Spain

Prof. Eusebi Calle (left), Prof. Jose L. Marzo (middle), and Prof. Bart Lannoo (right) delivering their keynote talks.

Presentation of the Best Paper Award (left to right: Jacek Rak, RNDM 2014 General Chair, Boris Bellalta, RNDM 2014 Organising Chair, Rein Houthooft - recipient of the award, and. James P.G. Sterbenz, RNDM 2014 Co-chair).

The 6th International Workshop on Reliable Networks Design and Modeling (RNDM 2014), technically co-sponsored by the IEEE Communications Society and endorsed by its Technical Committee on Computer Communication (TCCC), was organized in Barcelona, Catalonia, Spain on November 17-19, 2014. The host university was Universitat Pompeu Fabra (UPF), one of the most prestigious technical education institutions in Spain.

Established in 2009, this annual single-track event has rapidly become one of the leading workshops on network resilience and dependability, each time gathering world-class researchers from both academia and industry. Other technical co-sponsors of RNDM 2014 included: IFIP TC6, the IEEE Spain Section, the V.A. Trapeznikov Institute of Control Sciences of RAS, and the Russia (Northwest) Section BT/CE/COM Joint Chapter.

RNDM also offered two co-located half-day workshops: the 1st International Workshop on Survivable Content-Oriented and Cloud-Ready Networking (S2CN), and the 2nd International Workshop on Understanding the Inter-play between Sustainability, Resilience, and Robustness in Networks (USRR).

The main goal of the S2CN workshop was to stress the importance of survivability aspects in the context of increasingly growing content-oriented networks and cloud computing services. S2CN is intended to provide an international forum for both academia and industry on the topic of survivable content-oriented and cloud-ready networking. The S2CN workshop was supported by ENGINE, the European research centre of Network intelliGence for INnovation Enhancement, and the European Commission under the 7th Framework Programme, Coordination and Support Action, Grant Agreement Number 316097 (http://engine.pwr.edu.pl/).

As technical networks exhibit many inter-dependencies that are complex to measure and model but also lead to multi-objective decision/robust optimization problems where uncertainty becomes the transversal notion to capture, the aim of the USRR 2014 workshop was to develop a better understanding of the fundamental interplay between sustainability, resilience, and robustness essential to rejuvenate and improve current design and evaluation methods that are unable to cope with this fundamental dimension. USRR 2014 was supported by the EINS project, the FP7 European Network of Excellence (NoE) in Internet Science funded by the European Commission DG CONNECT.

A significant increase in community interest in RNDMs has been observed over the years. Despite being located near the European research community, RNDM has so far attracted many attendees and presenters from non-European countries, including the USA, Canada, Japan, China, and Uruguay. This year's edition was no exception.

A total of 66 regular submissions authored by researchers from over 30 countries were extensively reviewed by 70 TPC members and over 40 external reviewers. As a result, each RNDM 2014 paper received at least four reviews, while the average number of delivered RNDM 2014 reviews was 4.27. The 35 accepted manuscripts were organized as full and short papers into five RNDM 2014 technical sessions, entitled: Resilient Routing Strategies; Theory of Resilient Routing; Network Optimization; Resilience of Converged Services; and Network Reliability Assessment. There were two technical sessions of S2CN 2014: Survivability of Elastic Optical Networks; and Optimization of Survivable Content-oriented and Cloud-ready Networks. There were also two other sessions for the USRR 2014 workshop.

Participants of RNDM 2014.

The technical program of RNDM 2014 was enriched by two keynote talks: "Robustness Analysis of Networks under Large-scale Failures: Drawing Robustness Surfaces" by Prof. Jose L. Marzo and Prof. Eusebi Calle (University of Girona, ES); and "What is the Impact of Network Protection on the Energy-Efficiency?" by Prof. Bart Lannoo (Ghent University–iMinds, BE). There were also three invited talks: "Power Efficient Service Differentiation Based on Traffic-Aware Survivable Elastic Optical Networks", "Fragmentation-aware Survivable Routing and Spectrum Assignment in Elastic Optical Networks", and "Modeling and Impact of Flexible WDM Grid with Variable Channel Rates on Multiple Network Layers" by Ioan Turus (Technical University of Denmark, DK), Çiçek Çavdar (KTH, SE), and Achim Autenrieth (Adva Optical Networking, DE), respectively.

In addition, RNDM offered a panel discussion, "SDN – Revolution or Evolution in Network Reliability", addressing emerging resilience issues of software defined networking chaired by James P.G. Sterbenz with three panelists: Achim Autenrieth (Adva Optical Networking, DE), Krzysztof Walkowiak (Wroclaw University of Technology, PL), and Paul Smith (Austrian Institute of Technology, AT).

The RNDM 2014 Best Paper Award ceremony took place during the closing session. Four papers were nominated. The final decision was based on two factors: average overall scores of submitted manuscripts (based on recommendations of reviewers), and presentation quality (scored by chairs of RNDM technical sessions). This year the award was given to two papers: "Fault-tolerant Greedy Forest Routing for Complex Networks" by Rein Houthooft, Sahel Sahhaf, Wouter Tavernier, Filip De Turck, Didier Colle, and Mario Pickavet from Ghent University – iMinds, BE (presented by Rein Houthooft); and "Region-Based Fault-Tolerant Distributed File Storage System Design under Budget Constraint" by Anisha Mazumder, Arun Das, Chenyang Zhou, and Arunabha Sen from Arizona State University (US) (presented by Arunabha Sen).

(Continued on Newsletter page 4)

LECTURE TOUR REPORT

Amateur Radio Lectures In India, September 2014

By Miroslav Skoric YT7MPB, IEEE Austria Section

Many IEEE members are 'hams' (radio amateurs), and as Doug Zuckerman W2XD, 2008-2009 President of the IEEE Communications Society, pointed out in an email correspondence: "many of these [members] had ham radio as the launching point for their careers". That is one of the reasons I have been volunteering by promoting amateur radio communications.

My travels are usually connected to conference lectures, but to make this self-funded voyage more cost-effective, I decided to start with a two-day session on "Amateur Radio Digital Information and Communication Technologies", organized with the Surdar Vallabhai National Institute of Technology in Surat (known as NIT Surat), which is one of the most prestigious institutions of its kind in the Indian state of Gujarat. Thanks to Prof. Mrs. Upena Dalal and her associate Mrs. Shweta Shah, several female students were in the audience. In contrast with western countries, it seemed that Indian schooling in technology does not suffer from decreasing interest by girls and women.

The main event of my journey started a couple of days later. With the organizational support of Mohan Ram VU2MYH, Director of the National Institute of Amateur Radio (NIAR), one of the most influential amateur radio unions in India, I conducted a tutorial session with the 11th International Conference on Wireless and Optical Communications Networks (WOCN 2015), organized at Koneru Lakshmaiah University (KLU) in Vijayawada, the newly appointed capital of Andhra Pradesh state, and supported by the IEEE Hyderabad Section. Although the KLU campus is located in Vaddeswaram village, nearly half an hour by car from Vijayawada city center, it posses enough telecommunication equipment and has a good location for constructing an amateur radio relay facility.

After returning to Hyderabad, the main city of Telangana state, NIAR organized two lectures. The first was held at Gokaraju Rangaraju Institute of Engineering and Technology (also known as GRIET), where Jose Jacob VU2JOS, Deputy Director of NIAR, performed practical parts of the session. The second lecture was held at the Department of Electronics and Communication Engineering at Vardhaman College of Engineering. The overall logistics for both sessions was provided by N. Venkatesh, vice chair of the IEEE ComSoc/SPS Hyderabad Joint Chapter. After opening words by Prof. Zafar Ali Khan of the Indian Institute of Technology Hyderabad (Fig. 1, second from the right), Mr. Ram (Fig. 1, at the lecturn) and I prsented the rest of the lecture. The room was fully occupied by students and staff. Unfortunately, that part of India suffers from electrical blackouts, so the program was interrupted several times.

Fig. 3: Lecture at Indian Institute of Information Technology and Management (IIITM), Gwalior, Madhya Pradesh state.

Fig. 1: Lecture at Department of Electronics and Communication Engineering at Vardhaman College of Engineering, Hyderabad.

Fig. 2: Workshop with Department of Telecommunication Engineering at BMS College of Engineering in Bangalore.

A couple of days later, another two-day workshop was scheduled with the Department of Telecommunication Engineering at BMS College of Engineering in Bangalore, thanks to joint efforts of Dr. Srinivas Talabattula, chair of the IEEE ComSoc Bangalore Chapter, and Munir Mohammed, program specialist at the IEEE India office in Bangalore (Fig. 2, standing on the left).

The next stop on my journey was Chennai (formerly Madras). The first day I was a special guest of the IEEE Madras Section, by having a lecture introduced by Prof. Rama Rao from SRM University. The next morning the main event was a lecture at the huge SRM campus, located some 45 minutes by car outside the city. Approximately 100 people participated.

The final portion of this travel included three educational institutions in Gwalior, Madhya Pradesh state. Thanks to Jayant Bhide VU2JAU, the local amateur radio leader (Fig. 3, sitting second on the right), the first session was conducted with students and staff of the Indian Institute of Information Technology and Management (IIITM Gwalior). The next morning session we prsented in front of a younger audience at Gwalior Glory High School. The afternoon's session was presnted at the Department of Electronics & Communication Engineering at IPS College of Technology & Management. Once again, it was encouraging to see many female students in the rooms. The future seems to be bright.

Having good impressions, I can say that Indian education is eager for novelties and challenges of many kinds. They are not reluctant to ask for more. It is obvious that there will be more amateur radio tutorials and workshop sessions in years to come. The plans include establishing an "international conference on the amateur radio in education", as well as local events in the form of "summer schools". In that direction, people mentioned in this report, and some others who were not listed here, have addressed open calls for prospective participants in such events. Should you want to collaborate as a ham radio instructor, please do not hesitate to contact me (skoric@ieee.org).

CONFERENCE REPORT

5th International FOKUS FUSECO FORUM: Smart Communications Platforms for Seamless Smart City Applications

By Thomas Magedanz, General Chair, Fraunhofer FOKUS/TU Berlin, Germany

The FOKUS FUSECO FORUM (FFF) 2014 was the fifth event in the successful series of FFF and attracted approximately 230 academic and industry telecommunication specialists from 31 countries. Two full days with technical tutorials, interactive workshops, conferences, booths, and live demonstrations were offered. For the first time it was preceded by the First FUSECO Forum Asia, held in Bali, Indonesia in June 2014.

Tutorials, Interactive Workshops, and Demonstrations

Three main technical directions were addressed by a series of tutorials and workshops: cloud-based smart communication platforms; 5G and core network evolution based on SDN and NFV; and Internet of Things and machine to machine (M2M) solutions and services. Specifications, technical aspects, and practical use cases of these directions were thoroughly reviewed.

As in prior years, Fraunhofer FOKUS demonstrated the new advances in its Future Seamless Communication (FUSECO) Playground (www.fuseco-playground.org) testbed and its latest FOKUS OpenXXX toolkits, namely OpenSDNCore, Open5GCore, and OpenMTC, which address each of the main technological directions in telecom today: mobile broadband, 5G, cloud, SDN, NFV, and M2M/IoT.

Fraunhofer Fokus was particularly happy to announce the new End-to-End 5G Research Laboratory together with the Fraunhofer Heinrich Hertz Institute. 5GBerlin (www.5g-berlin.de) is an initiative for collaborative research toward 5G to be tested in one place, including 5G core, access and photonics technologies.

International FI-PPP Workshop

In conjunction with the FOKUS FUSECO Forum 2014, XiFi (www.fi-xifi.eu), the capacity building part of the FI-PPP program, organized an international FI-PPP workshop, a full day interactive session for the regional and international promotion of the European Future Internet – Public Private Partnership (FI-PPP) program on November 13, 2014. With participation of more than 50 guests and speakers, the international FI-PPP workshop successfully provided a comprehensive overview of worldwide Future Internet programs and the status of the FI-PPP program in particular.

Conference Opening/Keynote Sesssion with Prof. Magedanz (TU Berlin) and Guru Parulka (Stanford University).

Keynote, Conferences, and Panel Discussions

The FFF14 Conference Day was opened by a short welcome note given by Prof. Dr. Manfred Hauswirth, the new director of Fraunhofer FOKUS and a globally known expert on the Internet of Things and data analytics research communities.

Mr. Guru Puralkar from ONF/Standford University gave a keynote on software defined networking (SDN), network functions virtualization (NFV), cloud principles, and the role of open source software for driving innovation in the Internet and telecom world.

Five sessions followed, with representation from experts from the leading operators, vendors, consulting companies, and academia. Topics were: broadband access convergence in smart cities; evolution path to 5G based on SDN and NFV; human to human (H2H) communications in smart cities, an interesting comparison between virtualized solutions and WebRTC/OTT communications; cloud-based telco platforms enabling competitive smart city applications; M2M in smart cities and IoT as a key driver for innovative smart cities. Afterward there was a session on best practices from around the world for smart cities as Future Internet showcases.

Conclusion

Throughout the extensive review, discussion, and demonstration of the new technological advances in Telecom, it was confirmed that based on global Future Internet research activities, the ongoing virtualization and "cloudification" of service and network infrastructures is globally progressing rapidly and that new application domains beyond voice and video communications are emerging mainly based on Machine-to-Machine communications under the banner of smart cities and the Internet of Things.

Outlook: 2nd FUSECO Forum Asia and FOKUS FUSECO FORUM 2015

Based on the global relevance of the addressed topics and technologies, and following the continued success of the FOKUS FUSECO Forum series in the last decade, the 2nd FUSECO Forum Asia is planned for the end of May 2015 in Bali, Indonesia. The next FOKUS FUSECO Forum will be held in Berlin in the middle of November 2015. Key topics addressed by FFF15 will be 5G, SDN/NFV, IOT/M2M, and data analytics.

For more information, including more detailed event minutes, pictures, and presentations, or more information about the events to be held in 2015, please refer to: www.fuseco-forum.org/ and www.fusecoforum.asia.

RNDM 2014/*Continued from page 2*

As a tradition of RNDMs, in addition to IEEE Xplore publication, participants were also provided with printed as well as electronic proceedings. Authors of the top analytical RNDM 2014 papers were invited to submit the extended versions of their contributions to a special issue of *Networks Journal* (Wiley).

The next edition, RNDM 2015, will be held in Munich, Germany, on October 5–7, 2015. More information on RNDM 2015 is located at http://www.rndm.pl.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

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

IEEE Communications Society

Stefano Bregni, Vice-President Member Relations
Pedro Aguilera, Director of LA Region
Merrily Hartmann, Director of NA Region
Hanna Bogucka, Director of EAME Region
Wanjiun Liao, Director of AP Region
Curtis Siller, Director of Sister and Related Societies

Regional Correspondents Who Contributed to This Issue

Thomas Michael Bohnert (thomas.bohnert@zhaw.ch)
Ewell Tan, Singapore (ewell.tan@ieee.org)
Alexey Vinel, Finland (alexey.vinel@gmail.com)

A publication of the IEEE Communications Society
www.comsoc.org/gcn
ISSN 2374-1082

EMERGING APPLICATIONS, SERVICES AND ENGINEERING FOR CELLULAR COGNITIVE SYSTEMS

Muhammad Zeeshan Shakir

Octavia A. Dobre

Muhammad Ali Imran

Apostolos Papathanassiou

Zhongshan Zhang

Athanasios V. Vasilakos

Honggang Wang

Hiroshi Harada

The rapid advancements in wireless communications are expected to increase the demand for radio spectrum resources by orders of magnitude during the next decade. This problem must be addressed using technology and regulatory innovations for significant improvements in spectrum utilization in cellular networks. Emerging cognitive radio technology has been identified as a high-impact disruptive innovation that could provide solutions to the cellular traffic congestion problems and exploit the under-utilized spectral resources to pave an application-driven path toward the next generation of cellular technology.

Along with the advent of the 5G era in telecommunications systems, new emerging applications, services, and engineering (EASE) frameworks are now required to be facilitated with the advances in cognitive cellular systems (CCSs) to jointly address the socio-scientific challenges of global significance and catalyze the diversification of the world economy. This Feature Topic is a new addition in the scientific community to introduce the EASE framework. The goal of this Feature Topic is to showcase contributions from experts to further identify and present recent results and technical challenges related to new applications, and engineering, as well as the convergence of associated infrastructure for future implementation in CCS. The Feature Topic received a large number of submissions, and it has been a challenging task to select the best and most relevant papers. We have selected the top 15 research articles out of 47 submitted articles. The accepted articles are being published as two parts in *IEEE Communications Magazine*.

Part 1 of the Feature Topic presents the landscape of future spectrum requirements by exploiting emerging frameworks including spectrum aggregation, resource management and spectrum harvesting techniques to facilitate the ever-increasing demand for ubiquitous connectivity over the limited spectrum. This part also covers business models and network economic issues associated with the implementation of CCSs.

The first two articles are related to dynamic spectrum aggregation and spectrum access in future CCSs. The article by Hanna *et al.* presents potential methods for dynamic spectrum aggregation, and emphasizes the opportunities and issues for 5G communications in CCSs. The article focuses on two multicarrier techniques, enhanced non-contiguous orthogonal frequency-division multiplexing and non-contiguous filter-bank-based multicarrier. Issues beyond traditional spectrum access and shaping are considered, related to the aggregation dynamics, nonlinear effects in the transmission chain, as well as receiver synchronization and performance. Pros and cons are illustrated, regarding such issues as complexity, the need for lower peak-to-average power ratio and out-of-band power reduction, as well as interference-robust reception and synchronization. Additional directions of research are pointed out, such as increasing the flexibility of the RF front-end design to operate in multiple bands and handling the aggregated interference resulting from multiple cognitive radio transmissions. The article by Guizani *et al.* looks at the scalability challenge for dynamic spectrum access systems. To this end, the article presents viable distributed methods that are scalable to a very large number of nodes. The need for information exchange between distributed decision making nodes introduces delay, and this article summarizes how game theoretic techniques, learning (e.g., Q-learning), and filtering (e.g., particle filtering) approaches have aimed to achieve distributed cognitive solutions with minimal information exchange. The authors provide a comparison of learning and particle-filtering approaches for distributed dynamic spectrum allocation methods and show that the particle-filtering method achieves better throughput performance in fast changing dynamic environments.

The third article, by Saad *et al.*, presents a matching theory-based framework for dynamic resource allocation and management in emerging wireless networks. Centralized optimization techniques and game theoretic approaches have reached their limitations in solving the dynamic resource allocation problem in emerging highly dense wireless networks. Therefore, the primary objective of this article is to introduce the new engineering-oriented classes of matching theory, and provide robust and stable optimal treatment in the context of resource allocation problems for various heterogeneous-type networking environments. This interesting tutorial-oriented exposure of this Nobel Prize winning framework could potentially contribute to the further theoretical analysis of future wireless networks and their implementation.

The fourth article, by Zhang *et al.*, investigates interesting potential solutions for solving the challenge of spectrum scarcity. The authors propose an integrated cooperative framework for cooperative spectrum harvesting. They consider cooperation between the primary and secondary nodes as well as among a cluster of nodes in order to identify and harvest spectrum opportunities more efficiently. Their results show that significant improvements in spectrum efficiency and expected available time can be achieved using the proposed framework. An interesting aspect of the framework is the model to exchange credits between the cooperating nodes, which can potentially develop into a more practical business model in future CCSs. The impact of this research can be far reaching if extended to explore the potential of spectrum aggregation, utilization of narrow spectrum white spaces, and coexistence of licensed and license-exempt spectra in cellular scenarios of future generations.

In the fifth article, Mustonen *et al.* give a comprehensive review of the European licensed shared access (LSA) concept [4] spectrum sharing between a mobile network operator (MNO) and an incumbent spectrum user. The article views this novel application area of cognitive radio from the MNO point of view and provides valuable system design criteria that need to be considered by the MNO in order to succeed in utilizing the LSA opportunity as an additional means for meeting the constantly growing mobile traffic demand. The article is also valuable not only because it describes existing LTE/LTE-Advanced technologies, which can appear useful in introducing LSA in a cellular network, but also because it introduces the LSA management unit required to be implemented on top of the existing cellular network for enabling and optimizing LSA.

Network economics are the soul of CCSs, and therefore it is essential to carefully study and plan these aspects for CCSs to strive. It has been realized that the network architecture, algorithms, and protocols cannot simply be designed without considering the socio-economic aspects involved. The sixth article, by Jiang *et al.*, presents both the simultaneous and sequential behaviors in CCSs, with scenarios and examples to illustrate possible solutions. The authors highlight the network economic issues in CCSs from the perspectives of game theoretic modeling and mechanism design to reveal the fundamental problems and corresponding enabling techniques. The goal of this article is to provide an overview to understand the motivation, problem formulation, methodology, and solutions of the economic issues in CCSs.

Business modeling for database-assisted TV white space networks is regarded as a promising paradigm of dynamic spectrum sharing and can be exploited further to substantially alleviate the spectrum scarcity in future CCS. In the last article, Luo *et al.* present the concept of database-assisted TV white space network architecture where unlicensed devices obtain available spectrum via querying a certified geolocation database instead of performing traditional spectrum sensing. Two types of business models (i.e., spectrum market and information market) are discussed. The performance of both models is also evaluated, indicating that both the database operator's profit and the total network profit can be significantly improved by employing properly designed trading mechanisms.

Before closing Part 1 of this Feature Topic, which mainly covers emerging services and engineering frameworks in CCSs, we would like to thank all submitting authors for considering this Feature Topic as a potential venue for their research work, the reviewers for their high-quality evaluation, and the editorial/publishing team of *IEEE Communication Magazine* for their collaboration. We will be back with the remaining articles as Part 2 of this Feature Topic in the July 2015 issue, which will mainly cover emerging applications in CCS.

BIOGRAPHIES

MUHAMMAD ZEESHAN SHAKIR (muhammad.shakir@qatar.tamu.edu) has been an assistant research scientist at Texas A&M University at Qatar in Doha, Qatar, since July 2012. He received his Ph.D. in electronic and electrical engineering from the University of Strathclyde, Glasgow, United Kingdom, in 2010. From January 2006 to September 2009, he was the joint recipient of an industrial research fund and a prestigious overseas research scholarship by the University of Strathclyde. His research interests include design and deployment of diverse wireless communication systems, including hyper-dense heterogeneous small cell networks. He has published more than 75 technical journal and conference papers, and has contributed to six books, all in reputable venues. He is co-author of two research monographs. Most of his research has been sponsored by Qatar National Research Fund and national industrial partners. He has served as a lead Guest Editor for *IEEE Communication Magazine* and *IEEE Wireless Communications*. He has served as co-chair of several special sessions/workshops and symposia at flagship conferences, such as IEEE ICC and GlobalSIP.

OCTAVIA A. DOBRE is an associate professor with Memorial University, Canada. In 2000 she was the recipient of a Royal Society scholarship in the United Kingdom, and in 2001 she held a Fulbright Fellowship in the United States. Her research interests include cognitive radio systems, spectrum sensing techniques, transceiver optimization algorithms, and dynamic spectrum access. She has published over 130 journal and conference papers in these areas. She is a Senior Editor for *IEEE Communications Letters*, and has served as Editor and Guest Editor for other prestigious IEEE journals. She has been the Co-Chair of technical symposia at flagship conferences such as IEEE ICC and IEEE GLOBECOM.

MUHAMMAD ALI IMRAN is a reader (associate professor) in the Institute for Communication Systems, University of Surrey, United Kingdom. He is leading the physical layer work area for 5G innovation center and is curriculum design leader for the Engineering for Health center in Surrey. He has successfully led international projects encompassing the areas of energy efficiency, fundamental performance limits, sensor networks, and self-organizing cellular networks. He has supervised 20 successful Ph.D. graduates and published more than 150 peer-reviewed research papers. He is an Associate Editor for *IEEE Communications Letters* and the *IET Communications Journal*, and has served as a Guest Editor for other prestigious IEEE/IET journals.

APOSTOLOS (TOLIS) PAPATHANASSIOU is responsible for LTE PHY standardization and 5G technology development activities in the Next Generation and Standards (NGS) division of Intel's Communication and Devices Group (iCDG). He has more than 50 scientific contributions to international journals, conferences, and books, and more than 100 contributions to wireless standardization bodies such as 3GPP and IEEE 802.11/802.16. Previously at Intel, he led multiple standardization efforts in ITU-R and IEEE/WiMAX Forum. Before joining Intel, he worked on multiple-antenna PHY techniques and algorithms for 3G, WiFi, and satellite systems.

ZHONGSHAN ZHANG received his Ph.D. degree in electrical engineering in 2004 from Beijing University of Posts and Telecommunications. From February 2006 to March 2009, he was at the University of Alberta, Canada, as a postdoctoral fellow. He has also worked in DoCoMo Beijing Laboratories, Alcatel-Lucent Shanghai-Bell, and NEC China Laboratories as a researcher. He is currently a professor of the University of Science and Technology Beijing. His main research interests include self-organized networking, cognitive radio, and cooperative communications.

ATHANASIOS V. VASILAKOS is currently a professor with Kuwait University. He has served or is serving as an Editor for many technical journals, such as *IEEE Transactions on Network and Service Management*; *IEEE Transactions on Information Forensics and Security*; *IEEE Transactions on Systems, Man, And Cybernetics—Part B: Cybernetics*; *IEEE Transactions on Information Technology in Biomedicine*; *IEEE Transactions on Computers*; *IEEE Transactions on Cloud Computing*; *ACM Transactions on Autonomous and Adaptive Systems*; and the *IEEE Journal on Selected Areas in Communications*. He is also General Chair of the Council of Computing of the European Alliances for Innovation (www.eai.eu).

HONGGANG WANG is an assistant professor at the University of Massachusetts Dartmouth. His research interests include wireless health, body area networks, cybersecurity, mobile multimedia and cloud, wireless networks and cyberphysical systems, and big data in mHealth. He serves as an Associate Editor of *IEEE Transactions on Big Data*, the *IEEE IoT* (Internet of Things) *Journal*, and the *IEEE Access Journal*, and an Associate Technical Editor of *IEEE Communications Magazine*.

HIROSHI HARADA joined the Communications Research Laboratory in 1995, part of Japan's Ministry of Posts and Communications (currently National Institute of Information and Communication Technology, NICT). Since 1995, he has researched software defined radio, cognitive radio, dynamic spectrum access network, and broadband wireless access systems on the microwave and millimeter-wave band. Currently he is director of the Smart Wireless Laboratory at NICT and has been a visiting professor of the University of Electro-Communications, Tokyo, Japan, since 2005.

Dynamic Spectrum Aggregation for Future 5G Communications

Hanna Bogucka, Paweł Kryszkiewicz, and Adrian Kliks

ABSTRACT

We discuss dynamic spectrum aggregation methods, opportunities, and issues for future 5G communication in cognitive cellular networks. We focus on multicarrier technologies, which have been recognized as capable of flexible utilization of fragmented spectrum opportunities in unlicensed frequency bands, such as enhanced NC-OFDM and NC-FBMC. These techniques can be applied to shape the signal spectrum so that it is concentrated within used frequency bands, and the out-of-band transmitted power and associated interference generated to the incumbent coexisting systems are limited. We present key issues associated with spectrum aggregation that are recognized at the transmitter and the receiver. We also discuss solutions of these issues that can make spectrum aggregation viable for 5G systems.

INTRODUCTION

Constantly increasing demand by mobile devices for higher data rates, multimedia services support, and ever more bandwidth, as well as anticipated traffic of billions of devices constituting the Internet of Things are creating unprecedented challenges for future mobile communication systems. There seems to be general consensus on the future fifth generation (5G) wireless communication to aim at achieving 1000 times the system capacity, 10 times the energy efficiency, data rate, and spectral efficiency, and 25 times the average mobile cell throughput of today's 4G systems. Concerning the capacity and spectrum usage enhancement, they are viewed under the umbrella of network densification, a combination of spatial densification (by increasing the number of nodes and antennas per network node) and spectrum aggregation [1]. Spectrum aggregation refers to making use of possibly discontinuous frequency bands, and thus larger amounts of electromagnetic spectrum. It is known to be possible through a technique called carrier aggregation (CA), which has been proposed for a Long-Term-Evolution Advanced (LTE-A) standard in order to achieve throughput of 1 Gb/s in the downlink for 4G systems in a 20 MHz channel. It increases the usable spectrum by aggregating resource blocks (RBs) either within a given band or in different frequency bands [2]. This aggregation is handled at the medium access layer, and each component carrier uses its own physical layer protocol and hybrid automatic repeat requests. Each component carrier can take any of the transmission bandwidths supported by LTE Release 8, 6, 15, 25, 50, 75, or 100 RBs, corresponding to 1.4, 3, 5, 10, 15, and 20 MHz channel bandwidths, respectively. Two types of CA approaches have been proposed: contiguous and non-contiguous CA. In the contiguous CA approach, the multiple component carriers used are adjacent to each other, which enables the usage of a single inverse fast Fourier transform (IFFT) and FFT module, as well as one RF frontend. This is not the case for the non-contiguous CA approach, which does not require adjacent component carriers. Although CA applied in LTE-A is a step toward spectrum aggregation, its flexibility in aggregating any kind of spectrum fragments (available frequency bands) is limited, because only a limited and integer number of RBs can be used for data communication. Moreover, the proposed protocols do not allow for highly dynamic spectrum access and aggregation. To further expand LTE capacity, integration of unlicensed carriers (unlicensed spectrum) into the LTE system has been proposed, called Unlicensed LTE (U-LTE).

In general, as shown in Fig. 1a, spectrum aggregation can be considered on a macro and a micro scale. In the first case spectrum fragments with a fixed granularity defined by the RBs (e.g. bands I, II, and III in Fig. 1a) are merged together, and can be assigned to one end user as in the LTE-A or multiple-carrier (multiple-cell) high-speed packet access (HSPA) system using multiple physical carriers (HSPA-xC). In such a case, there is no option to utilize a fraction of an RB (predefined bandwidth). As opposed to this, small-scale and more flexible spectrum aggregation is envisioned, where the aggregated fragments of spectrum are interleaved with other incumbent system spectrum. Such a spectrum sharing concept is typical for future cognitive system coex-

The authors are with Poznan University of Technology.

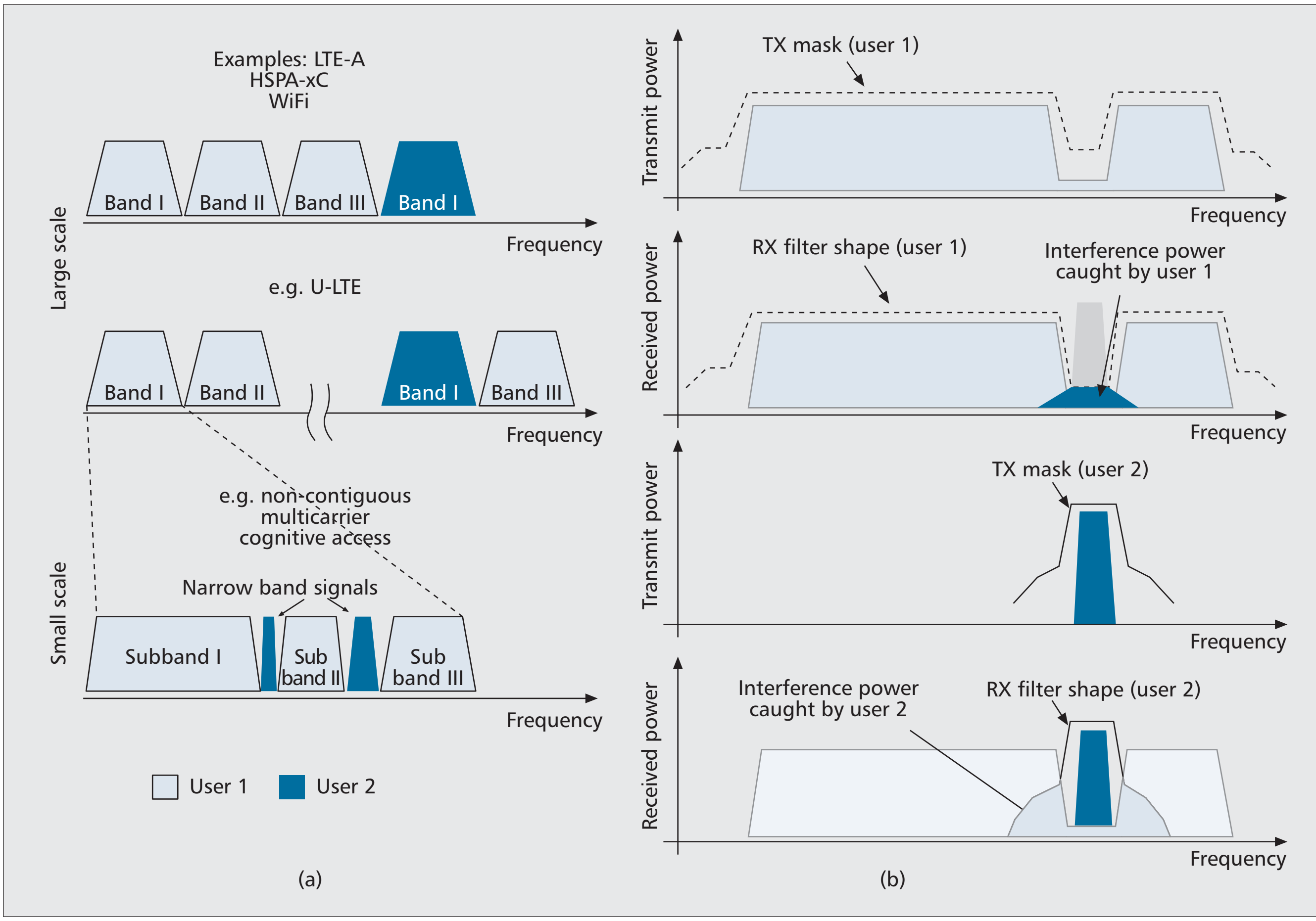


Figure 1. a) Cases of spectrum aggregation; b) example of aggregated spectra at the transmitter (TX) and receiver (RX) of users (1 and 2) of two coexisting systems.

istence, where user 1, transmitting signals within a given spectrum mask, geographically and spectrally neighbors the transmission and reception of user 2. In this case, as shown in Fig. 1b, interference generated to user 2 by user 1 depends on the out-of-band (OOB) power emission of user 1, as well as on the selectivity of user 2's reception filter. Similarly, user 2 can generate interference to user 1. Thus, appropriate measures have to be taken i) to minimize the OOB power of both transmissions, and ii) to possibly shape the signals, taking not just the required spectrum emission masks (SEMs) but also coexisting-system reception-filters characteristics into account.

New multicarrier transmission techniques using non-contiguous subcarriers, such as non-contiguous orthogonal frequency-division multiplexing (NC-OFDM) and non-contiguous filter-bank-based multicarrier (NC-FBMC), are known to be capable of flexible spectrum aggregation [3, 4]. By applying cognitive spectrum sharing using these techniques in both licensed and unlicensed frequency bands of the future heterogeneous networks, more frequency resources can be effectively used, and interference among cells and nodes can be avoided. Cognition here means operational context awareness, and making intelligent decisions (optimized in some sense) based on learning capability on accessing the spectrum (including its aggregation) satisfying dynamically changing SEMs, which are the key elements of the concept called cognitive radio. Dynamic aggregation of potentially non-contiguous fragments of bands in a wide frequency range poses a number of challenges for baseband processing, antenna, and RF front-end transceiver design, particularly in the changing radio environment. Below, we focus on baseband processing to meet these challenges with non-contiguous multicarrier technologies and novel algorithms enhancing spectral efficiency, interference robustness, and reception performance. The presented algorithms support the aforementioned cognitive radio embedded intelligence.

Spectrum Densification with Non-Contiguous Multicarrier Transmission

Let us consider the NC-OFDM and NC-FBMC techniques for their potential in achieving high spectral efficiency through spectrum aggregation in cognitive 5G communications. In an NC-OFDM transceiver, IFFT and FFT algorithms are employed for computationally efficient modulation and demodulation. Subcarriers that are not used due to fragmentation of the available frequency bands correspond to IFFT frequency

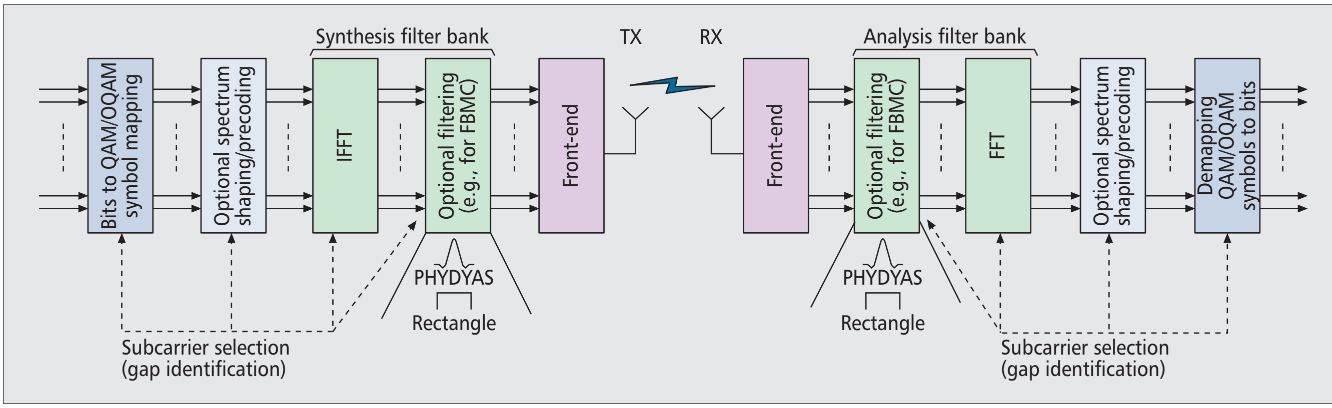


Figure 2. Generalized structure of an NC-OFDM and NC-FBMC system.

bins which are zeroed at the transmitter (TX). (Naturally, the same frequency bins are omitted at the receiver (RX).) Similarly, as in the traditional OFDM, the Cyclic Prefix (CP) is added to the NC-OFDM symbol at the output of the IFFT at the transmitter. In NC-FBMC, apart from IFFT/FFT modulator/demodulator blocks, additionally each subcarrier signal is filtered, which is usually implemented using polyphase filter banks (polyphase decomposition of a prototype filter): the synthesis and analysis filter banks at the TX and RX, respectively, as shown in Fig. 2. In general, FBMC can be optimized in either the time or frequency domain, depending on the target channel dispersion, and is often referred to as non-orthogonal subcarrier modulation due to overlapping of the adjacent symbols in the time domain and subcarriers in the frequency domain. This lack of orthogonality at the transmitter and application of the analysis filter bank at the receiver implies no need for CP, increasing the bandwidth efficiency of an FBMC system.

Both these techniques possess the ability to efficiently use fragmented spectrum opportunities and to suppress interference that may affect coexisting systems in the cognitive radio scenario and environment. To counteract the potential for significant OOB interference resulting from NC-OFDM transmission, several techniques have been proposed in the literature designed to suppress the subcarriers sidelobes and allow for dynamic steering of the sidelobes power suppression level and frequency range [4]. An (NC-)OFDM system applying any kind of OOB-power reduction is called enhanced (NC-)OFDM. On the other hand, NC-FBMC, which applies subcarrier filtering, handles OOB emission at a subcarrier level. This comes at the expense of increased computational complexity [5]. One can envision the OOB level for dynamically changing SEMs being adjustable by adaptively modifying filter characteristics. However, so far, this kind of pulse-shape dynamic adaptation of NC-FBMC is not mature enough for prospective applications. Given the possible constraints of limited computational and energy resources available at user equipment, a practical approach to this problem is needed that achieves a balance between interference mitigation efficiency and its associated costs.

Spectrum Aggregation and Shaping with Enhanced NC-OFDM

NC-OFDM combines the computational simplicity and algorithmic maturity of OFDM with the ability to aggregate fragmented spectrum by dynamic selection of subcarriers, and modulating unused ones with zeroes. This simple approach allows other users of the same cognitive radio system or other systems (not necessarily using carriers orthogonal to the considered NC-OFDM system) to operate in frequency band not subject to the considered NC-OFDM system spectrum aggregation, even in a relatively narrow spectrum notch. However, simply turning off subcarriers within a given band results in the interference generated in this band at the power spectral density (PSD) level of 20–30 dB (relative to in-band PSD), which may not be sufficient for the protection of coexisting systems operating in this band. This is because a single NC-OFDM subcarrier has the spectrum shape characterized by high-power sidelobes.[1]

There are a number of spectrum shaping methods in the literature to reduce NC-OFDM subcarrier-spectrum side-lobes in the OOB region. The most straightforward approach would be to use digital filtering. However, with dynamically changing spectrum availability, dynamic design of filters of high selectivity might be complex. In [4], an overview of the OOB power reduction methods is provided. They span from computationally simple but throughput-decreasing time-domain windowing to computationally complex but spectrally efficient precoding. A method that is particularly efficient in flexible reduction of the OOB power is optimized cancellation carrier selection (OCCS) [6], an extension of the well-known Cancellation Carriers (CCs) algorithm. It has acceptable complexity, adjustable OOB power attenuation, and backward compatibility with OFDM systems. The standard CCs method selects some subcarriers at the edges of available frequency bands to be CCs. These subcarriers are not modulated by regular data symbols, but by optimized complex values. The goal is to minimize the power of summed CCs and data carriers (DCs) in a given OOB frequency range. The CCs can be omitted at the receiver, or processed to improve the reception quality. OCCS is described in [6]. Note

[1] Note that the spectrum representation of a discrete subcarrier at the output of IFFT and digital-to-analog converter of finite sampling frequency is not strictly the Sinc function as for the continuous-time OFDM representation, and reflects the periodic nature of FFT spectrum.

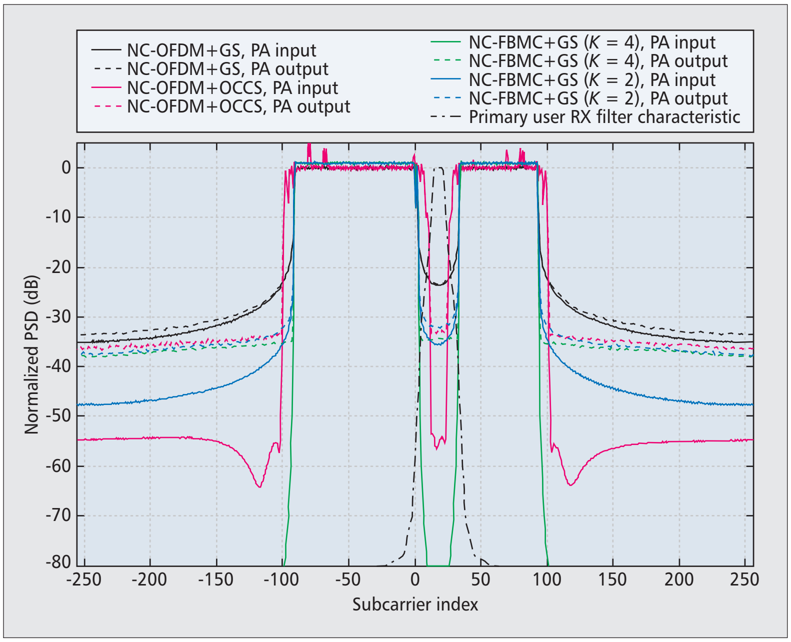


Figure 3. Example results of spectrum aggregation and shaping using NC-OFDM, enhanced NC-OFDM with OCCS, and NC-FBMC with overlapping factor K; QPSK mapping, PA Rapp power model, input backoff parameter IBO = 7 dB, and *smoothness* factor $p = 10$.

that this optimal method does not always select the spectrum-edge subcarriers. Importantly, calculation of CCs' modulating values for a single NC-OFDM symbol boils down to complex-valued matrix-by-vector multiplication before IFFT at the transmitter. The mentioned matrix can be pre-calculated for a given SEM, while the mentioned vector consists of symbols modulating DCs. Example results of spectrum shaping and aggregation are presented in Fig. 3 as the PSDs of the selected quadrature phase shift keying (QPSK)-mapped non-contiguous multicarrier transmission schemes. The black and red solid curves relate to NC-OFDM system using guard (nulled) subcarriers (GSs) and enhanced NC-OFDM using OCCS, respectively. Both compared systems use $M = 151$ DCs, and $L = 35$ subcarriers are used either as GSs or CCs. The order of the IFFT (modulator) of $N = 1024$, and the CP of $N_{CP} = 128$ samples is applied.[2] Note that the PSD in the OOB frequency region can be obtained at the level of –55 dB at the input of the considered power amplifier (PA) in the OCCS case; however, it can be arbitrarily lower depending on the number of subcarriers dedicated to OOB power cancellation.

Spectrum Aggregation with NC-FBMC

Recently considered as the potential successor of the widely applied OFDM, filtered multicarrier transmission realized by means of filterbank structures has been studied intensively for application in future, 5G, and cognitive radio networks in the past decade [5].[3] Although various realizations of this concept can be considered (e.g., complex exponentially modulated filterbanks, cosine-modulated filterbanks, transmultiplexers, perfect reconstruction filterbanks, oversampled filterbanks or modified discrete Fourier transform [DFT] filterbanks), the term filterbank multicarrier (FBMC) transmission has come to refer to schemes where offset-quadrature amplitude modulation (QAM: OQAM) symbols modulate independently shaped subcarriers using dedicated filters [7]. FBMC offers a significant improvement in OOB power reduction compared to conventional OFDM at the expense of higher complexity of the transceiver architecture. This is why this technique is being extensively considered for the future cognitive radio waveform design. An important issue in designing an NC-FBMC system is the selection of the pulse shape implying characteristics of the synthesis and analysis filters. Various pulses have been proposed in the literature, among which the extended Gaussian function together with its special case, Isotropic Orthogonal Transform Algorithm (IOTA), and the so-called PHYDYAS pulse belong to the most promising solutions [8]. A mixture of OFDM and FBMC, called universal filtered multicarrier (UFMC), was also recently proposed, where instead of subcarrier-based processing, the subcarriers are gathered into small groups and shaped jointly [8].

Should the pulses utilized for subcarrier filtering be characterized by good time-frequency localization, they guarantee low OOB emission, thus minimizing the energy induced in the adjacent frequency bands. Such a phenomenon can be observed in Fig. 3 (green solid curve); that is, very steep transition of the PSD curves between the occupied and unoccupied frequency bands can be achieved for the applied PHYDYAS prototype pulse shaping filter with the overlapping factor $K = 4$, outperforming NC-OFDM with OCCS at the PA input.[4] The PSD of the NC-FBMC signal in the OOB frequency region can be as low as –100 dB at the output of the modulator, depending on the applied transmitter filterbank characteristic. Thus, unlike in NC-OFDM, no additional subcarrier processing is required for an NC-FBMC system in order to keep the OOB power at the desired level.

It seems that both NC-OFDM and NC-FBMC have high potential to be applied for 5G signaling capable of fragmented spectrum aggregation. However, the mentioned computational complexity, nonlinear effects in a PA, as well as other practical issues at transmitters and receivers need further discussion.

Practical Issues of Dynamic Spectrum Aggregation

Complexity and Aggregation Dynamics

Energy efficiency is an important target of future cognitive 5G networks, and associated computational complexity becomes an important issue in system design. The complexity of the enhanced NC-OFDM and the NC-FBMC transceiver is drafted in Table 1. It has been estimated based on detailed analysis of the number of real additions and real multiplications required by the baseband algorithms listed in that table.[5] To this end, some assumptions have been made in this approximation: the set of aggregated subcarriers is known at the transmitter and the receiver after a technique called *randezvous* for setting

[2] These parameters are assumed for all other example results presented in this article.

[3] Research on FBMC technology potential for future cognitive radio communication has been funded by a number of EU projects: 5GNow, METIS, EMPhAtiC, PHYDYAS, URANUS, as well as other projects worldwide considered to be driving 5G standardization.

[4] Note that the compared NC-OFDM and NC-FBMC systems have the same rate.

[5] These algorithms have been implemented by the authors at the Laboratory of Wireless Communication, Poznan University of Technology.

	Processing task	Number of real multiplications per symbol	Number of real additions per symbol	Computational complexity (Landau notation)
		Enhanced NC-OFDM		
TX	IFFT modulator	$N \log_2 M - 3N + 4$	$3N \log_2 M - 3_N + 4$	$\mathcal{O}(N \log_2 M)$
TX	Spectrum shaping with OCCS (typically $M >> L$)	$4ML$	$4ML - 2L$	O(ML)
RX	FFT demodulator	$N \log_2 M - 3N + 4$	$3N \log_2 M - 3_N + 4$	$\mathcal{O}(N \log_2 M)$
RX	Channel estimation at RX	$\frac{2M}{G}\left(\frac{1}{\beta}+1\right)$	$\frac{5M}{G}+\frac{M}{\beta G}+2M$	$\mathcal{O}(M)$
RX	Equalization at RX	$4M$	$2M$	$\mathcal{O}(M)$
RX	Synchronization at RX	$4(N + N_{CP})$	$6(N + N_{CP})$	$\mathcal{O}(N)$
Off-line	Calculation of the OCCS coding matrix for a given spectrum mask	$\mathcal{O}(\gamma L^2) + 4ML\gamma^2 + 2ML - 3L + 4LJ + J + 4\gamma L^2 + 4LM\gamma^2 + 4L^2$	$\mathcal{O}(\gamma L^2) + 2ML\gamma^2 + 6ML\gamma - 2L - 3LM + 3\ LJ + 4\gamma L^2 + 2L^2$	$\mathcal{O}(\gamma L^2 + \gamma ML\gamma^2)$
		NC-FBMC		
TX	IFFT modulation with polyphase filtering (assuming real value filters coefficients and small K)	$2\left(N\log_2\left(\frac{M}{2}\right)-3N+4\right)+4MK$	$2\left(3N\log_2\left(\frac{M}{2}\right)-3N+4\right)+MK-M$	$\mathcal{O}(N \log_2 M)$
RX	FFT demodulation with RX polyphase filtering (assuming real value filter coefficients)	$2(N \log_2(M) - 3N + 4) + 4MK$	$2(3N \log_2(M) - 3N + 4) + 4MK - M$	$\mathcal{O}(N \log_2 M)$
RX	Channel estimation at RX; G and β are fixed	$\frac{4M}{G}\left(\frac{1}{\beta}+1\right)$	$\frac{10M}{G}+\frac{2M}{\beta G}+4M$	$\mathcal{O}(M)$
RX	Equalization at RX	$4ML_{eq}$	$4ML_{eq} - 2M$	$\mathcal{O}(ML_{eq})$
RX	Synchronization at RX (based on example presented in [14])	$3N/2$	N/A	$\mathcal{O}(N)$
f-line	TX/RX filter bank design for a given spectrum mask (for PHYDYAS filter)	$N(2K - 1)$	$N(K - 1)$	$\mathcal{O}(NK)$

* M, L: the numbers of data carriers (DCs), cancellation carriers (CCs); γ: the number of OOB spectrum sampling points that are minimized; K: the number of polyphase-filter coefficients; G: the number of NC-OFDM/NC-FBMC symbols between pilot-carrying symbols plus 1; β: the number of pilots in an NC-OFDM/NC-FBMC symbol transporting pilots; L_{eq}: the number of the equalizer taps.

Table 1. Computational complexity of the enhanced NC-OFDM and NC-FBMC TX-RX chain.

the parameters of the connection has been applied (or other configuration-managing technique in a cellular network scenario), the IFFT/FFT is implemented using the split-radix-2 algorithm,[6] typically $M >> L$, the number of coefficients of the polyphase-decomposed filters K is usually small, channel estimation is based on pilots, time and frequency synchronization is based on autocorrelation properties of a CP, and no additional self-interference cancellation algorithm is considered. The Landau notation is used to present the complexity of processes assuming that either multiplication or addition requires just one clock cycle (as in digital signal processors). Note that the OCCS method does not require any additional computations at the

[6] Implementation can be further streamlined by performing IFFT/FFT pruning [9] if the baseband transceiver is constructed on a programmable communication system.

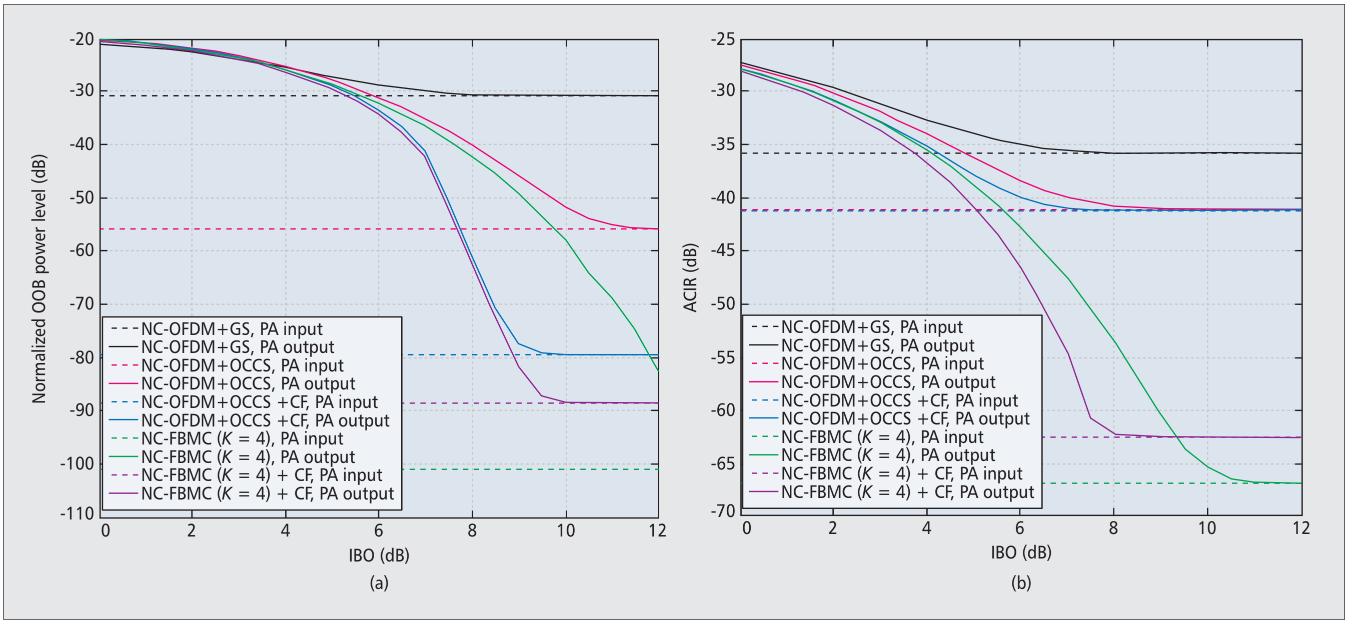


Figure 4. a) Normalized OOB power; b) ACIR for enhanced NC-OFDM and for NC-FBMC applying clipping-and-filtering (CF) method; QPSK mapping, PA Rapp model parameters: $p = 10$, IBO = 10 dB.

receiver. However, if we allow for the advanced reception of the OCCS method, the complex symbols transmitted using CCs can be used to improve reception quality of DCs and require $M(M + L - 1)$ complex additions and $M(M + L)$ complex multiplications [4].

Aggregation of fragments of bands changing positions on the frequency axis, and varying requirements concerning the spectrum masks (for adjacent channels' interference protection) entail additional calculations for the design of the proper spectrum shaping and aggregation algorithm. This calls for estimation of the amount of offline computations required to reshape the spectrum satisfying the upper limit associated with the dynamics of the required SEM changes. These estimates are also shown in Table 1. For enhanced NC-OFDM with OCCS, a redesign of the matrix used for calculation of CCs requires some matrix operations and solving a scalar-based nonlinear equation. Here, the most computationally demanding operation will be the matrix singular value decomposition (SVD). For the case of NC-FBMC, a prototype filter design may be needed if we allow for flexibility in pulse shape selection.

The Power Amplification Issue

The spectrum shaping and aggregation algorithms described above are designed assuming ideal RF front-ends. However, in practical RF front-ends a number of spectrum distorting effects are observed, such as in-phase and quadrature (IQ) signal components amplitude imbalance, local oscillator leakage, nonlinear distortion in PAs, or quantization noise in digital-to-analog converters (DACs). In particular multicarrier techniques are prone to nonlinear distortions of the PAs. This is because of high peak-to-average power ratios (PAPRs) of signals consisting of multiple independently modulated and summed subcarriers. High variations of the signal envelope may cause peak clipping in a nonlinear PA, energy spilling into adjacent frequency bands, and intermodulation effect (i.e., generation of additional spectrum components in frequencies that are linear combination of subcarrier frequencies). This source of OOB radiation is independent from that in the digital domain caused by the used impulse shapes. From the perspective of non-contiguous signal spectrum, the presence of this phenomenon destroys the steepness of the spectral characteristic achieved by either advanced OOB power reduction algorithms (in enhanced NC-OFDM) or per-subcarrier filtering (in NC-FBMC) (dotted-line curves in Fig. 3).

Application of high-class linear PAs in standard end-user mobile transmitters is very costly. Consequently, nonlinear effects are observed at a lower-class PA output, leading to degradation of the aggregated spectrum characteristics. This also means that there is no need for the application of sophisticated and computationally complex spectrum shaping methods in the digital domain since their performance will be significantly reduced in the analog domain. In order to cope with this problem, various solutions leading to either minimization of variations of the multicarrier signal envelope or linearization of the input-output power amplifier characteristics have been proposed. Their application in non-contiguous multicarrier schemes seems to be necessary.

In Fig. 4, the mean normalized PSD value observed in the whole OOB region at the output of the NC-OFDM and NC-FBMC transmitters is presented, as well as the Adjacent Channel Interference Ratio (ACIR, the ratio of the interference power affecting the victim coexisting system receiver to the total power from the source reaching the victim receiver) in NC-OFDM and NC-FBMC systems. These metrics are presented vs. the IBO parameter values for systems possibly applying the known PAPR reduction method called clipping-and-filtering (CF). The example victim system is narrowband, with the spectrum located in the NC-OFDM/NC-FBMC spectrum notch, as shown in Fig. 3.

It is visible in Fig. 4a that for low IBO, all considered systems emit similar OOB power at the PA output, while for a practical value of IBO (e.g., 7dB) there are visible differences between the considered systems, although small between enhanced NC-OFDM and NC-FBMC systems. Note that application of CF slightly increases the OOB power level, which is a well-known effect for all PAPR reduction methods. The plots of ACIR in Fig. 4b evaluate the effective interference power distorting coexisting systems — the effect of the interference power generated to the band of the victim receiver, and of the in-band power of NC-OFDM/NC-FBMC signal leaking to the victim receiver through the reception filter of limited selectivity. It is visible that NC-FBMC schemes outperform enhanced NC-OFDM with OCCS in this regard, as they use some GSs around the coexisting (victim) system band. It is a quite important result showing that even for very deep and steep spectrum shape around the spectrum notch of non-contiguous multicarrier signal spectrum, the effective interference will be limited by the coexisting system receiver characteristic. This has to be taken into account when designing spectrum shaping mechanisms.

Synchronization

Time and frequency synchronization for non-contiguous multicarrier transmission is another challenge for the spectrum-aggregating techniques in cognitive radio. Although a number of synchronization algorithms have been proposed for OFDM systems [10], they may not be suitable for operation of NC-OFDM. First, an algorithm has to be able to adapt to the dynamically changing set of occupied fragments of spectrum (subcarriers). Second, the NC-OFDM or NC-FBMC receiver within its synchronization algorithm must deal with high-power interference (originating from a coexisting system) inside the spectrum gap between the aggregated bands. There are two classical types of synchronization algorithms for multicarrier systems: based on a CP and based on the pre-designed frame-preamble. Their main limitation is lack of awareness of the distinct frequency bands of the useful and interfering signals. The synchronization metrics may treat the interference as white noise of high power, but in the worst case can drive the receiver to synchronize to the interfering signal. Preamble-based methods are more suitable for low signal-to-interference ratio (SIR) and low signal-to-noise ratio (SNR) scenarios; therefore, they are natural candidates for an NC-OFDM cognitive radio system.

A low computationally complex solution is autocorrelation-based, such as the Schmidl & Cox algorithm, which does not require knowledge of the symbols transmitted in the preamble [10]. It finds the peak of the autocorrelation function of the received signal, which is the estimate of time-domain synchronization. The argument of the complex number constituting the peak of autocorrelation function is used for carrier frequency offset (CFO) estimation. However, this algorithm was not designed for robustness against in-band interference, which is caused by the coexisting system operating in the spectrum notch between fragments of aggregated spectrum. The difficulty behind it is that the first stage of this synchronization algorithm is performed in the time domain, where one cannot extract the useful signal from the interference. While it has been shown that wideband interference of a given power can be destructive to synchronization performance, the narrowband interference (NBI) can be much more destructive. A narrowband signal has high values of the autocorrelation function, so "false synchronization" may occur (i.e., an NC-OFDM receiver synchronizes to the NBI).

The simplest approach would be to filter out the incoming interference in the NC-OFDM receiver because the useful signal and interference occupy separate bands. However, this filtering adds computational complexity, and requires computationally complex filter design when the CFO is not known (because this filter has to span the whole FFT band). The most interesting approach is to base synchronization on cross-correlation rather than autocorrelation or on the scheme proposed in [11]. Although it is typically more computationally complex and requires the receiver to know exactly the transmitted preamble, it can provide higher synchronization quality in the low-SNR/SIR regime. Example results are shown in Fig. 5a. A synchronization algorithm based on preambles, received-signal cross-correlation, and coarse and fine time and frequency synchronization steps performs well in the absence (SIR = ∞) and presence of NBI for SIR = 10 dB.

The main differences between the OFDM and FBMC systems from the perspective of synchronization lie in the lack of CP and the presence of self-interference in the latter case. Thus, CP-based synchronization algorithms are excluded from application in NC-FBMC. The problems of reliable symbol timing and CFO calculation in the context of FBMC transmission, either blind or data-aided, have been intensively studied for many years, with pilot-based or preamble-based synchronization considered [12–14]. The problems of dealing with high-power NBI are the same as for NC-OFDM, and it is a big challenge to derive fast and reliable synchronization algorithms for NC-FBMC schemes in the presence of interference from coexisting systems with the spectra located in the notch between used spectrum fragments. Some initial work was presented in [14].

Reception Quality

In some cases, the cost of spectrum aggregation and shaping is in lowering the power for data transmission. This is the case for the OCCS method in which some fraction of allowable power is used by CCs not conveying any information. This causes bit error rate (BER) performance degradation. However, as mentioned above, values modulating CCs can be treated as (pre)coding symbols correlated with the data symbols, and can be used to improve data reception quality. This requires knowledge of the precoding

Time and frequency synchronization for non-contiguous multicarrier transmission is another challenge for spectrum-aggregating techniques in cognitive radio. Although a number of synchronization algorithms have been proposed for OFDM systems [10], they may not be suitable for operation of NC-OFDM.

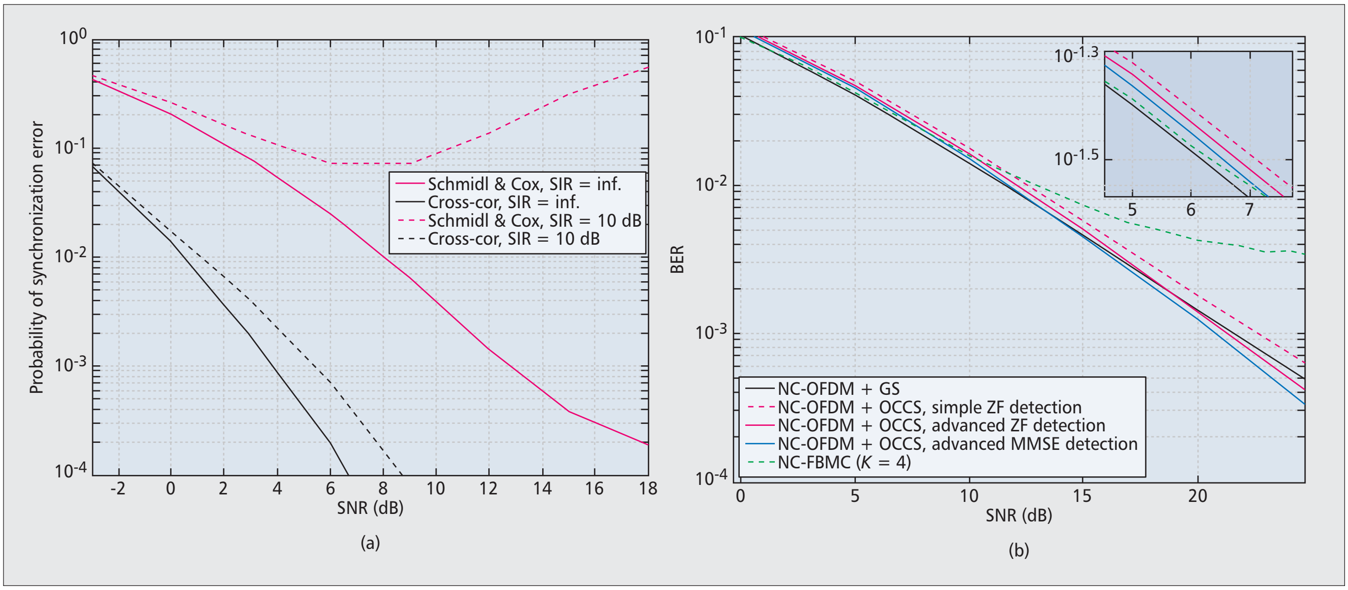


Figure 5. Reception performance: a) probability of synchronization error for NC-OFDM; b) BER for various spectrum aggregating techniques, simple and advanced detection; QPSK mapping.

matrix at the receiver and is associated with additional computations for decoding [4, 7]. Example results of such advanced reception with zero-forcing (ZF) and minimum mean square error (MMSE) detection are presented in Fig. 5b.

The reception quality of FBMC or NC-FBMC cognitive radio systems strongly depends on the prototype filter characteristic and the presence of self-interference among neighboring (on the time-frequency plane) pulses (called *intrinsic* interference). Typically, the error floor exists in the BER plot even for perfectly known channel estimates (example results in Fig. 5b) if there is no dedicated algorithm for channel equalization that deals with the presence of self-interference. This interference has to be removed in order to achieve a reasonable BER [13, 15]. Clearly, this particular aspect increases the reception complexity. Several works propose a dedicated FBMC subcarrier model that considers the interference from the adjacent subbands, as well as advanced algorithms for intrinsic interference management using MMSE or minimum least square error (MLSE). Typically the K-tap equalizer for each subcarrier can offer reliable performance in such a case. Furthermore, sophisticated techniques for channel estimation and equalization have been proposed; these include the application of special preambles, scattered pilots, and iterative decoding schemes. One of the solutions worth mentioning assumes the presence of the auxiliary pilots, which are introduced mainly for interference minimization [12]. Although a number of advanced solutions have already been discussed, the development of effective yet low-complexity algorithms for self-interference cancellation, channel estimation, and equalization for NC-FBMC is still a challenge.

Conclusions and Future Work

The enhanced NC-OFDM and NC-FBMC techniques are suitable candidates for future 5G cognitive communications capable of dynamic and flexible aggregate fragmented spectrum opportunities. In this article we have summarized their potential as well as challenging issues beyond traditionally considered spectrum access and shaping. These issues touch upon aggregation dynamics, nonlinear effects in the transmission chain, receiver synchronization, and performance. Above, we have shown how they can be achieved by novel baseband algorithms at the cost of increased complexity (i.e., flexible and adaptive transmission accounting for joint PAPR and OOB power reduction, as well as interference-robust reception and synchronization algorithms). Some of these algorithms have been presented above; however, there is still a lot of room for further improvements that need to be implemented before 5G networks are deployed. Specifically, future research should concentrate on increasing flexibility of the RF front-end design (including antennas) to operate in multiple bands. Lowering the ACIR floor at the output of a PA is another challenge. Moreover, we believe that the problem of handling the aggregated interference resulting from multiple cognitive radio transmissions requires novel algorithms and solutions.

Acknowledgments

The presented work has been supported by the National Science Centre in Poland within the PRELUDIUM project no. 2012/05/N/ST7/00164, and by the Polish Ministry of Science and Higher Education funds for Status Activity (DS) project on "Cognitive Radio Systems".

References

[1] N. Bhushan *et al.*, "Network Densification: The Dominant Theme for Wireless Evolution into 5G," *IEEE Commun. Mag.*, Feb. 2014, vol. 52, no. 2, pp. 82–89.
[2] S. Parkvall, A. Furuskar, and E. Dahlman, "Evolution of LTE toward IMT-Advanced," *IEEE Commun. Mag.*, vol. 49, no. 2, Feb. 2011.
[3] H. Bogucka *et al.*, "Spectrally Agile Multicarrier Waveforms for Opportunistic Wireless Access," *IEEE Commun. Mag.*, 2011, vol. 49, no. 6, pp. 108–15.

[4] P. Kryszkiewicz, H. Bogucka, and A. M. Wyglinski, "Protection of Primary Users in Dynamically Varying Radio Environment: Practical Solutions and Challenges," *EURASIP J. Wireless Comm. Net.*, vol. 2012:23, no. 1, pp. 1–20.
[5] B. Farhang-Boroujeny, "OFDM versus Filter Bank Multicarrier," *IEEE Sig. Processing Mag.*, vol. 28, no. 3, pp. 92–112.
[6] P. Kryszkiewicz and H. Bogucka, "Out-of-Band Power Reduction in NC-OFDM with Optimized Cancellation Carriers Selection," *IEEE Commun. Lett.*, vol. 17, no. 10, Oct. 2013, pp. 1901–04.
[7] P. Siohan, C. Siclet, and N. Lacaille, "Analysis and Design of OFDM/OQAM Systems Based on Filterbank Theory," *IEEE Trans. Sig. Pro*cessing, vol. 50, no. 5, May 2002, pp. 1170–83.
[8] V. Vakilian *et al.*, "Universal-Filtered Multi-Carrier Technique for Wireless Systems Beyond LTE," *IEEE GLOBECOM 2013*, 9–13 Dec. 2013, pp. 223–28.
[9] R. Rajbanshi *et al.*, "An Efficient Implementation of NC-OFDM Transceivers for Cognitive Radios," *Proc. Int'l. Conf. Cognitive Radio Oriented Wireless Networks and Commun.*, June 2006.
[10] M. Morelli, C.-C. Kuo, and M.-O. Pun, "Synchronization Techniques for Orthogonal Frequency Division Multiple Access (OFDMA): A Tutorial Review," *Proc. IEEE*, vol. 95, no. 7, 2007, pp. 1394–1427.
[11] H. Abdzadeh-Ziabari and M. Shayesteh, "Robust Timing and Frequency Synchronization for OFDM Systems," *IEEE Trans. Vehic. Tech.*, vol. 60, no. 8, 2011, pp. 3646–56.
[12] T. H. Stitz *et al.*, "Pilot-Based Synchronization and Equalization in Filter Bank Multicarrier Communications," *EURASIP J. Advances in Signal Processing*, vol. 2010, 2010, pp. 1–19.
[13] T. Fusco, A. Petrella, and M. Tanda, "Data-Aided Symbol timing and CFO Synchronization for Filter Bank Multicarrier Systems," *Ieee Trans. Wireless Commun.*, vol. 8, no. 5, May 2009, pp. 2705–15.
[14] C. Thein, M. Fuhrwerk, and J. Peissig, "About the Use of Different Processing Domains for Synchronization in Non-Contiguous FBMC Systems," *IEEE 24th Int'l. Symp. Personal Indoor and Mobile Radio Commun.*, 2013, 8–11 Sept. 2013, pp. 791–95.
[15] S. Rahimi and B. Champagne, "Joint Channel and Frequency Offset Estimation for Oversampled Perfect Reconstruction Filter Bank Transceivers," *IEEE Trans. Commun.*, vol. 62, no. 6, June 2014, pp. 2009–21.

Biographies

HANNA BOGUCKA (hbogucka@et.put.poznan.pl) received M.Sc. and Ph.D. degrees in telecommunications from Poznan University of Technology (PUT), Poland, in 1988 and 1995, respectively. Since 1988 she has been employed at PUT, currently in the Chair of Wireless Communications as a full professor with the Faculty of Electronics and Telecommunication. She has research interests in the area of wireless communications, cognitive radio, and green communication systems. She is the author of around 150 papers, and five monographs and handbooks in the area of radio communications.

PAWEL KRYSZKIEWICZ (pkrysz@et.put.poznan.pl) received his M.Sc. degree (with honors) in telecommunications from PUT in 2010. Since 2010 he has been employed under the Chair of Wireless Communication, PUT. Currently, he is an assistant and a PhD student. His research interests are in wireless communications, signal processing, and software defined radio. In particular, he is interested in multicarrier PHY layer design for 5G systems.

ADRIAN KLIKS (akliks@et.put.poznan.pl) received M.Sc. and Ph.D. degrees with honors in telecommunications from PUT in 2005 and 2011, respectively, and currently is employed as an assistant professor under the Chair of Wireless Communications, PUT. He has been involved in various industrial and international research projects. His scientific interests include advanced multicarrier communications, waveform design, small cells in heterogeneous networks, traffic offloading, and cognitive radio.

Lowering the ACIR floor at the output of a PA is another challenge. Moreover, we believe that the problem of handling the aggregated interference resulting from multiple cognitive radio transmissions requires novel algorithms and solutions.

Large-Scale Cognitive Cellular Systems: Resource Management Overview

Mohsen Guizani, Bassem Khalfi, Mahdi Ben Ghorbel, and Bechir Hamdaoui

Mohsen Guizani and Mahdi Ben Ghorbel are with Qatar University.

Bassem Khalfi and Bechir Hamdaoui are with Oregon State Univeristy.

ABSTRACT

This article presents recent advancements in resource management for large-scale DSA systems. Although the problem of spectrum and power allocation is well addressed in the literature, the need for more efficient algorithms still persists due to the exponential growth of the number of wireless devices. Thus, developing efficient distributed approaches has become an attractive solution that can follow the systems' rapid growth. Despite the number of economic-driven methods that have been presented, such as game theoretic solutions, these methods still rely on excessive information exchange, which results in high delays. Inspired by the success of behavioral techniques, mainly learning and filtering approaches, applications of these techniques to spectrum management have attracted more interest due to their distributivity and minimal requirements of information exchange.

INTRODUCTION

The unprecedented growth witnessed by wireless communication systems over the past decade has brought to light new concerns about the spectrum capability to handle the dramatic increase in the number of wireless devices. This was supported by a common belief that the spectrum has become overcrowded or is even running out of space. However, various spectrum measurement studies have shown that the problem is mainly due to lack of efficiency in spectrum utilization rather than scarcity of resources. Therefore, it is anticipated that by enhancing the awareness of wireless terminals about their surrounding environments, spectrum efficiency could be enhanced remarkably. This is the core idea of cognitive radio (CR), which has emerged as a potential candidate for enabling dynamic spectrum access (DSA) [1].

Conventionally, the different portions of the spectrum are allocated in a static manner to specified applications regardless of the activity of users within that system. With the current growth in mobile communication systems, a spectrum shortage is expected. However, in some applications, the spectrum is heavily underutilized. The DSA will allow opportunistically taking advantage of the temporary unused portions left in these applications to cover up the shortage of spectrum resources in other applications (e.g., cellular communication systems) or even to allow the coexistence of other applications. Hence, within the same portion of the spectrum, two systems with two levels of priority can coexist: a primary system (PS), with the highest priority to access the spectrum, and a secondary system (SS), with a lower level of priority to access the spectrum under certain constraints to protect the primary system privileges.

The main objective with DSA systems is to optimize the performance of the SS while protecting the quality of service of the PS. To achieve this, the key component to enable efficient DSA is dynamic spectrum management. It is responsible for:

- The awareness of the activity of the PS users
- The efficient spectrum and power allocation with regard to the target quality of service (QoS) at the SS users

This requires that the SS be self-aware, self-adaptive, and self-configurable.

Dynamic spectrum management encompasses four tasks, which are spectrum sensing, spectrum decision, spectrum sharing, and spectrum mobility. While spectrum sensing and decision are devoted to the identification of the portions of the spectrum that the SS will access with regard to various factors, spectrum sharing and mobility are dedicated to the access and seamless communication in the selected portions of the spectrum. Although these tasks have been intensively investigated in the literature [2], the need for efficient algorithms still persists with the growing density of users.

This article provides an overview of DSA systems, the dynamic spectrum management task, and the current emerging challenges related to resource allocation in DSA systems. Particularly, we focus on the developed algorithms for spectrum assignment and power allocation regarding the growing size of the system. An overview of the DSA system and spectrum management functionality is provided. Then we present a classification of the proposed distributed resource

 0163-6804/15/$25.00

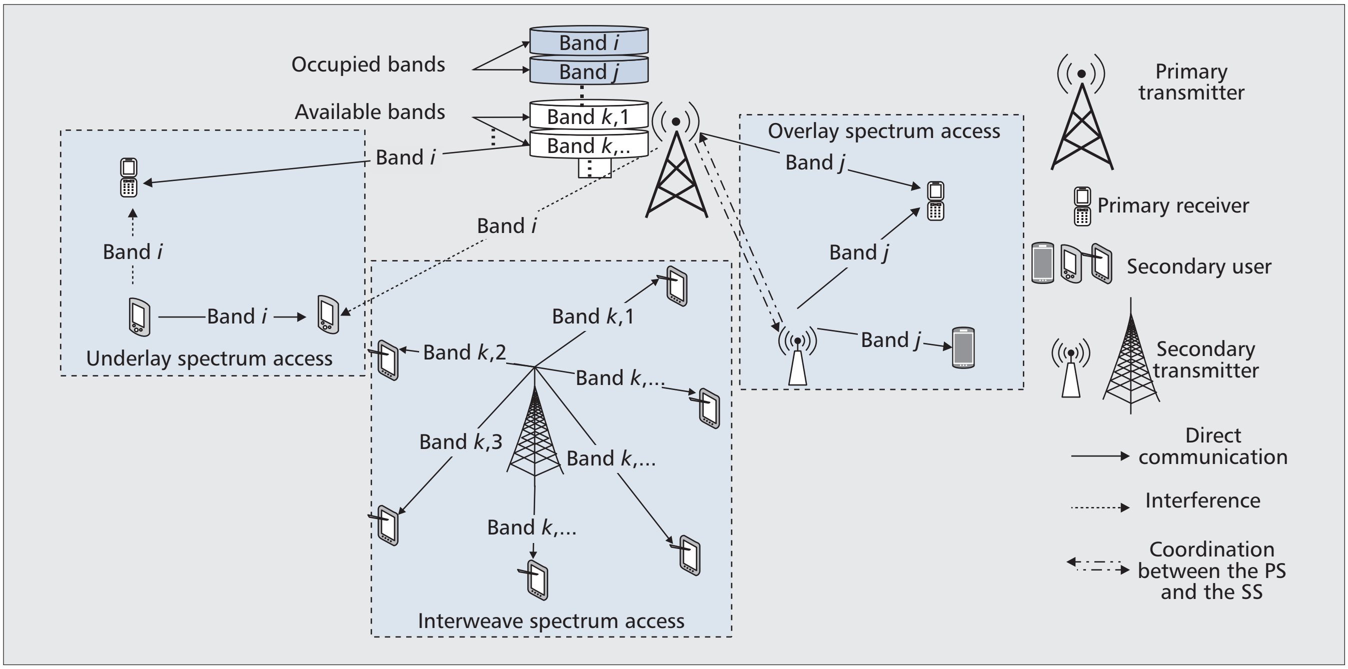


Figure 1. A large-scale DSA system consisting of a primary system with different spectrum sharing paradigms.

allocation methods along with the emerging challenges pertaining to scalability as well as spectrum and power efficiency.

Dynamic Spectrum Access Overview

DSA is envisioned to be the future of wireless communication networks. It will enable the coexistence of an intelligent license-exempt system (the SS) alongside the already existing licensed system (the PS). Users within the SS are known as secondary users (SUs), cognitive users (CUs), or DSA agents, while the users within the PS are referred to as primary users (PUs). The PUs are allowed to access the spectrum whenever they want without being aware of the presence of the SUs. However, the SUs should take into account the activity of the PUs. To do so, an SU should have adequate capabilities of being self-aware and self-adaptive to efficiently exploit the available opportunities in time, frequency, and space.

Based on how an SS behaves in the presence of a PS, three access paradigms, as shown in Fig. 1, have been intensively discussed in the literature [3].

Spectrum interweave: In this scenario, the SU avoids harming the PS with interference by only transmitting over a band if it is unused. No collaboration in this scenario is needed. This access model has attracted a lot of interest from the research community as well as regulators and standardization communities, and even found its way into practical implementation in different standards (802.22 for WRAN, 802.11af, etc.).

Spectrum overlay: In this scenario, there is a close collaboration between the PS and the SS, where the SS will help the PS transmit its data and at the same time profit from it by transmitting its own data while mitigating the interference that may bother the PS communication. However, security issues may arise with the adoption of this model since sharing the codebooks or even whole messages of the PS is required.

Spectrum underlay: This paradigm is known as interference control. The SS uses a transmit power control to keep the resultant interference below a given threshold that guarantees a minimum QoS for the PS. Although in this scenario the SS could access the spectrum at any time using any band anywhere, the challenge is how to accurately estimate the channel between the PS and the SS to limit the resultant interference.

In all the proposed models, the more awareness the SS acquires about the PS, the better it achieves its targeted throughput. On the other hand, this results in more complex problems to be solved in addition to the security issues that arise due to this information sharing. In this article, we focus mostly on the spectrum interweave scenario as it is less harmful to the PS and needs less collaboration from the PUs, as shown in Figs. 1 and 2. Hence, a DSA will refer to the cognitive system using this model. The mechanism that will allow the SS to identify the spectrum opportunities, the coordination between the different users, and the seamlessness of the communication is dynamic spectrum management. In what follows, we address the spectrum management functionality and emphasize its importance in DSA.

Resource Management for DSA Systems

Spectrum Management in DSA

Spectrum management represents the core of DSA. It encompasses all the tasks that will allow DSA users to identify the portions of the spectrum left temporarily by the PS, usually called spectrum holes. Furthermore, it controls the

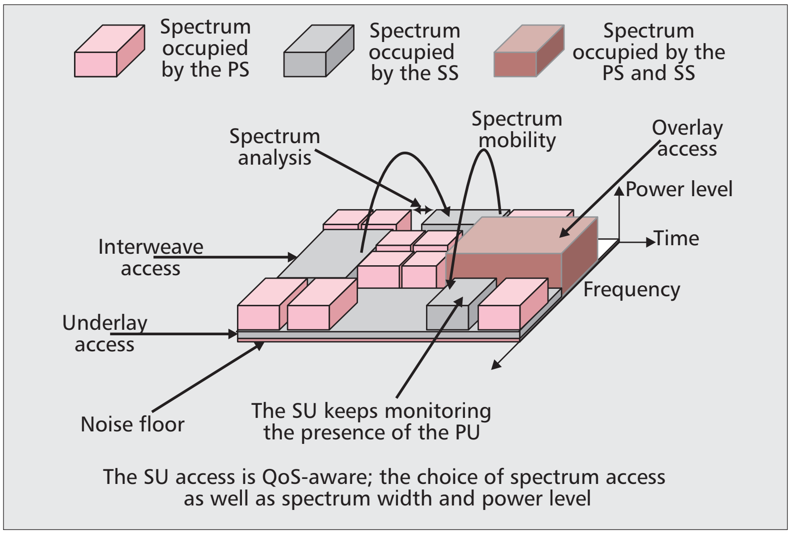


Figure 2. Spectrum management protocols for DSA systems.

access of these bands by different users of the DSA system as well as the coordination between them to avoid any collisions. Moreover, it takes into account the required QoS at the DSA agents and ensures their seamless communication. It could be performed by either a central unit called a central spectrum manager (CSM) or distributively where each user will play the role of the spectrum manager.

There are four vital tasks in spectrum management: spectrum sensing, spectrum decision, spectrum sharing, and spectrum mobility.

In spectrum sensing, the DSA system senses the different portions of the spectrum to detect whether there is PS activity or not. Various methods have been proposed in the literature and classified over various criteria, including individual or cooperative, simple or sophisticated, centralized or distributed, and single band or multiband [4, 5].

In the spectrum decision, based on the measurements made in spectrum sensing, the DSA system decides on the occupancy of the spectrum. This is the most critical task in spectrum management as it could result in a false alarm or misdetection. In fact, if the DSA declares the band vacant while a PU is present and active, mutual interference will harm both systems. In the other case, if there is no primary activity but the SS declares the band occupied, one might miss the opportunity to take advantage of that available slot, which would reduce the system's performance in terms of spectrum efficiency.

In spectrum sharing, the task consists of controlling access to the spectrum and coordinating between the different SUs while accessing the spectrum [6]. The importance of spectrum sharing is the efficient assignment of spectrum among the users. The objective is to minimize the interference between users and to take into account their requirements in terms of QoS.

The final task is spectrum mobility, which is aimed at supervising the activity of the PS while the SS is accessing the spectrum. In fact, since the PS has priority to access the spectrum, it is the responsibility of the SS to detect when a PU is back to using the channel. In such a situation, the SS must switch to another band, stop, or reduce the communication in that band in order to give priority to PUs.

In Fig. 2, we present the different spectrum management phases for a DSA system, depending on the used CR paradigm.

Spectrum Management Challenges in Large-Scale Systems

Conventionally, the spectrum sharing and mobility tasks have attracted less research interest than spectrum sensing and decision in DSA systems. This is because it was thought that once the spectrum was declared unused, conventional access techniques could be employed. However, particular attention has recently been paid to spectrum and resource allocation, driven by the exponentially increased number of wireless devices. Therefore, large-scale or dense DSA systems are more likely to be formed. Such systems will be characterized by a very high number of users, but on the other hand a very limited number of available bands. This requires the development of efficient algorithms that account for system scalability. Hence, sharing the bands is a required solution but controlling the resultant interference represents a real challenge.

The spectrum management process with regard to the large-scale component is anticipated to be different. Conventional spectrum management processes are performed in a centralized manner. However, relying on a CSM makes the resource allocation problem computationally complex and difficult to solve in a relatively acceptable timeframe for practical implementations. The authors in [7] showed that the problem of spectrum and modulation allocation is a nonlinear integer programming problem. Hence, the solution is to use a heuristic algorithm that allows the achievement of a close to optimal solution with polynomial complexity. By doing so, the processing delay is minimized, but the problem of overhead still persists. This may cause an unacceptable delay, which affects the scalability of the system. Decentralized resource allocation is more attractive from this perspective. The decision would be made locally by each user by exchanging some information with the DSA system.

Moreover, the heterogeneity of users and services in terms of QoS requirements combined with the scarcity of resources adds one more component to the resource allocation problem [8]. The DSA system should be QoS-aware such that the DSA agents needing higher levels of QoS are prioritized in the allocation of the "best" bands. On the contrary, a DSA agent with low QoS requirements is accommodated by bands with weaker channel gains. On the other hand, although the detection of a PU activity is part of the spectrum sensing and decision tasks, it is important to consider that in the process of resource allocation. In fact, if an SU with high QoS requirements is transmitting in a band and the PU takes it back, the SU should be switched to another band that allows it to achieve its required QoS level.

Power consumption is another critical resource in wireless communications that has been attracting continuous attention since the emergence of the concept of green communications [9]. Power allocation should be optimized for economic and environmental concerns. Hence, the power should be optimized locally such that it allows the target throughput to be reached without inducing unacceptable interference. This helps reduce the effect on the environment as well as the power cost of the global system.

In the following sections, we provide a classification of the already proposed methods that have addressed the problem of resource management in DSA.

Efficient Resource Management

The problem of resource management is the most critical task in wireless communication systems in general and in DSA systems in particular. At this level, we denote by spectrum management only the two tasks of spectrum sharing and spectrum mobility. Broadly speaking, these methods could be classified into centralized and distributed resource spectrum management techniques from a network architecture perspective.

Centralized Approaches

Based on the primary user behavior, centralized approaches fit the case where the PS activities in the channel do not change constantly. This is the case when accessing the TV white bands, also called spectrum holes, in IEEE 802.22. To do so, after deciding on the available bandwidth, the CSM devotes a part of the spectrum to each DSA agent. For a successful resource allocation scheme, the CSM should perform the following:

- The CSM acquires full knowledge of all the channel gains between the different users and their target throughput as well as their constraints. This could be done using a dedicated control channel.
- The CSM coordinates between the different users to avoid them harming each other with interference.
- The CSM controls the PS activity to keep the SUs enjoying transmitting without sudden interruption.
- The CSM should employ an intelligent and efficient algorithm that succeeds to allocate resources to the different users within the system in a relatively short time.
- There should be a backup scenario in case of failure or immediate evacuation of the channel.

Optimal Approaches for Resource Allocation — In optimal resource allocation, the CSM succeeds in efficiently allocating the spectrum and power for each user. This could be done for relatively small networks where the variability of the channels within the system is small compared to the data processing at the CSM. Integer nonlinear programming tools are used for solving the joint spectrum and bit allocation in DSA systems [10]. However, if the network population increases, the algorithm suffers from the exponential increase of computation time. Hence, suboptimal approaches become more attractive.

Heuristic Approaches to Resource Allocation — Due to the inconvenience of optimal approaches for large-scale systems with regard to computation time, heuristic approaches are more attractive. In [7], the authors considered evolutionary algorithms for an orthogonal frequency-division multiplexing (OFDM) DSA system when trying to minimize the total power consumption. The authors considered a system where the available portions of the spectrum are accessed using OFDM. The CSM jointly allocates the subcarriers and bit modulation among the different users. The PU activity was also taken into account, where at each time episode some bands may be occupied with the PU and leave in the next episode. This will motivate us to consider reallocation at each episode.

Since the optimal problem is NP-hard, the authors employed evolutionary algorithms to solve the problem suboptimally. The Genetic Algorithm (GA), considered to be a heuristic algorithm, was employed by taking advantage of the mutation and crossover to prevent the local optima and thus reach the global optimal in a finite number of iterations. A second method applied to solve the allocation problem is the "ant colony" optimization. This is inspired by the way ants find the optimal path between food and their nest where the shortest path is identified based on an amount of hormone. When applied to DSA, the vertices model the bands, while the edges represent the combination between the users and the modulation levels. Thus, modeling the fact that at each subcarrier different modulation schemes could be used, at each vertex different edges are issued.

The ant colony optimization was shown to achieve near-optimal performance in terms of power saving compared to the optimal solution and better performance when compared to the genetic algorithm.

Decentralized Approaches

Although heuristic approaches could solve the computational complexity in a centralized resource allocation problem, the whole system suffers from scalability issues. In fact, the system is subject to a large amount of signaling overhead to ensure the coordination and control of the communication between the different users. This results in a high delay, which is not in line with the real-time characteristic of DSA (it was envisioned to exploit the instantaneous non-used portions of the spectrum) [11]. Although the use of a hierarchical structure (clustering) may improve the scalability issue of the centralized approach, the need for a more efficient structure that supports real-time implementation is more pressing. Relying on a distributed DSA structure was seen as the solution to the scalability issue by taking advantage of more cognition.

The employed methods for resource allocation in DSA in a decentralized manner are mainly based on economic or heuristic approaches.

Economic-Based Approaches — Although economic approaches could be considered to be centralized approaches [12], they have attracted more attention in solving distributed resource management problems. The authors in [11] pre-

In optimal resource allocation, the CSM succeeds in efficiently allocating the spectrum and power for each user. This could be done for relatively small-networks where the variability of the channels within the system is small compared to the data processing at the CSM.

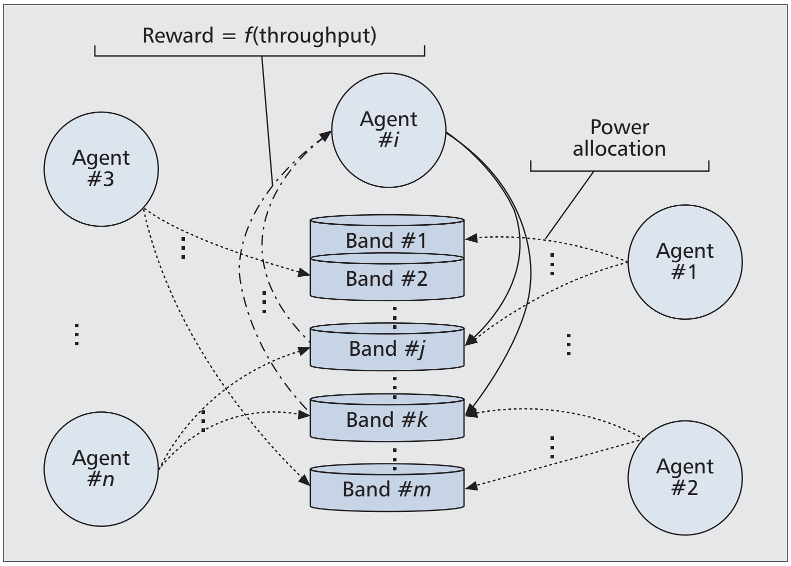


Figure 3. Q-learning for large-scale DSA.

sented game theory as a powerful tool to analytically deal with decentralized spectrum allocation. It is a class of optimization techniques useful for solving optimization problems with conflicting resources. The basic concept is to find the best strategy for each player in order to maximize a given payoff. When considering DSA systems, it could be employed at different levels of the network where at each time the players could model a particular entity of the network [12]. Game theory could model the negotiation or the cooperation between the decision makers to solve the problem of resource management at the highest level between the SS and the PS. In the context of a large DSA system, the players could represent the DSA agents (the pair of a transmitter and its corresponding receiver) in their competition between each other for the available resource or even the competition when accessing the spectrum using the underlay approach with the PUs. The strategy represents the set of bands that will be selected, while the payoff may differ according to the context but could be taken as the achieved throughput or the energy efficiency in some contexts.

The major limitation of game theoretic approaches is that an exchange of information between users is still needed, even if it is minimal. For instance, each user should know the cost function (payoff) of the other players, which is, from a practical perspective, very difficult to acquire unless a control channel is dedicated to this information exchange. Hence, we focus next on approaches that need limited feedback from the network.

Heuristic-Based Approaches: Learning- and filtering-based approaches have great potential for enabling efficient resource management in large-scale DSA systems.

Reinforcement Learning — Reinforcement learning could be formulated as a class of Markov decision problems where a given number of agents try to learn from their environment based on the actions they take. The agent, based on the set of actions that could be taken, visits a given number of states and computes the associated reward. The objective at the end is to select the appropriate action that maximizes the accumulated reward. This accumulated value of reward is the combination of the past accumulated rewards (exploitation) and the future received reward (exploration).

When applied in the context of spectrum management in large-scale DSA, each DSA agent selects the best channel bands that allow it to receive the highest accumulated reward. For instance, Q-learning has been applied to promote efficient spectrum allocation targeting the maximization of the per-agent reward [13]. Specifically, in these attempts, efficient objective functions have been designed, depending on the traffic model, and shown to achieve high performance in terms of scalability and learnability in addition to guaranteeing a distributed implementation. In Fig. 3, we detail the main steps of the Q-learning method when employed for resource allocation in DSA systems.

Although this method is very simple to apply with limited complexity, it may suffer from some inconvenience with fast time-varying channels. In fact, for a given channel state, the performance of the learning could be efficient if all the states have been explored. However, if the channels vary rapidly with time, this will deteriorate the learnability of the approach as the perceived reward for a given state will change with time.

Algorithm	Distributive	Optimal	Complexity	Comments
Integer linear programming [10]	No	Optimal	Exponential	Discrete power allocation
Evolutionary algorithms [7]	No	Suboptimal	Linear	
Extended Kalman filtering [14]	No	Close-optimal	Linear	Valid for Gaussian noise
Game-theoretic methods [11]	Yes	Optimal	Linear	Cooperation needed
Q-learning [13]	Yes	Close-optimal	Linear	Performance's degradation with rapid changing environments
Particle filtering [15]	Yes	Close-optimal	Linear	Complexity proportional to number of particles

Table 1. Characteristics of spectrum management approaches for DSA systems.

Hence, this method is very effective with quasi-static or slow varying channels only.

Filtering — Filtering-based algorithms could be very useful for addressing the limitations pertaining to the varying nature of the channels thanks to their tracking capabilities. In [14], extended Kalman filtering (EKF), which is a derivative of Kalman filtering adapted to nonlinear scenarios, has been applied to the tracking of the transmission power and channel variations distributively. However, in general, Kalman filtering and its derivatives fail to reach acceptable performance, especially with non-Gaussian noise. The estimation may suffer from large biases, convergence, or lack of robustness. Particle filtering (PF) emerges as a key candidate to overcome these deficiencies.

The PF algorithm includes three main steps, as shown in Fig. 4:
- Generating the particles according to a given importance density
- Weighing the particles using current observations and the previous weights
- Resampling to avoid particle degeneracy, making the particles with large weights more dominant than the ones with small weights

Like Kalman-based filters, two model equations are needed for PF:
- A state evolution equation to characterize the time evolution of the state. This could be the prediction equation of the channel allocation given the previous allocation from the channel fading temporal correlation.
- An observation equation that relates the observation to the state and will be useful in the weighing and resampling phases of the filtering algorithm.

PF was applied for distributed resources management for DSA in [15] where each particle represents a possible channel allocation among the users. Hence, each user generates a given number of particles according to a given importance density. The prediction equation of the channel allocation could be derived using the previous allocation and exploiting the channel fading temporal correlation. For the observation equation, it could model the objective function that each user wants to maximize. For a general setup, joint multiband selection and power allocation could be employed along with the consideration of the PU's activity.

Summary and Comparison

We summarize in Table 1 the main performance characteristics of the different algorithms.

In Fig. 5, we investigate the achieved reward compared to the case when Q-learning is employed. This figure shows the efficiency of the filter-based approach to follow channel variations and better allocate channels to optimize the performance in a changing environment.

Future Research Directions

Even though spectrum management in large-scale DSA systems is a well investigated subject, efficient spectrum assignment and power allocation are still needed.

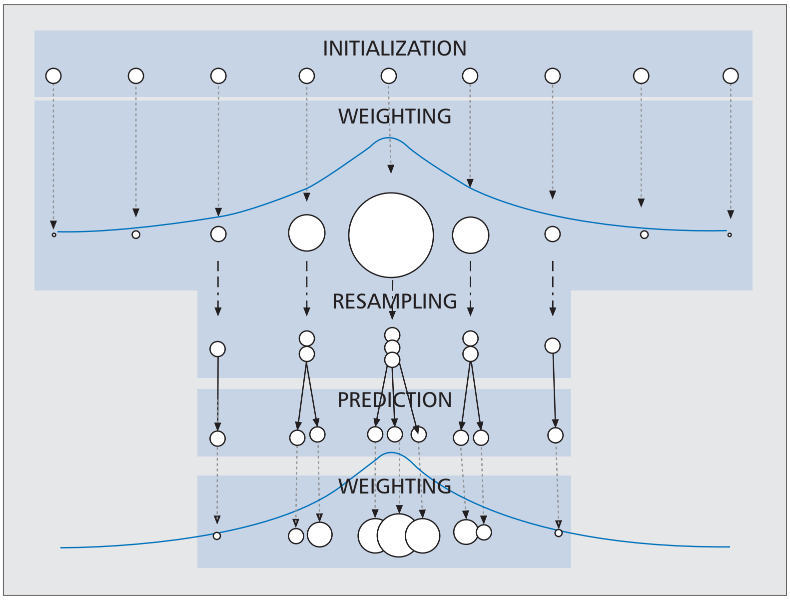


Figure 4. Particle filter approach for large-scale DSA.

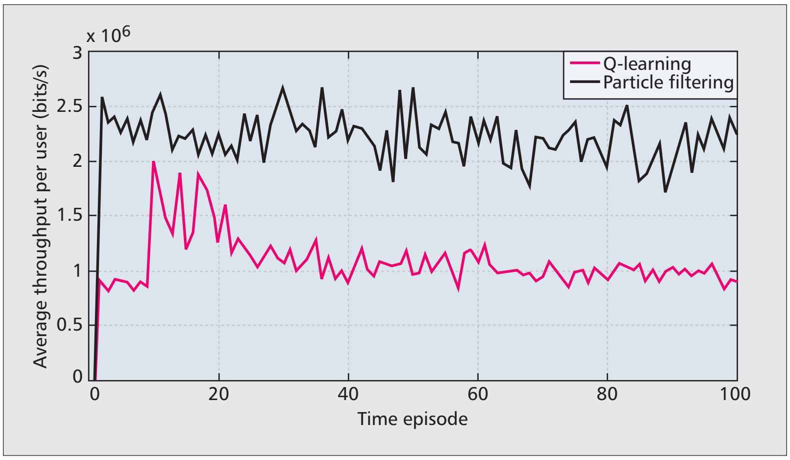


Figure 5. The achieved per-agent throughput reward with multiband selection, power allocation, and primary user activity: particle filtering vs. Q-learning. The system parameters are $n = 100$ users, $m = 11$ bands, $N = 20$ particles, bandwidth 6 MHz, PU's activity in 10 percent of the bands. Each user may select up to two bands.

It is very important to consider a holistic framework where all the components of the DSA system are taken into consideration, which, in spite of the ever growing literature discussing DSA, is still missing.

In the following, we present some of the directions that need further research investigation in order to enhance the performance of resource management of DSA systems.

Traffic prediction models: The system should use traffic prediction models to anticipate the network traffic. In particular, predicting the PUs' activity is an important piece of information for the resource management to have advanced knowledge of the channels' occupancy.

Users' interactions in a distributed model: Even though a distributed model is adopted, a resource allocation system may require the presence of a permanent channel to ensure commu-

In distributed approaches, heuristic methods based on behavioral evolution of resource allocation reward represent a promising solution due to their ease of implementation and low requirements in terms of information exchange and ability to track environment changes.

nication success and prevent other users from interfering with the PS. Some issues about the identification of the opportunities in the spectrum have already been addressed using cooperative communication (the trade-off between sensing complexity and sensing time reporting should be carefully addressed).

Channel variability: Channel variability represents a real challenge. Redoing the allocation at each channel variation may not be efficient, especially in fast changing environments. Tracking techniques offer a faster adaptation capability, but modeling channel variations is a persistent problem. Developing efficient prediction models for channel variations will allow resource allocation to be enhanced through better prediction of suitable resource allocations. It can be employed in either learning or filtering techniques.

QoS awareness: Spectrum resource assignment should not only account for the user's absolute satisfaction. If a user gets resources that exceed its requirements in terms of QoS, switching to another band or reducing allocated power may leave some resources for other users with higher requirements. The challenge resides in designing efficient utility functions that allow obtaining optimal performance with regard to requirements.

Scalability: Not only do scalable protocols allow large systems to be handled in an acceptable time, but also to efficiently use resources according to the system size. For instance, reducing the exchange data and processing delays, and increasing the reusability of the bands will allow maximizing the optimization criteria.

CONCLUSION

This article presents an overview of the main methods for resource management for large-scale DSA systems. Centralized approaches are not adapted for large DSA systems due to their high computational complexity. In distributed approaches, heuristic methods based on behavioral evolution of resource allocation reward represent a promising solution due to their ease of implementation and low requirements in terms of information exchange and ability to track environment changes. In particular, we compare the learning- and particle-filtering-based DSA algorithms, and show that particle filtering could achieve a better performance in terms of throughput especially in fast changing environments. Research is still open in this direction to propose efficient approaches to predict environment changes based on previous observations, notably the primary users' occupancy, channels' conditions, and other users' behaviors.

ACKNOWLEDGMENT

This work was made possible by NPRP grant # NPRP 5-319-2-121 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] J. Mitola and G. Q. Maguire, "Cognitive Radio: Making Software Radios More Personal," *IEEE Personal Commun.*, vol. 6, no. 4, Aug 1999, pp. 13–18.

[2] E. Z. Tragos *et al.*, "Spectrum Assignment in Cognitive Radio Networks: A Comprehensive Survey," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 3, 2013, pp. 1108–35.

[3] A. Goldsmith *et al.*, "Breaking Spectrum Gridlock with Cognitive Radios: An Information Theoretic Perspective," *Proc. IEEE*, vol. 97, May 2009, no. 5, pp. 894–914.

[4] B. Hamdaoui, "Adaptive Spectrum Assessment for Opportunistic Access in Cognitive Radio Networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, Feb. 2009, pp. 922–30.

[5] G. Hattab and M. Ibnkahla, "Multiband Spectrum Access: Great Promises for Future Cognitive Radio Networks," *Proc. IEEE*, vol. 102, no. 3, Mar. 2014, pp. 282–306.

[6] B. Hamdaoui and K. G. Shin, "OS-MAC: An Efficient MAC Protocol for Spectrum-Agile Wireless Networks," *IEEE Trans. Mobile Computing*, vol. 7, no. 8, Aug. 2008, pp. 915–30.

[7] H. Ahmadi and Y. Chew, "Evolutionary Algorithms for Orthogonal Frequency Division Multiplexing-Based Dynamic Spectrum Access Systems," *Computer Networks*, vol. 56, no. 14, Sept. 2012, pp. 3206–18.

[8] X. Jiang *et al.*, "Quality of Service-Aware Coordinated Dynamic Spectrum Access: Prioritized Markov Model and Call Admission Control," *Wireless Commun. and Mobile Computing*, vol. 13, no. 5, 2013, pp. 510–24.

[9] G. Gür and F. Alagöz, "Green Wireless Communications via Cognitive Dimension: An Overview," *IEEE Network*, vol. 25, no. 2, Mar. 2011, pp. 50–56.

[10] Y. Rahulamathavan *et al.*, "Optimal Subcarrier and Bit Allocation Techniques for Cognitive Radio Networks Using Integer Linear Programming," *Proc. IEEE Wksp. Statistical Signal Processing*, Aug. 2009, pp. 293–96.

[11] G. Salami *et al.*, "A Comparison Between the Centralized and Distributed Approaches for Spectrum Management," *IEEE Commun. Surveys & Tutorials,* vol. 13, no. 2, 2011, pp. 274–90.

[12] Z. Ji and K. Liu, "Cognitive Radios for Dynamic Spectrum Access — Dynamic Spectrum Sharing: A Game Theoretical Overview," *IEEE Commun. Mag.*, vol. 45, no. 5, May 2007, pp. 88–94.

[13] M. NoroozOliaee, B. Hamdaoui, and K. Tumer, "Efficient Objective Functions for Coordinated Learning in Large-Scale Distributed OSA Systems," *IEEE Trans. Mobile Computing*, vol. 12, no. 5, May 2013, pp. 931–44.

[14] D. Zhang and Z. Tian, "Adaptive Games for Agile Spectrum Access based on Extended Kalman Filtering," *IEEE J. Sel. Topics Signal Processing*, vol. 1, no. 1, June 2007, pp. 79–90.

[15] M. Ben Ghorbel *et al.*, "Resources Allocation for Large-Scale Dynamic Spectrum Access System Using Particle Filtering," *Proc. IEEE GLOBECOM Wksp.*, Dec. 2014.

BIOGRAPHIES

MOHSEN GUIZANI [S'85, M'89, SM'99, F'09] (mguizani@ieee.org) is currently a professor in the Computer Science & Engineering Department of Qatar University. He has also served in academic positions at the University of Missouri-Kansas City, University of Colorado-Boulder, Syracuse University, and Kuwait University. He received his B.S. (with distinction) and M.S. degrees in electrical engineering, and M.S. and Ph.D. degrees in computer Engineering in 1984, 1986, 1987, and 1990, respectively, all from Syracuse University, New York. His research interests include wireless communications and mobile computing, computer networks, cloud computing, cyber security, and smart grid. He currently serves on the Editorial Boards of several international technical journals, and is the founder and Editor-in-Chief of the journal *Wireless Communications and Mobile Computing* published by John Wiley. He is the author of nine books and more than 400 publications in refereed journals and conferences (with an h-index = 30 according to Google Scholar). He received two best research awards. He is a member of the IEEE Communication Society and a Senior Member of ACM.

BASSEM KHALFI [S'14] is currently a Ph.D student at Oregon State University, Corvallis. He received his Diplome d'Ingenieur from Ecole Superieure de Communications de Tunis (SUP'COM) in Ariana, Tunisia, in 2012 and his M.Sc. from Ecole Nationale d'Ingenieurs de Tunis (ENIT) in Tunis, Tunisia, in 2014. His research interests include resource allocation in dynamic spectrum access systems and performance analysis for cooperative spectrum sharing.

MAHDI BEN GHORBEL [S'10, M'14] (mahdi.benghorbel@kaust.edu.sa) is currently a postdoctoral fellow at Qatar University, Doha, since September 2013. He received his

Diplome d'Ingenieur from Ecole Polytechnique de Tunisie (EPT), Tunis, Tunisia, in 2009 and his Ph.D. in electrical engineering from King Abdullah University of Science and Technology (KAUST), Saudi Arabia, in 2013. He received the Excellency Fellowship for his studies at EPT and the top student award of his university in June 2009. He also received a Provost Award and Discovery Fellowship to join KAUST in September 2009. His research interests include optimization of resource allocation and cooperative spectrum sensing performance for cognitive radio systems.

BECHIR HAMDAOUI [S'02, M'05, SM'12] (hamdaoui@eecs.orst.edu) is presently an associate professor in the School of EECS at Oregon State University. He received his Graduate Engineer Diploma (1997) from the National School of Engineers, Tunis, Tunisia. He also received M.S. degrees in both ECE (2002) and CS (2004), and a Ph.D. degree in computer engineering (2005), all from the University of Wisconsin-Madison. His research interests span various topics in the areas of wireless communications and computer networking systems. He has won the NSF CAREER Award (2009), and is presently an Associate Editor for *IEEE Transactions on Wireless Communications* (2013–present) and the *Wireless Communications and Computing Journal* (2009–present). He has also served as an Associate Editor for *IEEE Transactions on Vehicular Technology* (2009–2014) and the *Journal of Computer Systems, Networks, and Communications* (2007–2009). He served as Program Chair for SRC at ACM MobiCom 2011 and for many IEEE symposia/workshops, including ICC, IWCMC, and PERCOM. He also served on the TPCs of many conferences, including INFOCOM, ICC, and GLOBECOM. He is a member of the IEEE Computer, IEEE Communications, and IEEE Vehicular Technology Societies.

Matching Theory for Future Wireless Networks: Fundamentals and Applications

Yunan Gu, Walid Saad, Mehdi Bennis, Merouane Debbah, and Zhu Han

Yunan Gu and Zhu Han are with the University of Houston.

Walid Saad is with Virginia Tech and Kyung Hee University.

Mehdi Bennis is with the University of Oulu.

Merouane Debbah is with Huawei France R&D.

The work of Yunan Gu and Zhu Han was supported by the U.S. National Science Foundation under Grants CMMI-1434789, ECCS-1405121, CNS-1443917, CNS-1265268, CNS-0953377, and NSFC 61428101. The work of Walid Saad was supported by the U.S. National Science Foundation under Grants CNS-1460316 and CNS-1460333.

ABSTRACT

The emergence of novel wireless networking paradigms such as small cell and cognitive radio networks has forever transformed the way in which wireless systems are operated. In particular, the need for self-organizing solutions to manage the scarce spectral resources has become a prevalent theme in many emerging wireless systems. In this article, the first comprehensive tutorial on the use of matching theory, a Nobel Prize winning framework, for resource management in wireless networks is developed. To cater for the unique features of emerging wireless networks, a novel, wireless-oriented classification of matching theory is proposed. Then the key solution concepts and algorithmic implementations of this framework are exposed. The developed concepts are applied in three important wireless networking areas in order to demonstrate the usefulness of this analytical tool. Results show how matching theory can effectively improve the performance of resource allocation in all three applications discussed.

INTRODUCTION

Smartphones, tablets, and other handheld devices are exponentially increasing the traffic load in current wireless networks. To meet this increasing demand, several new communication paradigms have emerged:

- Cognitive radio (CR) networks, in which cognitive devices can adaptively opportunistically access the wireless spectrum, thus improving spectral utilization
- Small cell networks that boost wireless capacity and coverage via a viral deployment of low-cost small cell base stations
- Large-scale device-to-device communications that can occur over both cellular and unlicensed bands

This change in the wireless landscape is gradually leading to a future multi-tiered heterogeneous wireless architecture, as seen in Fig. 1.

Effectively managing resource allocation in such a complex environment warrants a fundamental shift from traditional centralized mechanisms toward self-organizing and self-optimizing approaches. The need for this shift is motivated by practical factors such as the increasing density of wireless networks and the need for communications with low latency. Even recent emerging centralized paradigms such as cloud-based radio access networking (RAN) will still require some form of self-organization due to country-specific backhaul constraints. In consequence, there is a need for self-organizing systems in which small cell base stations and even devices can have some intelligence to rapidly make resource management decisions.

Indeed, there has been a recent surge in the literature that proposes new mathematical tools for optimizing resource allocation in many emerging wireless systems. Examples include centralized optimization and game theory. Centralized optimization techniques can provide optimal solutions to resource allocation problems, and their algorithmic implementations have matured over the past few years. However, they often require global network information and centralized control, thus yielding significant overhead and complexity. This complexity can rapidly increase when dealing with combinatorial integer programming problems such as channel allocation and user association. Moreover, centralized optimization may not be able to properly handle the various challenges of dense and heterogeneous wireless environments such as the one in Fig. 1.

The aforementioned limitations of optimization have led to an interesting body of literature that deals with the use of noncooperative game theory for wireless resource allocation [1]. Despite their potential, such approaches present some shortcomings. First, classical game-theoretic algorithms such as best response will require some form of knowledge of other players' actions, thus limiting their distributed implementation. Second, most game-theoretic solutions, such as the

Nash equilibrium, investigate one-sided (or unilateral) stability notions in which equilibrium deviations are evaluated unilaterally per player. Such unilateral deviations may not be practical when investigating assignment problems between distinct sets of players. Last but not least, the tractability of equilibria in game-theoretic methods requires having some structure in the objective functions, which for practical wireless metrics may not always be satisfied.

Recently, *matching theory* has emerged as a promising technique for wireless resource allocation, which can overcome some limitations of game theory and optimization [2–6]. Matching theory is a Nobel Prize winning framework that provides mathematically tractable solutions for the combinatorial problem of matching players in two distinct sets [7–9], depending on the individual information and preference of each player. The advantages of matching theory for wireless resource management include:

- Suitable models for characterizing interactions between heterogeneous nodes, each of which has its own type, objective, and information
- The ability to define general "preferences" that can handle heterogeneous and complex considerations related to wireless quality-of-service (QoS)
- Suitable solutions, in terms of stability and optimality, that accurately reflect different system objectives
- Efficient algorithmic implementations that are inherently self-organizing and amenable to fast implementation

However, reaping the benefits of matching theory for wireless networks requires advancing this framework to handle their intrinsic properties such as interference and delay. Despite the surge in research that applies matching theory for wireless, most existing works are restricted to very limited aspects of resource allocation. This is mainly due to the sparsity of tutorials that tackle matching theory from an engineering perspective. For instance, most references, such as [7–9], focus on matching problems in microeconomics. In addition, although [10] provides an interesting introduction to matching theory for engineering, it does not explicitly explore the challenges of future wireless systems.

In this tutorial, we aim to provide a unified treatment of matching theory oriented toward engineering applications in general, and wireless networking in particular. The goal is to gather the state-of-the-art contributions that address the major opportunities and challenges in applying matching theory to the understanding of emerging wireless networks, with emphasis on both new analytical techniques and novel application scenarios. Beyond providing a self-contained tutorial on classical matching concepts, we introduce a new classification that is oriented toward next-generation wireless systems. For each class of matching problems, we provide the basic challenges, solution concepts, and potential applications. Then we conclude by summarizing the potential of matching theory as a tool for resource management in wireless networks.

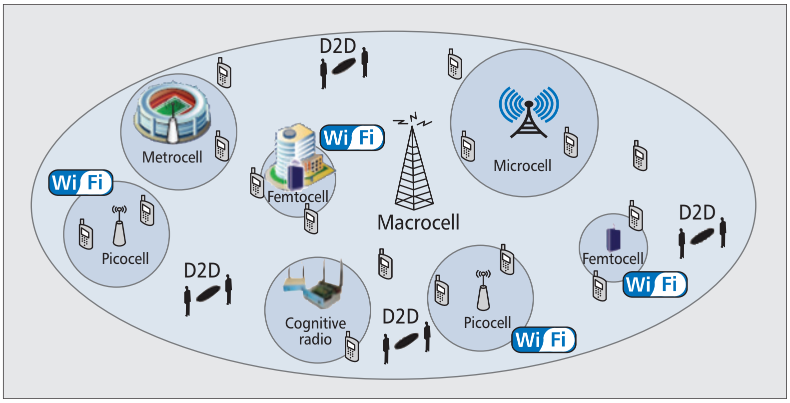


Figure 1. A future wireless network with a mixture of small cells, cognitive radio devices, and heterogeneous spectrum bands.

Matching Theory: Fundamentals and Conventional Classification

Basic Matching Definitions

The basic wireless resource management problem can be posed as a *matching problem* between resources and users. Depending on the scenario, the resources can be of different abstraction levels, representing base stations, time-frequency chunks, power, or others. Users can be devices, stations, or smartphone applications. Each user and resource has a quota that defines the maximum number of players with which it can be matched. The main goal of matching is to optimally match resources and users given their individual, often different, objectives and learned information. Each user (resource) builds a ranking of the resources (users) using a *preference relation*. The concept of a preference represents the individual view that each resource or user has of the other set, based on local information. In its basic form, a preference can simply be defined in terms of an objective utility function that quantifies the QoS achieved by a certain resource-user matching. However, a preference is more generic than a utility function in that it can incorporate additional qualitative measures extracted from the information available to users and resources.

A matching is essentially an allocation between resources and users. The basic solution concept for a matching problem is the so-called *two-sided stable matching*. A matching is said to be two-sided stable if and only if there is no blocking pair (BP). A BP for a stable marriage case is defined as a pair of user and resource (u, r) where u prefers r to its currently matched user j, and r prefers u to its currently matched resource k. Thus, u will leave i to be matched to r, and r would rather be matched to user u than user k. The implication of stability in a wireless network is further discussed later. This definition of stability can extend to all types of matching problems.

Conventional Classification

The classical classification of matching problems is based on the values of the player quotas as follows:

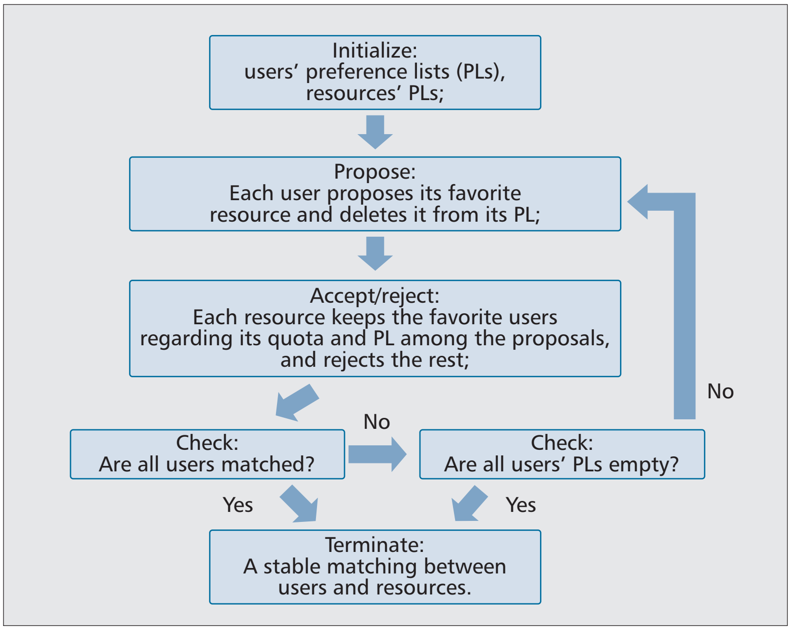


Figure 2. Deferred acceptance algorithm.

1. **One-to-one matching:** Each player can be matched to at most one member of the opposite set. The most prominent example is the stable marriage problem in which men and women need to be matched for marriage.
2. **Many-to-one matching:** Here, in one of the sets, at least one player can be matched to multiple players of the opposing set, while in the other set, every player has exactly one match. One example is the college admissions problem in which one student can be matched to one university while a university can recruit multiple students.
3. **Many-to-many matching:** At least one player within each of the two sets could be matched to more than one member in the other set. Many-to-many matching is the most general type of problem and has many examples such as creating partnerships in peer-to-peer networks.

There are other classifications for matching problems, such as based on the partitioning of players and the preference requirement for players. However, such classes can often be derived as special cases of the above matching problems.

Basic Algorithmic Solution: Deferred Acceptance

The seminal result in matching theory shows that *at least one stable matching exists* for general preferences in conventional one-to-one and one-to-many games [11]. This work also introduced an efficient algorithm, known as the deferred acceptance (DA) algorithm (polynomial time for one-to-one and empirically very fast for one-to-many), which can find such a matching. DA is an iterative procedure, shown in Fig. 2, in which players in one set make proposals to the other set, whose players, in turn, decide to accept or reject these proposals, respecting their quota. Users and resources make their decisions based on their individual preferences (e.g., available information or QoS metric). This process admits many *distributed implementations*, which do not require the players to know each other's preferences [11]. When the preferences are strict (no indifference), the stable matching is also Pareto optimal for the proposing players [11]. Extensions that balance the roles of proposing and receiving players or handle many-to-many cases have been developed such as in [10, 12].

From an information exchange point of view, even though DA requires players to submit proposals to one another, it does not require a centralized controller. During the information exchange (proposals), each player is required to only collect information on the players in whom they are interested from the opposite set to perform a ranking according to its preferences. The players need not observe the actions or preferences of other players. After building preference lists, the players take actions based on the local information they have collected without requiring any synchronization in time. The convergence of DA to stable matching is guaranteed irrespective of the order of play and without any synchronization. Therefore, a DA-based approach can be implemented in a distributed manner without requiring a central information collection center. For such distributed implementations, the results on the polynomial time convergence of one-to-one matching would still hold as corroborated by some recent studies [3, 13].

Matching in Wireless Networks: Fundamentals

Wireless-Oriented Classification

To capture the various wireless resource management features, we condense the rich matching literature into three new proposed classes of problems, illustrated in Fig. 3, having the following properties:

Class I: Canonical matching: This constitutes the baseline class in which the preference of any resource (user) depends solely on the information available at this resource (user) and on the users (resources) with which it is seeking to match. This is useful to study resource management within a single cell or for allocating orthogonal spectrum resources. This is particularly applicable, for example, to CR networks, in which one must allocate orthogonal licensed channels to a number of unlicensed users.

Class II: Matching with externalities: This class allows desirable matchings to be found when the problem exhibits "externalities," which translate into interdependencies between the players' preferences. For example, in a small cell network, whenever a user is associated with a resource, the preference of other users will automatically change, since this allocated resource can create interference at other resources using the same frequency. Thus, the preferences of any player depend not only on the information available at this player, but also on the entire matching of the others. We distinguish between two types of externalities: conventional externalities and peer effects. In the former, the dependence of the preferences is between players

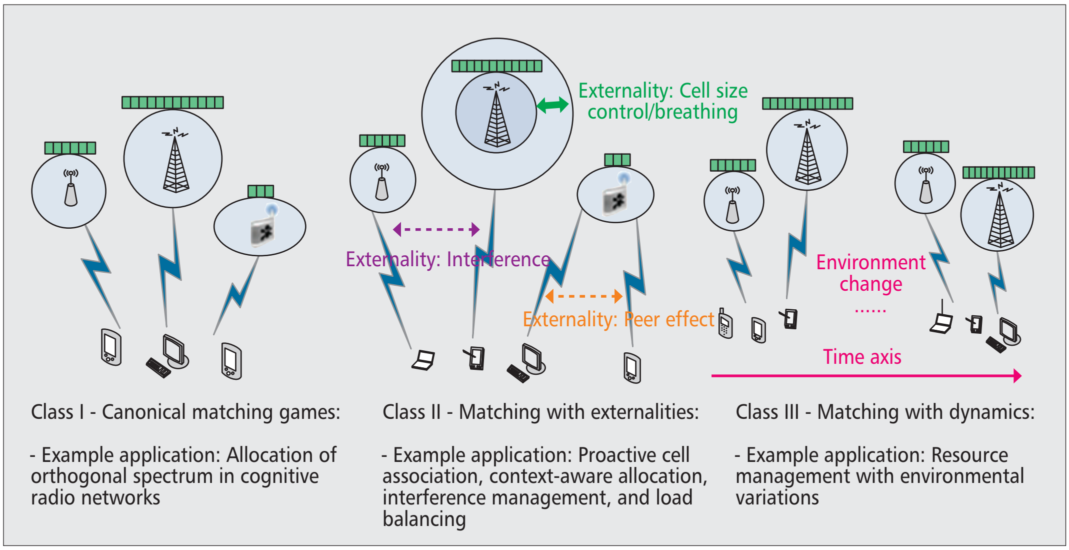


Figure 3. Novel wireless-oriented classification of matching theory.

The seminal result in matching theory shows that at least one stable matching exists for general preferences in conventional one-to-one and one-to-many games. This work also introduced an efficient algorithm, known as the deferred acceptance algorithm (polynomial time for one-to-one and empirically very fast for one-to-many), which can find such a matching.

matched to different players in the other set, such as in the interference example. In the latter, the preference of a user for a resource will depend on the identity and number of other users matched to the same resource. Such peer effects are abundant in a wireless environment due to factors such as delay.

Class III: Matching with dynamics: The third class, matching with dynamics, is suitable for scenarios in which one must adapt the matching processes to dynamics of the environment such as fast fading, mobility, and time-varying traffic. Here, at each time, the preferences of the players might change; thus, the time dimension must be accounted for in the matching solution. However, for a given time, the matching problem can be of either class I or class II.

Mathematically, the formulation of problems in all three classes will follow the basics of an earlier section. For class I, the preferences of one player set simply depend on the other player set. However, for class II, the preferences will now depend not only on the matched user, but also on the entire matching, due to externalities. For class III, one can introduce a time-dependent state variable in the matching. Subsequently, the preferences will now be time- and state-dependent if the problem has both dynamics and externalities. The transition between states depends on the application being studied. For example, if the state represents the activity pattern of a licensed user, the transition would follow a classic Markov model. In contrast, if the state represents a dynamically varying fast fading channel, one can use differential equations to represent the state transition.

Matching Theory in Wireless: Discussions

In wireless resource management, the matching stability notion discussed earlier implies robustness to deviations that can benefit both the resource owners and the users. In fact, an unstable matching can, for example, lead to undesirable cases in which a base station can swap its least preferred user with another since this swap is beneficial to both the resource and the user. Having such network-wide deviations ultimately leads to an unstable network operation. Remarkably, a recent result [2] has shown that classical schemes such as proportional fair often yield unstable matchings, which further motivates the need to analyze and optimize stable matchings for self-organizing wireless systems. This concept is very useful in matching problems and is broadly applicable to all classes.[1]

While the existence of a stable matching is guaranteed for canonical games in the one-to-one and one-to-many cases, such results do not readily map to many-to-many or to classes II and III. However, although DA and its variants were originally conceived for canonical matching, one can also use them as the nexus of new matching algorithms, tailored to the nature of wireless networks. Such algorithms can be used to establish the existence of stable matchings for classes II and II as well as to find outcomes with desirable efficiency properties.

Here, we note that there is no general existence result for stable matching with externalities. However, to handle externalities, one can utilize an iterative DA process that continuously updates the preferences based on the currently perceived matching. By exploiting the structure of externalities via wireless concepts such as interference graphs (e.g., who interferes with whom), one can analyze the convergence and stability of the resultant matching. Naturally, by building on such methods one can expand the realm of matching theory to handle externalities. Similarly, by integrating notions from stochastic games into matching, practical dynamic algorithms can be devised to find matchings that can cope with time-varying changes and are stable over time. The basic idea is to cast the matching problem as a stochastic game and then explore the rich literature on dynamic game theory [1] to solve this problem while ensuring that the solution will converge to a *two-sided stable matching* rather than a classical Nash equilibrium. The solution would now essentially be a dynamic and stochastic version of DA.

Although the above discussed matching solu-

[1] This concept can also be connected to other stronger or weaker stability notions (e.g., setwise or Nash stability) that have various interpretations.

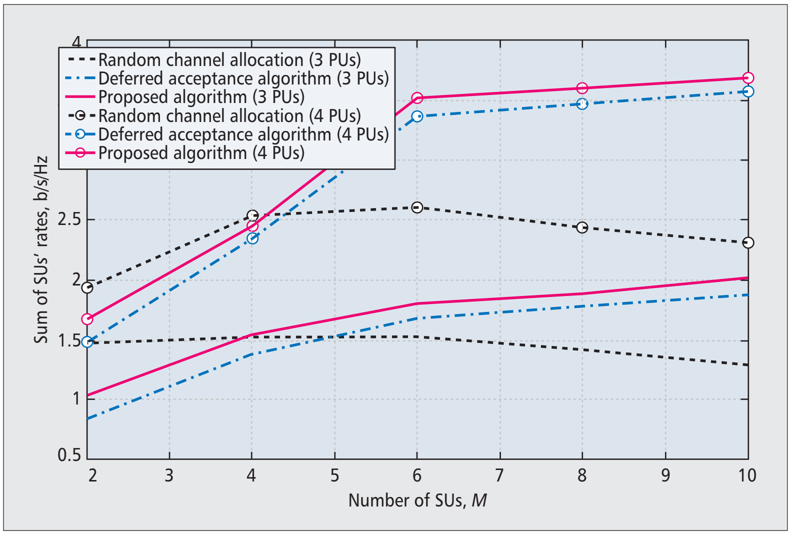


Figure 4. Example result showing how matching improves channel allocation in cognitive radio networks.

tions could provide stable matchings, they also have some limitations:

- Similar to game theory, matching problems can admit multiple stable solutions; thus, the selection of a desirable matching is a key design issue.
- The optimality of the stable solution may not be always guaranteed; however, one can utilize known techniques, such as pricing or optimized utility designs, to drive the matching solution toward an optimal and stable point,
- The exchange of proposals during DA requires additional signaling in a wireless network; however, one can exploit some structure of the problem to reduce the number of proposals as done in [3].

Matching Theory in Wireless Networks: Applications

Cognitive Radio Networks

Cognitive radio networks present a primary application of matching theory due to:

- The necessity of decentralized operation
- The need for dynamic spectrum access, which requires efficient resource management solutions
- The two-sided nature of CR in which licensed primary users (PUs) own channels that must be accessed by unlicensed secondary users (SUs)

Indeed, in CR networks, centralized optimization solutions are undesirable since PUs and SUs often belong to different operators and cannot be centrally controlled. On the other hand, in some CR problems such as PU-SU association, using a noncooperative game can lead to stable solutions in which matching PUs and SUs is done without requiring the two-sided consent of both PUs and SUs.

The suitability of matching theory for CR has been corroborated by a number of recent works such as [3, 4]. In particular, [3] presented one of the first works in this area. In this work, a one-to-one matching problem is formulated between a number of SUs and a number of PUs (channels). The channels are assumed to be orthogonal; hence, the game is a canonical matching game. The preferences of both users and channels are based on the same utility function, which primarily captures the rate of transmission. Under this model, it is shown that:

- A unique stable matching exists.
- A modified version of the DA algorithm can be used to find the stable allocation in a time efficient manner.

This work was extended in [4] to account for energy efficiency.

Recently, we also studied a one-to-one matching problem between SUs and PUs in which the SUs rank the PUs based on their confidence in sensing the PUs' channels. In particular, using a soft-decision Bayesian framework, we quantified the accuracy of the sensing of each channel and incorporated this metric into the SUs' preferences. Prior to matching, each SU evaluates its appreciation of the PU channel by capturing the effect of confidence in sensing as well as rate. Then the PUs actively participate in the association process based on two cases:

- When inactive, the PU prefers to grant its band to the SU with the highest sensing detection and rate to better exploit the channel.
- When active, the PU prefers to protect its band, and thus will attempt to limit or deny association.

Here, we show that the matching is canonical and adopt a modified DA algorithm that allows the PUs to handle the aforementioned property. As shown in Fig. 4, for the studied scenario, the matching-based algorithm yields significant performance gains, in terms of the SUs sum rate, when compared to classical random channel allocation schemes. Moreover, the modified DA algorithm also presents sum rate improvements over classical DA (similar gains can be seen in terms of convergence time).

Clearly, CR networks present an important avenue for matching theory. Many extensions to the existing works can be envisioned, particularly by exploring matching with externalities (under interference constraints) and dynamic matching (given time-varying PU activity).

Heterogeneous Small-Cell-Based Networks

Heterogeneous (small-cell-based wireless) networks (HetNets) present an important application of matching theory due to their heterogeneity and scale. Also, there has been increased interest recently in developing context-aware or user-centric HetNets that can exploit new dimensions such as social metrics to improve resource allocation. Such context awareness further motivates the need for distributed solutions that account for the individual context available at each node — similar to how matching captures individual preferences. Given this striking analogy between matching theory and resource management in HetNets, the proposed classes can be used to address a variety of problems, which include interference manage-

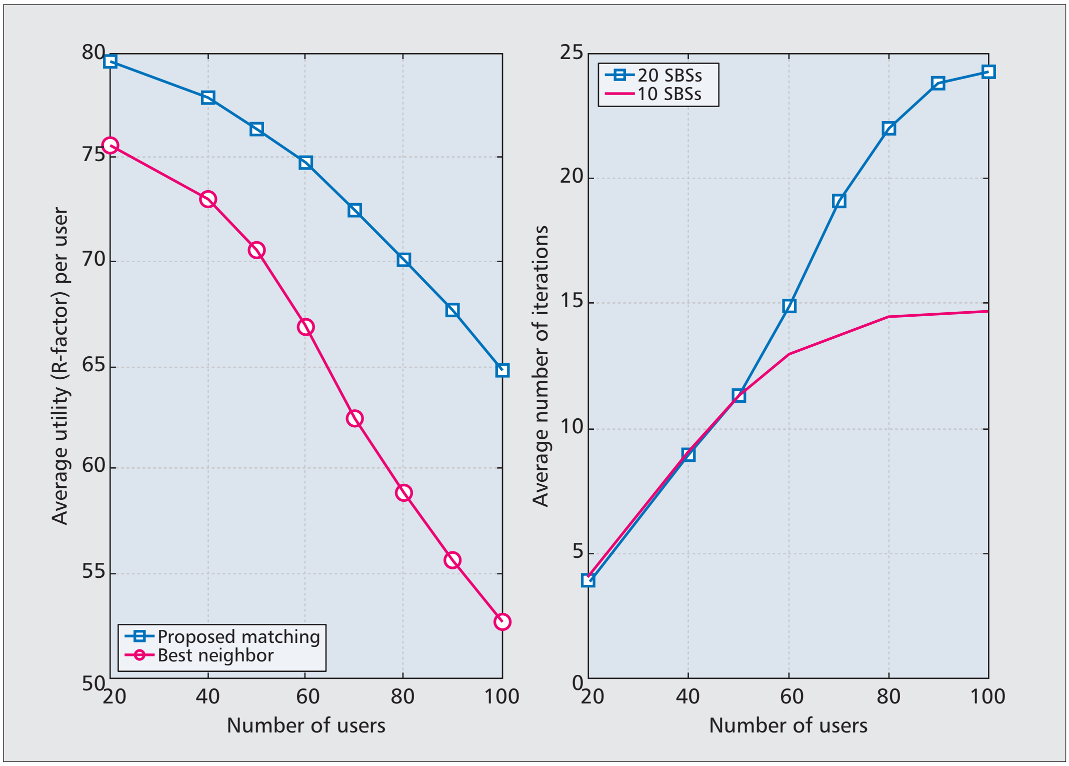


Figure 5. Example result showing how matching theory can be used to improve uplink cell association in HetNets.

HetNets present an important application of matching theory due to their heterogeneity and scale. Also, there has been an increased recent interest in developing context-aware or user-centric HetNets that can exploit new dimensions such as social metrics to improve resource allocation.

ment, handover management, caching, and cell association.

Here, matching is preferred over optimization due to:
- The density and scale of HetNets, which motivate self-organizing solutions
- The need to account for the context present at each small cell base station (SBS) and device instead of a single global utility function
- The centralized optimization approach, which will generally yield a combinatorial problem, particularly, in the presence of heterogeneous context, which limits its applicability here

Moreover, although a noncooperative game is also applicable here, it will have a number of limitations that include the need to observe (at least partially) all players' preferences and the fact that the solution concepts would not account for two-sided stability as previously explained.

In [14], we studied the problem of cell association in the uplink of a HetNet. The basic model here is an uplink HetNet model in which each user needs to decide to which SBS it should be assigned . The problem is formulated as a one-to-many matching model in which a user can be associated with only one SBS, and an SBS can admit a certain quota of users. The users' preferences over SBSs capture the bit error rate and delay trade-off that they can achieve, while the SBSs' preferences favor load balancing by pushing users to the smaller cells without jeopardizing QoS. Such a load balancing is essentially a form of cell biasing in which an SBS would offload some users from the macrocell and service them directly. Here, we also consider the delay at each SBS due to the increasing load and the limited capacity of the backhaul that connects the SBSs to the core network. Therefore, although orthogonal spectrum is considered, due to the delay, the matching problem is shown to have *peer effects,* and thus belongs to class II, matching with externalities.

Due to peer effects, applying DA and its variants will not yield a stable matching. Instead, we developed a new algorithm that starts with a distributed DA-based process using initial preferences based on the worst case delay. Then, as the nodes measure externalities, they modify their preferences and change their choices by transferring to other SBSs. Then we show that due to the presence of transfers in the model, delay-sensitive users will trade-off two-sided stability for a weaker stability that achieves better QoS. Figure 5 shows simulation results for 2 macrocells and 10 SBSs. We can see that the matching-based approach outperforms the benchmark best neighbor scheme often adopted in classical cellular systems, with up to 23 percent improvement in average user utility. Figure 5 also shows a reasonable convergence time that grows slowly with network size.

One can extend this framework of matching with peer effects or, more generally, externalities to many other areas in HetNets. For example, in [5], the framework is extended to account for interference and perform downlink cell association for a context-aware network in which preferences capture a palette of information including application type, hardware size, and physical layer metrics. In addition, as shown in in [13], one can explore canonical matching

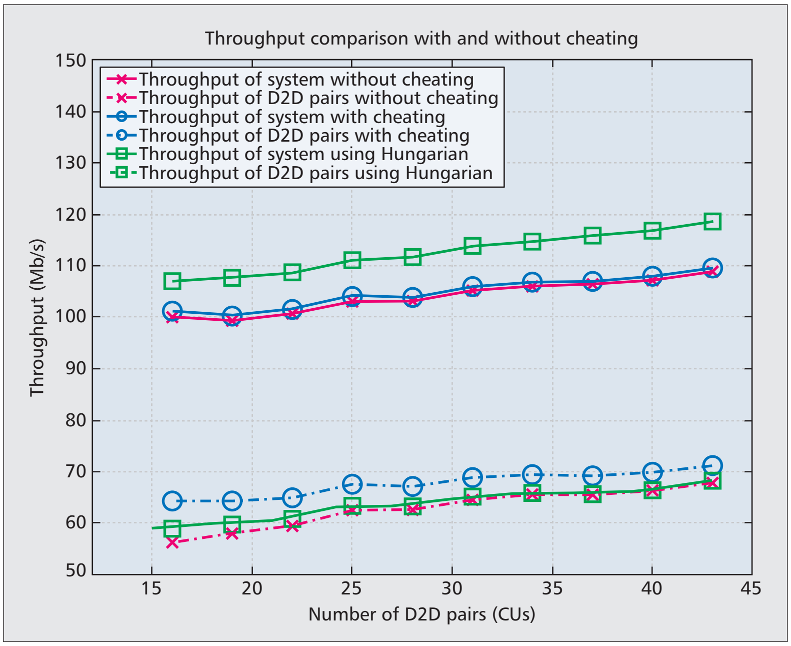


Figure 6. Example result showing how cheating further improves both DU and system utilities.

models to study the association not only at the radio level, but also at the level of operators.

In a nutshell, matching-theoretic models, in all three classes, can serve as a fundamental analytical tool for future cellular systems. Beyond the examples discussed above, one can envision several new models such as many-to-many matching models for caching, dynamic matching models for handling mobility, and stochastic matching models that smartly combine matching with stochastic geometry.

Device-to-Device Communications

One promising technology to overcome the ever-increasing wireless capacity crunch is device-to-device (D2D) communications. Using D2D, mobile users can communicate directly over cellular spectrum bands while bypassing the base stations (BSs). As D2D users may share spectral bands with one another as well as with the cellular network, the introduction of D2D in cellular networks will lead to new challenges in terms of interference management and resource allocation. Thus, it will provide an important application for matching theory.

Optimizing resources for D2D communication using centralized optimization can result in more overhead due to information exchange and centralized computation. In particular, centralized optimization will require not only dynamic information collection at the BS from all possible D2D pairs, but will also need to deal with increased computation complexity at the BS level. Formulating the D2D problem as a noncooperative game will be limited by the fact that it will still rely on individual stability and the need for D2D users to observe the preferences of other players. To counter these limitations, it is of interest to develop matching-theoretic models for D2D communications. To this end, we observe that D2D typically involves two types of users: cellular users (CUs) and D2D users (DUs). In the underlay mode of D2D operation, the CUs are exclusively assigned licensed spectrum chunks from the BSs, while the DUs must share the spectrum with the CUs.

Here, we formulate a two-sided matching problem between the CUs and DUs. Each CU and DU starts by building its preference list by observing the necessary information (e.g., the channel condition, the transmission power, and specific QoS requirements) on the other types of players. Here, the preferences of CUs over DUs are defined as monetary payment from the DUs or the incurred interference on CUs. The DUs build their preferences over CUs based on the channel conditions or achievable transmission rate. A CU-DU matching is said to be unacceptable if the system's QoS requirements are violated. Players that are unacceptable are then removed from each other's preferences. Then each player (CU or DU) sorts the acceptable players in descending order of its preferences. After setting up the preferences, proper matching algorithms must be developed to achieve the required system objectives such as maximizing the throughput. For example, when using the DA algorithm, as explained in Fig. 2, DUs will propose to the CUs who, in turn, will accept or reject the received applications. The complexity of this iterative process depends on the total number of acceptable pairs m, which is $\mathcal{O}(m)$ [8].

However, beyond considering a global utility, in some scenarios, one is interested in optimizing the performance of one of the two sets of players, such as the DUs. Based on our work in [15], we proposed the idea of incorporating a form of "cheating" in the preferences in an effort to improve the DUs' utilities. Cheating is done by enabling the DUs to smartly change their preferences so as to reap more performance gains. As shown in Fig. 6, the use of such cheating strategy can improve the DUs' utility compared to the DA algorithm, and it also outperforms the Hungarian algorithm, which achieves the optimal system utility without guaranteeing system stability. In addition, as done in [6], one can extend such D2D models to cases in which D2D communication must explore, beyond physical layer parameters, the social ties of the users. Here, as shown in [6], one can cast the problem as a class II problem to capture peer effects that reflect how socially connected the users are who utilize D2D communication with an anchor device (i.e., used for content distribution or caching by the BSs). For such a model, one can enhance the DA algorithm to account for peer effects and show its convergence to a two-sided stable matching.

D2D is undoubtedly an important application area for matching theory with a promising set of future problems.

Conclusion

In this article, we have provided the first comprehensive tutorial on using matching theory for developing innovative resource management mechanisms in wireless networks. First, we have provided the fundamental concepts of matching theory and discussed a variety of properties that allow the definition of several classes of matching scenarios. Then we have proposed three new engineering-oriented classes of matching theory that can be adopted in wireless networking environments. For each class, we have developed the basic concepts and solutions that can be used to address related problems. We have provided a detailed treatment on how to use such matching-theoretic tools in specific wireless applications. In a nutshell, this article is expected to provide an accessible and holistic tutorial on the use of new techniques from matching theory for addressing pertinent problems in emerging wireless systems.

References

[1] Z. Han *et al.*, *Game Theory in Wireless and Communication Networks: Theory, Models and Applications*, Cambridge Univ. Press, Oct. 2011.

[2] E. A. Jorswieck, "Stable Matchings for Resource Allocation in Wireless Networks," *Proc. 17th Int'l. Conf. Digital Signal Processing*, Corfu, Greece, July 2011.

[3] A. Leshem, E. Zehavi, and Y. Yaffe, "Multichannel Opportunistic Carrier Sensing for Stable Channel Access Control in Cognitive Radio Systems," *IEEE JSAC*, vol. 30, no. 1, Jan. 2012, pp. 82–95.

[4] O. Naparstek, A. Leshem, and E. A. Jorswieck, "Distributed Medium Access Control for Energy Efficient Transmission in Cognitive Radios," arXiv preprint arXiv:1401.1671, 2014.

[5] F. Pantisano *et al.*, "Matching with Externalities for Context-Aware User-Cell Association in Small Cell Networks," Atlanta, GA, Dec. 2013.

[6] O. Semiari *et al.*, "On Self-Organizing Resource Allocation for Social Context-Aware small cell networks," *Proc. 1st KuVS Wksp. Anticipatory Networks*, Stuttgart, Germany, Sept. 2014.

[7] A. Roth and M. A. O. Sotomayor, *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Cambridge Univ. Press, Mar. 1992.

[8] D. F. Manlove, *Algorithmics of Matching Under Preferences*, World Scientific, 2013.

[9] R. W. Irving, P. Leather, and D. Gusfield, "An Efficient Algorithm for the Optimal Stable Marriage," *J. ACM*, vol. 34, no. 3, July 1987, pp. 532–43.

[10] H. Xu and B. Li, "Seen as Stable Marriages," *Proc. IEEE INFOCOM*, Shanghai, China, Mar. 2011.

[11] D. Gale and L. S. Shapley, "College Admissions and the Stability of Marriage," *Amer. Math. Monthly*, vol. 69, no. 1, Jan. 1962, pp. 9–15.

[12] K. Hamidouche, W. Saad, and M. Debbah, "Many-to-Many Matching Games for Proactive Social-Caching in Wireless Small Cell Networks," *Proc. 12th Int'l. Symp. Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, Hammamet, Tunisia, May 2014.

[13] S. Bayat *et al.*, "Multiple Operator and Multiple Femtocell Networks: Distributed Stable Matching," *IEEE ICC*, Ottawa, Canada, June 2012.

[14] W. Saad *et al.*, "A College Admissions Game for Uplink User Association in Wireless Small Cell Networks," *Proc. IEEE INFOCOM*, Toronto, Canada, Apr.–May 2014.

[15] Y. Gu *et al.*, "Cheating in Matching of Device to Device Pairs in Cellular Networks," *Proc. IEEE GLOBECOM*, Austin, TX, Dec. 2014.

Biographies

YUNAN GU (ygu6@uh.edu) received her B. S. degree in electronic engineering from Nanjing University of Science and Technology in June 2011 and her M. S. in computer science from Texas Southern University in 2013. Now, she is a Ph.D student in the Electronic and Computer Engineering Department at the University of Houston. Her research interests include wireless communications and matching theory.

WALID SAAD (walids@vt.edu) received his B.E., M.E., and Ph.D. degrees in computer and communications engineering from the Lebanese University in 2004, the American University of Beirut (AUB) in 2007, and the University of Oslo in 2010, respectively. Currently, he is an assistant professor at the Bradley Department of Electrical and Computer Engineering at Virginia Tech. He is also an International Scholar with the Department of Computer Engineering, Kyung Hee University, South Korea. His research interests include wireless and social networks, game theory, cybersecurity, cyber-physical systems, and cognitive radio networks.

MEHDI BENNIS (bennis@ee.oulu.fi) received his M.Sc. degree from the Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland, and the Eurecom Institute, France, in 2002. In 2004, he joined the Centre for Wireless Communications at the University of Oulu, Finland, as a research scientist. He obtained his Ph.D in December 2009 on spectrum sharing for future mobile cellular systems. His research interests include radio resource management, heterogeneous and small cell networks, game theory, and machine learning. He is currently the Principal Investigator of the European CELTIC project SHARING.

MÉROUANE DEBBAH (merouane.debbah@huawei.com) entered the Ecole Normale Supérieure de Cachan, France, in 1996, where he received his M.Sc and Ph.D. degrees. From 2003 to 2007, he was with the Mobile Communications department of the Institut Eurecom, Sophia Antipolis, France, as an assistant professor. Since 2007, he has been a full professor at Supelec. Since 2014, he is vice-president of the Huawei France R&D center and director of the Mathematical and Algorithmic Sciences Lab. His research interests are in information theory, signal processing, and wireless communications.

ZHU HAN [S'01, M'04, SM'09, F'14] (zhan2@uh.edu) received his B.S. degree in electronic engineering his Tsinghua University in 1997, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1999 and 2003, respectively. Currently, he is an associate professor in the Electrical and Computer Engineering Department at the University of Houston, Texas. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, wireless multimedia, security, and smart grid communication.

As D2D users may share spectral bands with one another as well as with the cellular network, the introduction of D2D in cellular networks will lead to new challenges in terms of interference management and resource allocation. Thus, it will provide an important application for matching theory.

Cooperative Heterogeneous Framework for Spectrum Harvesting in Cognitive Cellular Network

Ning Zhang, Haibo Zhou, Kan Zheng, Nan Cheng, Jon W. Mark, and Xuemin (Sherman) Shen

Ning Zhang, Haibo Zhou, Nan Cheng, Jon W. Mark, and Xuemin Shen are with the University of Waterloo.

Kan Zheng is with Beijing University of Posts and Telecommunications.

Abstract

With the proliferation of mobile devices and emerging data-hungry applications, mobile data has been increasing dramatically. To accommodate massive mobile data, the cellular network has been straining to meet the need due to the scarcity of spectrum. As a promising technology, cognitive radio can be leveraged by the cellular network to harvest spectrum holes on demand. By employing cognitive radio, the cellular network becomes a cognitive cellular network. In this article, we first provide an overview of the cognitive cellular network, including the network architecture and main applications. Then existing spectrum harvesting approaches are reviewed, and the limitations are identified. To better explore spectrum access opportunities, three types of cooperation-based approaches are introduced for different scenarios, based on which an integrated cooperative framework is devised to fully harvest spectrum holes. Simulation results are provided to evaluate the performance of the proposed cooperative approaches.

Introduction

We have witnessed a boost in the growth of the cellular network, which continues to flourish worldwide. Billions of wireless devices are proliferating, and various wireless applications, such as mobile social networks, online gaming, and high-definition video streaming, are emerging. As a result, mobile traffic keeps increasing dramatically, which has overwhelmed the cellular network. In the future, mobile traffic is expected to skyrocket further. According to the Cisco Visual Networking Index (VNI) [1], mobile traffic is expected to grow at a compound annual growth rate (CAGR) of 61 percent from 2013 to 2018. It is predicted that mobile traffic will continuously grow up to 1000 times by 2020 from that in 2010. The massive number of mobile devices and ever increasing mobile traffic pose a great challenge for the cellular network to provide users with quality guaranteed services.

To accommodate the explosive growth in mobile traffic and devices, a large amount of wireless spectrum is needed. The spectrum demand for mobile systems is predicted to be around 1280–1720 MHz by 2020 [2], while the current mobile systems only own the spectrum around 230–430 MHz. To address the challenge, one way is to add more spectrum resource. The Federal Communications Commission (FCC) is trying to free up additional spectrum of 500 MHz for mobile networks, which still cannot bridge the gap. Another path is to improve spectrum efficiency by developing new physical and link layer techniques, which usually incurs high complexity, and the improvement is limited. Alternatively, spectrum harvesting could harvest unused spectrum bands to meet the ever growing spectrum demand [3], considering that a large portion of spectrum is underutilized temporally and spatially. Spectrum harvesting can be enabled by cognitive radio (CR), which is able to acquire information from surroundings through sensing and adapt to environments by adjusting the operating parameters [4, 5]. By employing CR technology, the cellular network can harvest and operate over the unused spectrum/spectrum holes from other systems, such as TV white space (TVWS). The cellular network equipped with CR technology is also referred to as the cognitive cellular network (CCN) [6].

In the CCN, through spectrum harvesting, additional spectrum can be utilized for traffic offloading or capacity relief. Furthermore, the harvested spectrum may present different propagation or penetration features, which can be exploited to support diverse applications with various quality of service (QoS) requirements. For instance, if the communication distance is long, a low frequency band can be exploited, while a high and wide frequency band is selected if high throughput is preferred. Besides spectrum harvesting, the CCN can also support many promising applications. For instance, CR can be utilized to mitigate intercell interference among the macrocell and small cells, which are expected to be deployed densely and underlay the macrocell in the cellular network. With CR, small cell base stations (BSs) can perform spectrum sensing before accessing the spectrum bands to mitigate interference [7]. In addition, users can adjust the transmission parameters to achieve different objectives (e.g., energy efficiency or high transmission rate) based on its operating context.

As the key enabling technology, spectrum sensing plays a vital role to identify unused spectrum bands for spectrum harvesting. Cognitive cellular users (CCUs) can access the spectrum band from other systems only when it is detected idle. Otherwise, the normal operation of the primary users, or PUs (the licensed users in other systems), will experience interference. However, due to fading and shadowing, spectrum sensing of individual users is not perfect, and sensing errors are inevitable, which adversely affect the performance of both CCUs and PUs [8]. In this article, we introduce user cooperation to overcome the limitations of individual spectrum sensing and better harvest unused spectrum for the CCN. We first present an overview of the CCN, including the network architecture and main applications. Then the existing spectrum harvesting approaches for exploiting additional spectrum are reviewed, and the limitations are discussed accordingly. To overcome the limitations, cooperation-based spectrum harvesting approaches for various scenarios are introduced; for example, cooperative spectrum sensing can be utilized to better detect unused spectrum, and cooperative relaying can be leveraged to explore access opportunities even though the PUs are active. Then an integrated cooperative framework is devised to combine these cooperation-based schemes, aiming to fully explore spectrum access opportunities. Simulation results are provided to evaluate the performance of the proposed cooperation schemes.

Overview of the Cognitive Cellular Network

Network Architecture

The future cellular network, that is, the fifth generation (5G) network, is envisioned to be a heterogeneous network [9]. Therefore, the CCN has a heterogeneous network architecture, as shown in Fig. 1, where different small cells underlay the macrocell: picocells, femtocells, and so on. The BSs of the macrocell and small cells, as well as the user equipments (UEs), are equipped with CR technology. For ease of presentation, all the network entities with CR are called cognitive cellular users (CCUs), including cognitive UEs and BSs. The BSs of macrocell and small cells are connected to the external spectrum trading market server, which facilitates the spectrum trading/leasing between the CCN and other systems with licensed spectrum bands. CCUs can harvest spectrum holes through either spectrum sensing or spectrum trading.

The operation of the CCN can be centralized, where the BS coordinates CCUs for spectrum harvesting and spectrum sharing. For instance, it can coordinate CCUs to perform spectrum sensing and then combine the sensing results to make a final decision [10]. With the harvested spectrum and the cellular bands, the BS performs resource allocation for CCUs based on their traffic types and QoS requirements. For example, when the signal reception for a CCU is very poor in an indoor environment, the BS can switch the CCU to other spectrum bands with better penetration features (e.g., TV bands).

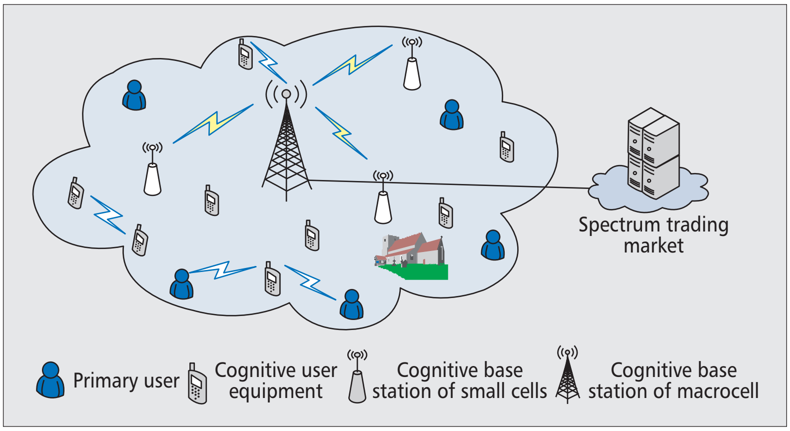


Figure 1. Network architecture.

The CCN can also operate in a decentralized fashion, where CCUs acquire available spectrum and access without a central controller. For instance, CCUs can sense the spectrum bands and select suitable spectrum bands based on their own needs. For instance, when the sender and receiver are far away, the low spectrum bands can be utilized due to the desirable propagation and penetration features. Moreover, different portions of unused spectrum band can be aggregated by the CCU to have a large spectrum band by means of spectrum bonding and aggregation.

Main Applications

Besides spectrum harvesting, CR also facilitates other applications.

Intercell Interference Mitigation: By deploying small cells densely, network capacity can be significantly improved. However, the main limiting factor for dense deployment of small cells is intercell interference among neighboring cells. One way to mitigate interference is spectrum splitting, which could lead to inefficient spectrum usage. With CR, small cells can sense the surroundings to detect the spectrum usage status and access the cellular spectrum bands only when they are not occupied. Moreover, small cells can also explore the unused spectrum bands owned by other systems, while the macrocell exclusively uses the cellular spectrum bands. By doing so, intercell interference can be mitigated, and different cells can coexist.

Transmission Adaption: The two main characteristics of CR are cognitive capability and reconfigurability. With cognitive capability, the CCU can get the information regarding the surrounding environments (the communication network, channel conditions, etc.). With

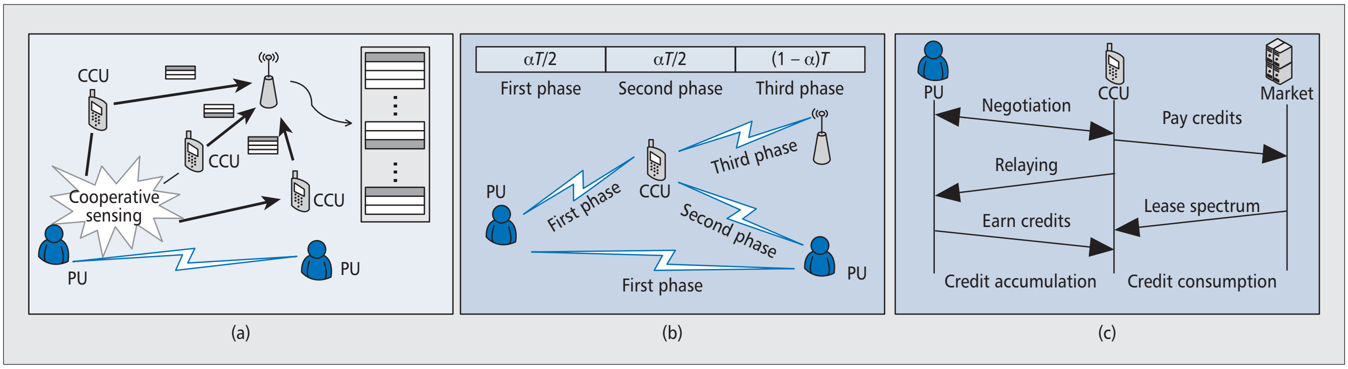


Figure 2. Cooperation-based spectrum harvesting.

reconfigurability, the CCU can dynamically and autonomously adjust the operating parameters to satisfy the QoS requirement in terms of throughput, reliability, delay, and so on, being aware of the operating environments. For the future cellular network, in order to achieve network access any time and anywhere, the CCU can detect the surrounding available network service and change the operating parameters for access.

Spectrum Harvesting Approaches and Challenges

The basic and main target of the CCN is to harvest spectrum holes to support mobile traffic growth. In this section, the existing spectrum harvesting approaches are reviewed.

Spectrum Harvesting Approaches

Geolocation database: Due to the transition from analog TV to digital TV, TVWS is considered to be allowed for free secondary access. Since TV programs are predetermined, the channel availability information can be provided by a geolocation database, which is governed by certified third parties [11]. The procedure for the geolocation database approach is as follows. The CCU sends a request to the geolocation database with location information. Then the database sends back the channel availability information for the given location.

In terms of implementation, Google and Microsoft have launched their geolocation database products, Google spectrum database and Microsoft spectrum observatory, respectively.

Spectrum sensing: Spectrum sensing is the key approach to spectrum harvesting. The CCUs perform spectrum sensing to detect spectrum holes and prevent interference to the PUs. The CCUs access the channel only when the channel is detected to be idle. Since spectrum holes can be in a specific time, frequency band, or spatial location, spectrum sensing can be performed in the time, frequency, and space domains. Spectrum sensing can also be used to acquire information regarding the carrier frequency, modulation schemes, bandwidth, and so on. Popular spectrum sensing methods include energy detection, matched filtering, and cyclostationary detection [8].

Spectrum trading: Spectrum trading allows the spectrum resource to be exchanged among different parties, on either a long- or short-term basis. The resources for trading can be the available frequency bands, the maximum allowable transmission power, and the time duration for access in the spectrum bands [12]. The CCUs can gain temporary exclusive rights to use the spectrum by paying a certain amount of money when PUs are willing to lease the spectrum for monetary rewards. During transmission, the CCUs will not be interrupted by the PUs, so their QoS can be guaranteed to some degree.

Limitations

Geolocation Database: To acquire channel availability, geolocation information is required, which is usually obtained from the global positioning system (GPS). However, some CCUs might not have GPS functions. Moreover, the geolocation information might be inaccurate due to poor GPS signals at some locations. In addition, there are two types of PUs in the TV spectrum band: TV broadcasters and microphones. Since it is difficult to predict the activities of microphones, they might be interfered with by CCUs using this approach. Last but not least, the database can only provide channel availability for TV bands [13].

Spectrum sensing: Spectrum sensing is considered as the main approach to harvesting spectrum holes. However, the detection performance of spectrum sensing is limited by several factors such as multipath fading and shadowing [8]. Specifically, when the CCU experiences multipath fading or shadowing, the reception of a PU's signal will be significantly degraded, which reduces the detection accuracy dramatically. In addition, for CCUs that are out of the transmission range of the primary transmitter, it is impossible to detect the PU's presence. Then the CCUs might start to transmit, causing harmful interference to the primary receiver. Moreover, once the PU reclaims the channel, the CCUs have to vacate the current channels and find other idle channels, resulting in intermittent transmission.

Spectrum trading: CCUs can have exclusive spectrum access rights through spectrum trading so that the QoS can be guaranteed. Long-term spectrum trading still has the problem of spectrum waste because the CCUs might have no

Cooperation Schemes	Scenarios	Partners	Objective
Cooperative spectrum sensing	Have traffic and cellular bands are not sufficient	CCUs with CCUs	Detect idle bands where no active PUs are present
Cooperation with PUs for access	Have traffic and no sufficient bands even through sensing	CCUs with active PUs	Gain access time through cooperation with PUs
Cooperation with PUs for credits	No traffic now but may have in the future	CCUs with active PUs	Earn credits for future use

Table 1. Comparison among different cooperation schemes.

traffic at a specific time due to the bursty nature of the traffic. For short-term or even real-time spectrum trading, the spectrum utilization can be improved. However, from the perspective of the CCN or CCUs, besides the operational and maintenance costs, additional money has to be paid when purchasing the spectrum through the market. Over a long period the cost could be fairly high, resulting in a burden on the CCN or CCUs. From the viewpoint of PUs, they might not only pursue monetary benefits through spectrum trading. In some cases, the PUs also desire to improve the transmission performance (e.g., increase throughput whenever they have a large volume of traffic). The existing trading mechanism does not consider these types of PUs' needs.

Cooperation-Based Approaches

The geolocation database approach can be complemented by spectrum sensing to protect microphones and explore spectrum holes in other bands. However, the detection performance of individual spectrum sensing is limited, as mentioned before. To overcome these issues, user cooperation can be leveraged, mainly in two forms: cooperative spectrum sensing [10] and cognitive cooperative relaying [14]. For the former, cooperation is carried out among CCUs, where multiple CCUs cooperate with each other to enhance the sensing performance for better detecting spectrum holes, as shown in Fig. 2a. For the latter, cooperation is performed between CCUs and active PUs, where CCUs cooperate with PUs to improve the transmission performance of the PUs and gain spectrum access opportunities as a reward, as shown in Fig. 2b.

Cooperative spectrum sensing aims to harvest the spectrum holes where no active PUs are present, while cognitive cooperative relaying tries to explore spectrum access opportunities when PUs are active. In addition, CCUs can acquire exclusive spectrum rights through the spectrum trading market. To compensate for the cost in spectrum marketing, CCUs can cooperate with PUs to improve the PUs' transmission performance and request credits when they are idle. That is, through cooperation, idle CCUs can accumulate credits, and the PUs' performance can be enhanced. The earned credits can be used by CCUs in the spectrum trading market when needed in the future, as shown in Fig. 2c. The comparison among different cooperation schemes is provided in Table 1.

Cooperation-Based Spectrum Harvesting Approaches

In this section, the three cooperation-based spectrum harvesting approaches are introduced.

Cooperation for Spectrum Sensing

Cooperative spectrum sensing can improve the detection performance in terms of increasing the detection probability and reducing the false alarm probability by exploiting spatial diversity and multiuser diversity. In cooperative spectrum sensing, each CCU performs local sensing and then forwards the results to a fusion center (e.g., the BS) to make a final decision. The fusion rules to combine the sensing results include AND rule, OR rule, and so on. Through cooperation, the deficiency of individual observations can be mitigated. Since CCUs are usually in a multi-channel environment, the fundamental issue is how to coordinate CCUs for multi-channel sensing. In what follows, the cross-entropy (C-E) method is applied to schedule CCUs to sense different channels for better harvesting spectrum holes.

Suppose that there are K primary channels that N CCUs can explore for transmission. The states of the channels alternate between ON (busy) and OFF (idle), which are modeled by an ON-OFF model with different transition rates. To better harvest spectrum holes, the BS coordinates the CCUs to perform cooperative spectrum sensing. Specifically, each individual CCU adopts energy detection for spectrum sensing and reports the sensing results to the BS, which makes a combined decision. Considering the AND rule is adopted as the fusion rule, a channel is deemed busy when all the CCUs report the OFF state. The detection probability and false alarm probability[1] of CCU i to sense channel j are denoted by $p_d(i, j)$ and $p_f(i, j)$, respectively, which are functions of the associated channel conditions. The expected available time for a given channel j is $T^j_{off} p^j_{off}(1 - F^j_f)$, where T^j_{off} is the sojourn time of the OFF state, p^j_{off} is the associated probability, and F^j_f is the cooperative false alarm probability. Note that $F^j_f = \prod_{i \in \mathbf{S}_j} p_f(i, j)$, where $\mathbf{S}_j$ is the set of CCUs selecting channel j for sensing.

The objective of the CCN is to maximize the expected available time of all the channels while sufficiently protecting the PUs. To this end, we define a channel selection matrix $\mathbf{I} = (I_{i,j})_{N\times K}$, where $I_{i,j} = \{0,1\}$ indicates whether or not CCU

[1] The detection probability and false alarm probability are the probabilities that the PU is detected to be present when it is actually present or absent, respectively.

It is well known that cooperative relaying can improve the transmission rate, save energy, enhance reliability, and so on. Leveraging cooperative relaying, CCUs can improve the PUs' performance by acting as relays to gain spectrum access opportunities as a reward.

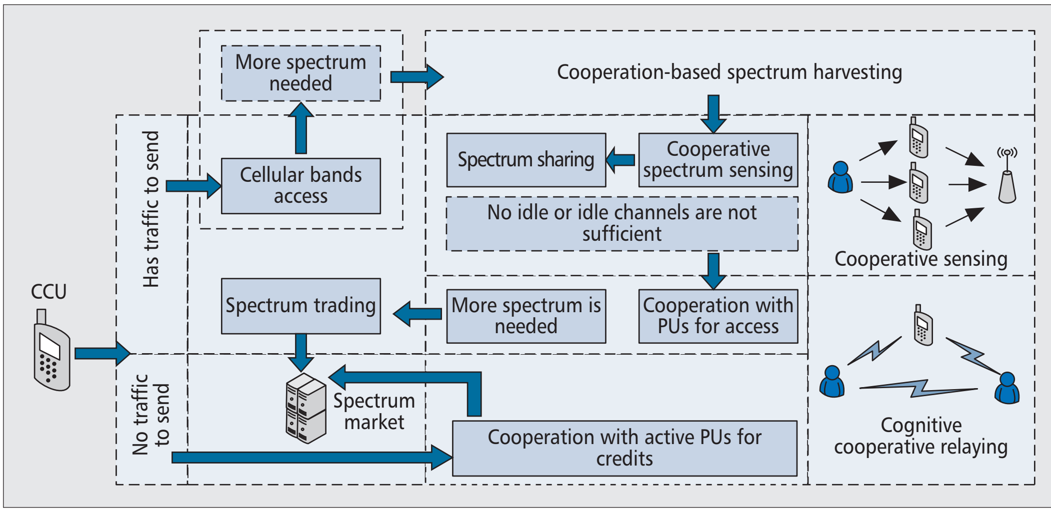


Figure 3. An integrated cooperative framework for spectrum harvesting.

i selects channel j for sensing. **I** can be determined by applying the C-E method of stochastic optimization [15] to maximize the average idle time. The main idea of the C-E method is as follows. In the initial phase, all the CCUs select the sensing channels following the uniform distribution. According to the distribution, a set of channel selection samples are generated, which are utilized to calculate the fitness (i.e., the total expected available time). Then, based on the outcome of all the samples, those channel selection samples with higher fitness will be utilized to update the distribution parameters to produce a "better" sample in the next iteration. By performing this procedure iteratively, it can converge to an optimal or near-optimal deterministic solution (i.e., the channel selection results).

Cooperation with PUs for Access

It is well known that cooperative relaying can improve the transmission rate, save energy, enhance reliability, and so on. Leveraging cooperative relaying, CCUs can improve the PUs' performance by acting as relays to gain spectrum access opportunities as a reward. As shown in Fig. 4b, cooperative communication can be performed in a three-phase manner. Specifically, the CCU acts as a relay to perform cooperative communication with the PU in the first two phases (i.e., αT) to improve the letter's transmission performance. Then the PU grants the third period of time with $(1 - \alpha)T$ to the CCU as a reward. Through cooperation, a "win-win" situation is created, where the PU's performance is improved while the CCU obtains spectrum access rights. For a multi-user scenario, how to perform cooperation between multiple PUs and CCUs needs to be studied.

Suppose that multiple PUs desire to improve their throughput through cooperation, each of which owns a time slot of T. The CCUs can cooperate with the PUs to gain spectrum access opportunities in a three-phase manner. Each PU selects a CCU for cooperation, and amplify-and-forward (AF) mode is adopted. Cooperation between a single PU and a generic CCU is modeled as a Stackelberg game, where the PU acts as the leader and the CCU acts as the follower. The PU's utility is defined as the achievable throughput through cooperation, while the CCU's utility is defined as the throughput achieved in rewarding time minus the energy cost. Note that the energy cost is given by $\lambda_1 P_C(1 - \alpha T/2)$, where λ_1 is the cost rate for transmission power and P_C is the transmission power of the CCU for cooperation. To maximize the utilities, the PU selects the time allocation coefficient α, while the CCU chooses the transmission power P_C. By analyzing the game, the best α and P_C can be determined, which are utilized in the multi-user cooperation scenario.

For the multi-user scenario, a secondary-centric cooperative scheme (SCC) is introduced to coordinate the users' cooperation to maximize the aggregate throughput of all cooperative CCUs. Note that the achievable throughput for a given CCU can be calculated when it cooperates with a certain PU using the cooperation parameters obtained before. For multiple CCUs and PUs, the problem is to select the suitable CCU for each PU to maximize the total throughput of CCUs, which can be represented as a maximum weight bipartite matching. In the bipartite graph, the weight on each edge represents the throughput of the CCU, if the corresponding CCU and PU (represented by vertices) cooperate with each other. Finding the best partner is equivalent to finding the maximum weight matching. To this end, the well-known Hungarian algorithm can be applied to determine the optimal matching (e.g., pairs of CCUs and PUs) such that the sum of weights is maximized [14].

Cooperation with PUs for Credits

When CCUs have no traffic, they can still harvest spectrum in the following way. The CCUs can cooperate with PUs to improve PUs' transmission performance and then request credits as compensation. The earned credits can be utilized by CCUs for spectrum trading in the future when they have traffic. In other words, the CCUs can earn credits through cooperation with PUs and consume credits in the spectrum trading market when needed, as shown in Fig. 4c.

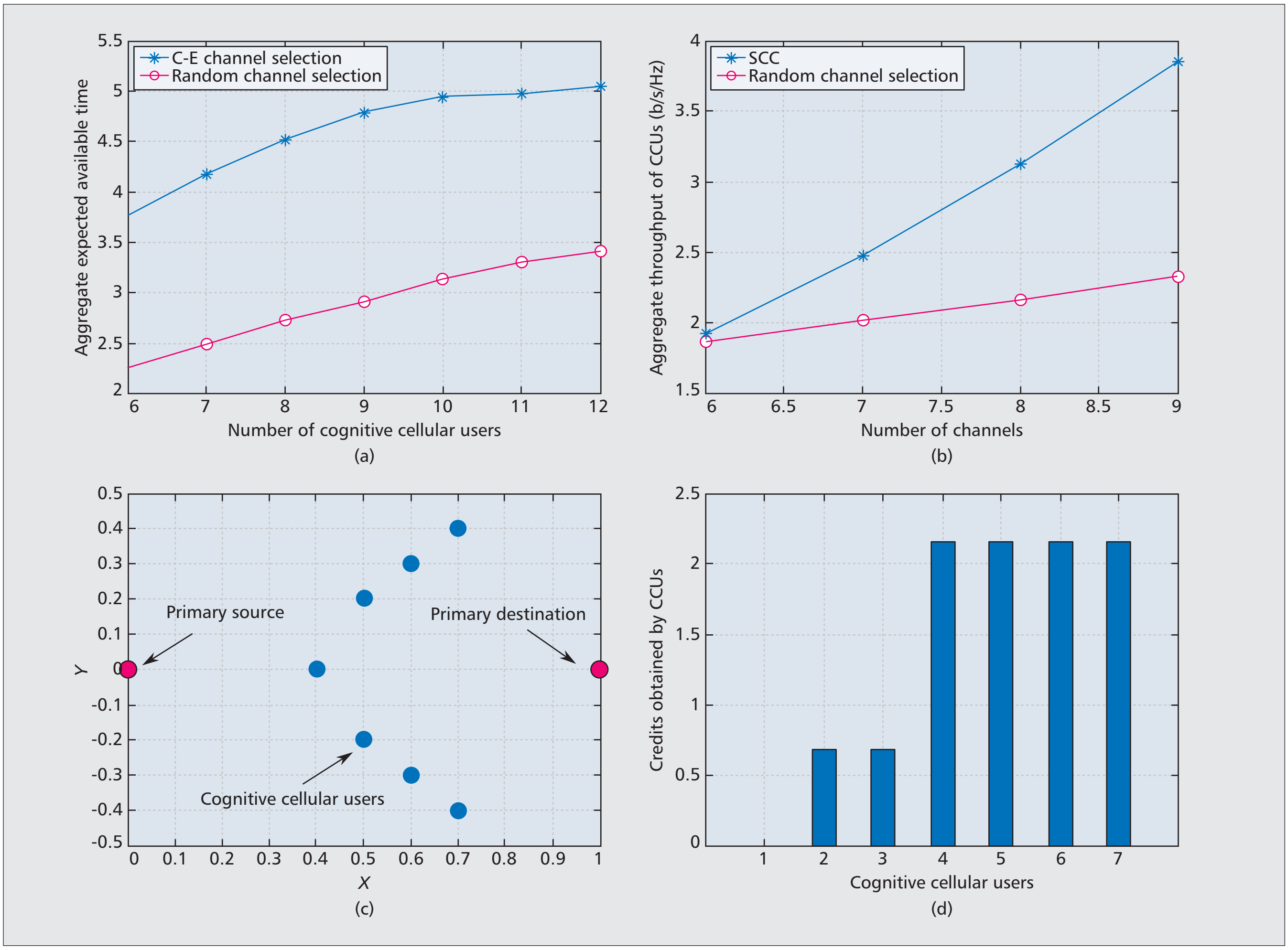


Figure 4. Performance of the cooperation based spectrum harvesting schemes: a) aggregate expected available time vs. number of CCUs; b) aggregate throughput vs. the number of channels; c) network scenario for simulation; d) credits earned by CCUs through cooperation.

Suppose a PU intends to increase its throughput, and a group of CCUs are interested in acting as cooperative relays. To compensate the cooperative CCUs, the PU selects an amount of credits to be shared by CCUs according to their contribution. Since the CCUs act as relays to increase the PUs' throughput, the contribution of CCU i can be approximately given by $P_C^i h_d^i$, where P_C^i is the transmission power of CCU i for cooperation, and h_d^i is the channel gain between CCU i and the primary receiver. Then the utility of the PU is given by $U_p = \lambda_2 R_p - R_m$, where R_p is the achievable throughput through cooperation, λ_2 is the profit per throughput, while R_m is the credits granted to the CCUs. The utility of CCU i is given by

$$U_i = \frac{P_C^i h_d^i}{\sum_{j \subseteq \mathbb{C}} P_C^i h_d^j} R_m - \lambda_1 P_C^i,$$

where $\mathbb{C}$ is the set of cooperative CCUs.

To determine the parameters for cooperation (i.e., the payment of the PU and the transmission power of the CCUs), a two-layer game is used to analyze the negotiation procedure. At the top layer, a buyer-seller game is utilized to model the process by which the PU pays for the service provided by the CCUs. At the bottom layer, for a given payment, the CCUs determine their transmission powers to share the credits in a distributed way, which can be analyzed by a non-cooperative power selection game G. The game G is defined as $G = \{\mathbb{C}, \{\mathbb{S}_i\}, \{U_i\}\}$, where $\mathbb{C}$ is the set of players, U_i is the utility function, and $\mathbb{S}_i$ is the strategy set of CCU i (i.e., the transmission power CCU i can choose).

The two-layer game can be solved by the backward induction method. First, for a given amount of credits, the non-cooperative power selection game is analyzed to get the solution, that is, the Nash equilibrium (NE). The NE strategies are the transmission powers selected by the CCUs. Then, based on the NE strategies, the PU can select the best payment to maximize its utility U_p in the buyer-seller game. After that, all the cooperation parameters can be determined (i.e., the payment and transmission power).

An Integrated Cooperative Framework

With the aforementioned cooperation schemes, an integrated cooperative spectrum harvesting framework is devised to better harvest spectrum

For our future work, we will develop security-aware cooperation-based spectrum harvesting since a partner may be an adversary and misbehave during cooperation. Moreover, how to jointly utilize cellular spectrum and harvested spectrum bands to optimize overall system performance will be investigated.

opportunities, as shown in Fig. 3. When the CCU has no traffic, it can accumulate credits through cooperation with active PUs. Once the CCU has traffic and the QoS requirement cannot be satisfied using only the cellular bands, such as when downloading a high volume of data, additional spectrum bands need to be explored. Then the CCUs can perform cooperative spectrum sensing to detect the unused spectrum for access. If no idle spectrum bands are detected or the idle spectrum bands are not sufficient for QoS requirements, the CCUs can seek spectrum access opportunities through cooperation with the active PUs or keep sensing after a period of time. Otherwise, the CCU can obtain some spectrum from the spectrum trading market by paying credits.

Performance Evaluation

In this section, simulation results are provided to evaluate the performance of the proposed cooperation schemes. In a 2 km × 2 km area, the PUs are randomly located inside the circle with 1 km radius, while the CCUs are randomly distributed outside the circle. The transmission power of PUs, the noise power, the path loss coefficient μ, and the minimum required false alarm probability are set to be 10 W, –80 dB, 3.5, and 10 percent, respectively. Figure 4a shows the aggregate expected available time of the CCN through cooperative spectrum sensing when the number of primary channels is set to 9. It can be seen that the C-E-based user scheduling can achieve a longer available time than the random channel selection algorithm whereby all the CCUs just randomly select a channel to perform spectrum sensing. This is because the C-E algorithm can adaptively adjust the channel selection stochastic policy to increase the aggregate expected available time.

Figure 4b shows the aggregate throughput of the CCN achieved through cognitive cooperative relaying when the number of CCUs is 10. It can be seen that the SCC scheme can obtain a higher aggregate throughput, compared to the random channel selection algorithm whereby the CCUs just randomly select a channel to seek spectrum access opportunities. This is because the SCC scheme selects the best CCUs for each channel to maximize the total throughput.

Figure 4d shows the credits earned by the CCUs through cooperation with the PUs. The primary source and destination pair is placed at (0, 0) and (1 km, 0), respectively, while the location of a set of CCUs without traffic is shown in Fig. 4c. The maximum reward and λ_1 are set to be 20 and 10, respectively. It can be seen that six CCUs can obtain a ceratin amount of credits through relaying the PU's message. This means that the CCUs can accumulate credits through cooperation, which can be utilized for spectrum trading when needed.

Conclusion and Future Work

In this article, we have provided an overview of the cognitive cellular network. Three types of cooperation-based spectrum harvesting schemes have been introduced for different scenarios to better harvest spectrum opportunities. An integrated cooperation-based spectrum harvesting framework has been devised by considering the three cooperation-based schemes to fully explore the spectrum access opportunities. For our future work, we will develop security-aware cooperation-based spectrum harvesting since a partner may be an adversary, which misbehaves during cooperation. Moreover, how to jointly utilize cellular spectrum and harvested spectrum bands to optimize overall system performance will be investigated.

References

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast," White Paper, 2014.

[2] ITU-R Rep. M.2078, "Estimated Spectrum Bandwidth Requirements for the Future Development of IMT-2000 and IMT-Advanced," 2006.

[3] M. Pan *et al.*, "Spectrum Harvesting and Sharing in Multi-Hop CRNs Under Uncertain Spectrum Supply," *IEEE JSAC*, vol. 30, no. 2, 2012, pp. 369–78.

[4] S. Haykin, "Cognitive Radio: Brain-Empowered Wireless Communications," *IEEE JSAC*, vol. 23, no. 2, 2005, pp. 201–20.

[5] N. Zhang *et al.*, "Dynamic Spectrum Access in Multi-Channel Cognitive Radio Networks," *IEEE JSAC*, vol. 32, no. 11, 2014, pp. 2053–64.

[6] Y. Liu *et al.*, "Deploying Cognitive Cellular Networks Under Dynamic Resource Management," *IEEE Wireless Commun.*, vol. 20, no. 2, 2013, pp. 82–88.

[7] W. Wang, G. Yu, and A. Huang, "Cognitive Radio Enhanced Interference Coordination for Femtocell Networks," *IEEE Commun. Mag.*, vol. 51, no. 6, 2013.

[8] I. F. Akyildiz, B. F. Lo, and R. Balakrishnan, "Cooperative Spectrum Sensing in Cognitive Radio Networks: A Survey," *Physical Commun.*, vol. 4, no. 1, 2011, pp. 40–62.

[9] M. Z. Shakir *et al.*, "Green Heterogeneous Small-Cell Networks: Toward Reducing the CO_2 Emissions of Mobile Communications Industry Using Uplink Power Adaptation," *IEEE Commun. Mag.*, vol. 51, no. 6, June 2013, pp. 52–61.

[10] M. Z. Shakir, A. Rao, and M.-S. Alouini, "Generalized Mean Detector for Collaborative Spectrum Sensing," *IEEE Trans. Commun.*, vol. 61, no. 4, 2013, pp. 1242–53.

[11] M. Fitch *et al.*, "Wireless Service Provision in TV White Space with Cognitive Radio Technology: A Telecom Operator's Perspective and Experience," *IEEE Commun. Mag.*, vol. 49, no. 3, 2011, pp. 64–73.

[12] L. Gao *et al.*, "Spectrum Trading in Cognitive Radio Networks: A Contract-Theoretic Modeling Approach," *IEEE JSAC*, vol. 29, no. 4, 2011, pp. 843–55.

[13] Y.-C. Liang *et al.*, "Cognitive Radio Networking and Communications: An Overview," *IEEE Trans. Vehic. Tech.*, vol. 60, no. 7, 2011, pp. 3386–3407.

[14] N. Zhang *et al.*, "Risk-Aware Cooperative Spectrum Access for Multi-Channel Cognitive Radio Networks," *IEEE JSAC*, vol. 32, no. 3, 2014, pp. 516–27.

[15] R. Y. Rubinstein and D. P. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo SIMULation and Machine Learning*, Springer, 2004.

Biographies

Ning Zhang (n35zhang@uwaterloo.ca) earned his Ph.D. degree from the University of Waterloo, Ontario, Canada, in 2015. He received his B.Sc. degree from Beijing Jiaotong University and his M.Sc. degree from Beijing University of Posts and Telecommunications (BUPT), China, in 2007 and 2010, respectively. His current research interests include dynamic spectrum access, 5G, physical layer security, and vehicular networks.

Haibo Zhou (h53zhou@uwaterloo.ca) received his Ph.D. degree in information and communication engineering from Shanghai Jiaotong University, China, in 2014. He is currently a postdoctoral fellow with the Broadband Communications Research (BBCR) Group, University of Waterloo. His current research interests include resource management and protocol design in cognitive radio networks and vehicular networks.

KAN ZHENG (zkan@bupt.edu.cn) received B.S., M.S., and Ph.D. degrees from BUPT in 1996, 2000, and 2005, respectively, where he is currently a professor. He has worked in companies including Siemens and Orange Labs R&D. His current research interests are in the field of wireless communications, with an emphasis on resource allocation in heterogeneous networks, M2M networks, and vehicular communication networks. He holds 12 patents, and he has published more than 200 journal and conference papers and books.

NAN CHENG [S'13] (n5cheng@uwaterloo.ca) (n5cheng@uwaterloo.ca) is currently a Ph.D. candidate in the Department of Electrical and Computer Engineering, University of Waterloo. He received his B.S. and M.S. degrees from Tongji University, China, in 2009 and 2012, respectively. Since 2012, he has been a research assistant in the Broadband Communication Research group in the ECE Department, University of Waterloo. His research interests include vehicular communication networks, cognitive radio networks, and resource allocation in smart grid.

JON W. MARK [M'62, SM'80, F'88, LF'03] (jwmark@uwaterloo.ca) received his Ph.D. degree in electrical engineering from McMaster University in 1970. In September 1970 he joined the Department of Electrical and Computer Engineering, University of Waterloo, where he is currently a Distinguished Professor Emeritus. He served as the Department Chairman during the period July 1984–June 1990. In 1996 he established the Center for Wireless Communications (CWC) at the University of Waterloo and is currently serving as its founding director. He has been on sabbatical leave at the following places: IBM Thomas J. Watson Research Center, Yorktown Heights, New York , as a visiting research scientist (1976–1977); AT&T Bell Laboratories, Murray Hill, New Jersey, as a resident consultant (1982–1983): Laboratoire MASI, Université Pierre et Marie Curie, Paris, France, as an invited professor (1990–1991); and the Department of Electrical Engineering, National University of Singapore, as a visiting professor (1994–1995). He has previously worked in the areas of adaptive equalization, image and video coding, spread spectrum communications, computer communication networks, ATM switch design, and traffic management. His current research interests are in broadband wireless communications, resource and mobility management, and cross-domain interworking. He is a Fellow of the Canadian Academy of Engineering. He is the recipient of the 2000 Canadian Award for Telecommunications Research and the 2000 Award of Merit of the Education Foundation of the Federation of Chinese Canadian Professionals. He has been an Editor of *IEEE Transactions on Communications* (1983–1990), a member of the Inter-Society Steering Committee of *IEEE/ACM Transactions on Networking* (1992–2003), a member of the IEEE Communications Society Awards Committee (1995–1998), an Editor of *Wireless Networks* (1993–2004), and an Associate Editor of *Telecommunication Systems* (1994–2004).

XUEMIN (SHERMAN) SHEN [M'97, SM '02, F'09] (sshen@uwaterloo.ca) received his B.Sc.(1982) degree from Dalian Maritime University, China, and his M.Sc. (1987) and Ph.D. (1990) degrees from Rutgers University, New Jersey, all in electrical engineering. He is a professor and University Research Chair, Department of Electrical and Computer Engineering, University of Waterloo. He was Associate Chair for Graduate Studies from 2004 to 2008. His research focuses on resource management in interconnected wireless/wired networks, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. He is a co-author/editor of six books, and has published more than 600 papers and book chapters in wireless communications and networks, control, and filtering. He has served as Technical Program Committee Chair/Co-Chair for IEEE INFOCOM '14, IEEE VTC '10-Fall, Symposia Chair for IEEE ICC '10, Tutorial Chair for IEEE VTC '11-Spring and IEEE ICC '08, Technical Program Committee Chair for IEEE GLOBECOM '07, General Co-Chair for Chinacom '07 and QShine '06, Chair of the IEEE Communications Society Technical Committees on Wireless Communications, and P2P Communications and Networking. He also serves or has served as Editor-in-Chief of *IEEE Network*, *Peer-to-Peer Networking and Application*, and *IET Communications*; a Founding Area Editor for *IEEE Transactions on Wireless Communications*; an Associate Editor for *IEEE Transactions on Vehicular Technology*, *Computer Networks*, *ACM/Wireless Networks*, and others; and a Guest Editor for *IEEE JSAC*, *IEEE Wireless Communications*, *IEEE Communications Magazine*, *ACM Mobile Networks and Applications*, and others. He received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004, 2007, and 2010 from the University of Waterloo, the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada, and the Distinguished Performance Award in 2002 and 2007 from the Faculty of Engineering, University of Waterloo. He is a registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, and a Distinguished Lecturer of the IEEE Vehicular Technology and Communications Societies.

An Evolution toward Cognitive Cellular Systems: Licensed Shared Access for Network Optimization

Miia Mustonen, Marja Matinmikko, Marko Palola, Seppo Yrjölä, and Kari Horneman

Miia Mustonen, Marja Matinmikko, and Marko Palola are with VTT Technical Research Centre of Finland.

Seppo Yrjölä and Kari Horneman are with Nokia.

Abstract

This article reviews the application of the recent European Licensed Shared Access (LSA) concept for spectrum sharing between a mobile network operator (MNO) and an incumbent user. LSA, as a new area of application of cognitive technology, provides the MNO an opportunity to access new frequency resources on a shared basis. The article presents critical design criteria of LSA from the MNO point of view in order to allow future cognitive cellular networks to efficiently exploit shared spectrum bands. We describe the role of LSA bands in the context of heterogeneous networking, and identify the Long Term Evolution (LTE) and LTE-Advanced enabling technologies that support the introduction of LSA. Such technologies include traffic steering, carrier aggregation, and self-organizing networking. Additionally, we introduce an LSA management unit controlled by the MNO, to be implemented on top of the existing LTE/LTE-Advanced architecture, and we discuss the functionalities required for the optimization and automation of LSA resource management. We also depict the interrelations of the tasks between the LSA management unit and the supporting LTE/LTE-Advanced technologies. Based on the findings in this article, the ongoing cellular system evolution is shown to form a solid base for the introduction of new shared spectrum bands for cognitive cellular systems.

Introduction

The growing demand for mobile data traffic challenges mobile network operators (MNOs) to make the most of their current and future network and spectrum investments. Access to new spectrum bands, as well as the introduction of cognitive features at various network levels, will be critical factors in meeting that demand. As acquiring exclusive spectrum in bands suitable for mobile communications is challenged by incumbent use, the Licensed Shared Access (LSA) concept is gaining significant interest in regulation [1, 2], standardization [3], industry, and research [4]. LSA makes spectrum available for an MNO in a timely manner by allowing the MNO to share spectrum with incumbent users in a fully controlled manner by guaranteeing quality of service (QoS) for both. In LSA, the incumbent user maintains a higher level of usage rights to the spectrum and is allowed to reclaim it, at any location and at any time, in a dynamic manner. Sharing is enabled by built-in mechanisms ythat address the varying spectrum availability in time, location, and frequency dimensions.

The current operational environment for MNOs is characterized as heterogeneous networks, consisting of a portfolio of multiple radio access technologies (RATs), spectrum bands, and cell layers. The coordination and optimization of those resources will be critical factors for the success of MNOs, as depicted in [5]. The integration of shared spectrum bands with variable availability will further complicate this coordination and will call for cognition in the management of resources. On the business side, the use of shared bands offers the potential for new business opportunities both for dominating as well as challenger MNOs [6]. On the technical side, some Long Term Evolution (LTE) and LTE-Advanced enabling technologies support the management and utilization of variable spectrum resources. For example, self-organizing networks (SON) encompass multiple features that simplify operational tasks and optimize the usage of available resources [7]. Since the incumbent user may request the MNO to terminate transmissions in the LSA resources at any time, the integration of shared bands into the cellular network architecture requires an additional logical management unit on top of the existing architecture [3]. Moreover, new coordination protocols are needed to facilitate cooperation between incumbents and MNOs.

This article addresses the inclusion of licensed shared bands into MNOs' heterogeneous networks. The rest of the article is organized as follows. First, spectrum opportunities for MNOs are discussed, focusing on the new LSA concept.

 0163-6804/15/$25.00 © 2015 IEEE

Next, the LSA concept is analyzed from the MNOs' point of view, by discussing design criteria and current LTE/LTE-Advanced features supporting LSA. Finally, a new logical management unit is introduced along with guidelines on the use of LSA resources by future cognitive cellular networks.

Emerging Regulatory Approaches for Efficient Use of Spectrum

In the current spectrum regulatory framework, frequency bands are mainly either exclusively licensed, or license-exempt [1] (see Fig. 1). The license-exempt frequency bands are accessible without QoS guarantees to users operating according to the prevailing regulatory conditions, e.g. transmission power limits. This offers easy access to the spectrum, and thus increases competition and fosters innovation. However, the dedication of a band as license-exempt is irreversible and may limit the future usage of the band. The rights associated with an exclusive spectrum license provide certainty to the MNO, and allow long-term investment planning, but also include obligations such as national coverage in rural areas. These obligations, together with the limited number of licenses and potentially very high license cost, restrict the potential MNOs and shape accordingly the competitive environment. Due to the rapid growth in mobile traffic demand, it is becoming increasingly challenging to find adequate spectrum resources. The International Telecommunication Union (ITU) attempts to address this growth by identifying bands for the use of International Mobile Telecommunications (IMT) in its World Radiocommunication Conferences (WRCs). However, due to existing incumbent usage, not all IMT identified bands can be used exclusively for mobile services in a harmonized manner, either globally or throughout Europe.

As a result, new licensed based sharing concepts are emerging in order to make the IMT bands available to MNOs on a shared basis, by taking advantage of the development of cognitive technologies. These sharing approaches lower the entry barrier to spectrum compared to exclusive licensing, and therefore increase competition by enabling spectrum access to new players as well as to new innovative services. Conversely, via licensing, the sharing schemes maintain the rights of use to the incumbent users and provide certainty to all stakeholders. In the Unites States, activities are ongoing on a spectrum access system (SAS) model, enabling shared use initially for the 3.4–3.6 GHz band [8]. In this model, three different levels of access rights are defined, and a database is used to enable sharing. In Europe, regulatory and standardization activities are ongoing, aiming to enable sharing based on LSA [1]. LSA, as initially introduced by the European Commission (EC), is a general regulatory framework that enables shared access to spectrum for a limited number of licensed users [2]. This article focuses on the LSA concept as a promising path toward cognitive cellular systems.

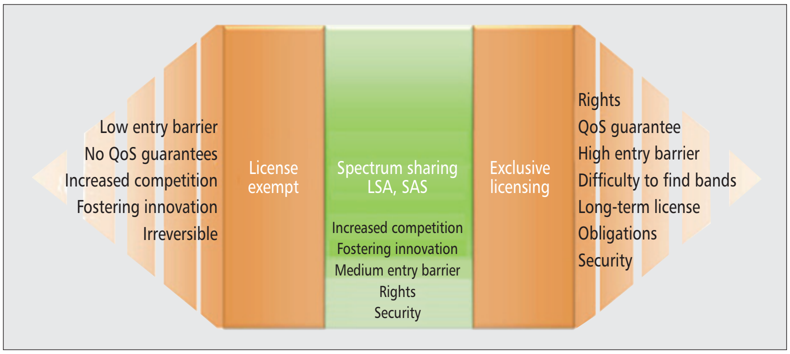


Figure 1. Current and emerging regulatory approaches.

As the first use case, LSA is under investigation in regulation and standardization for enabling access to wireless broadband systems to the 2.3–2.4 GHz band. This is also Band 40 of the 3rd Generation Partnership Project (3GPP) for time-division LTE (TD-LTE). The incumbent usage of the band varies between different European countries. The most common incumbent uses are telemetry and programme making and special events (PMSE), that is to say, wireless cameras. According to an economic estimate, the value of mobile access to the band on a harmonized basis across Europe using LSA ranges from 6.5 to 22 billion euros [9].

In standardization, the European Telecommunications Standards Institute (ETSI) has defined system requirements for LSA [3], and is continuing its work by determining system architecture and high-level procedures. On the spectrum regulatory side, the European Conference of Postal and Telecommunications (CEPT) started developing, in April 2014, harmonized technical conditions for the deployment of wireless broadband services in the 2.3–2.4 GHz band, while ensuring long-term incumbent use. CEPT has developed a report on the technological and regulatory options for sharing between wireless broadband applications and the relevant incumbent services in the band [10], as well as another report providing guidance on the implementation of a sharing framework between wireless broadband and PMSE [11]. Live experiments have been conducted with existing mobile network equipment to prove the feasibility of mobile use on the band using LSA [4]. For European MNOs, this band provides an opportunity for harmonized operations, due to its global identification for IMT. Currently, incumbent usage prevents exclusive licensing to MNOs in a harmonized manner in Europe.

Mobile Network Operators and Licensed Shared Access

LSA is a regulatory approach for fully controlled spectrum sharing, where the incumbent enables the use of its unoccupied spectrum by others, presuming that harmful interference is avoided [1]; the concept and its work flow are depicted in [12]. The availability of LSA resources may be limited either to a certain geographical area or a time frame, but when available guaranteed QoS is provided to the MNO. The resources of the

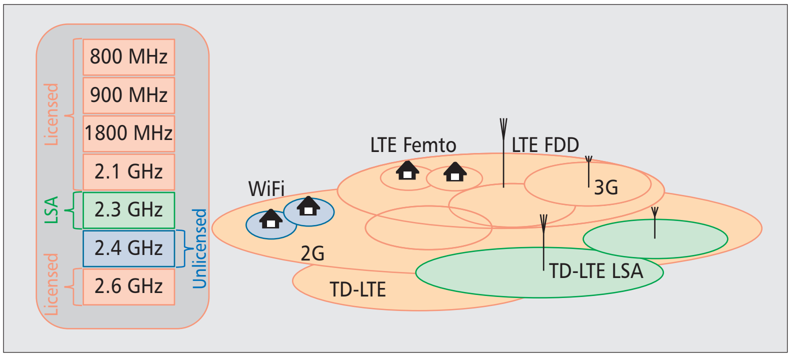


Figure 2. Heterogeneous networks for MNO.

MNO consist of a combination of several spectrum bands and RATs, as illustrated in Fig. 2. The licensed bands illustrated in Fig. 2 have been identified for IMT by the ITU and are available for exclusive usage for mobile services. In these heterogeneous networks, macro cells are complemented by a multitude of smaller (micro, pico, and femto) cells. Coexistence of LTE frequency division duplexing (LTE FDD) and TD-LTE networks, combined with traffic offloading and load balancing between them, allows the MNO to use the available spectrum and other network assets efficiently. The MNO will integrate the LSA resources, with possibly variable availability, as an additional resource layer to its heterogeneous network environment. Using LSA an MNO could potentially access any current, or future, IMT bands that are not exclusively available due to incumbent usage. An example of such a band in Europe is the 3.4–3.6 GHz band, where the incumbent service is fixed satellite service (FSS).

Design Criteria for Adopting LSA

One of the key benefits gained from LSA is that it does not require any changes to the air interfaces or internal procedures of the MNO. This allows the use of existing equipment and simplifies implementation, without the need to change 3GPP standardization. Currently, it is foreseen that for LSA two additional functional components are needed on top of the existing cellular network infrastructure. First, an LSA database is needed for storing and updating the information on LSA spectrum availability, as well as for coordination. Second, a management unit is needed to enable secure access to spectrum and to protect the incumbent user. Those two components form the LSA system [3]. From the MNO perspective, a new interface needs to be standardized for the purpose of information transfer from the LSA database to the LSA management system. In addition to the technical components, the LSA concept is based on an individual LSA license and the related sharing framework negotiated between the key players involved: the MNO, regulator, and incumbent [1]. The following critical factors need to be taken into account in the negotiations:

Security: Appropriate measures are necessary to secure data and interfaces of the MNO network. The information transmitted to and from the LSA system, as well as the stored data, needs to be secured. The coordination between MNOs and the protection of the incumbent user should be done in such a way that the MNO network deployment information is not needed. Due to the confidentiality and sensitivity of this information, the LSA database could be under the purview of the regulator.

Predictable QoS: When and where LSA resources are available, the MNO should be able to provide predictable QoS to its users. In practice, the MNO should be guaranteed interference-free operation, as well as a sufficient level of spectrum availability. For this, collaboration between the incumbent and the MNO is necessary, e.g. information exchange on spectrum availability. The MNO should have exclusive access to a certain LSA resource at a given place and time without competition [1].

Responding to Variable LSA Resource Availability: In case the MNO is requested to evacuate the LSA resource, the MNO should have mechanisms in place to provide QoS for its users, by using alternative networks or resources. The evacuation request(s) should be known as early as possible, for network planning purposes. To achieve this goal, the MNO needs a management unit containing an up-to-date list of available LSA resources. This management unit will be in charge of network planning for the LSA resources, and should therefore be fully under the control of the MNO [14].

Network Confidentiality: The information shared between the key players of the LSA should be limited to a minimum. Details of the MNO network, such as base station (BS) positions or transmit powers, are not required for LSA. Confidentiality applies to the MNO's business sensitive information as well, such as traffic characteristics.

Economies of Scale: As with the support of any new spectrum band, LSA requires investment on the BSs supporting the new bands. Additionally, new management functions are needed to support LSA resources. To provide certainty for the investment, there should be long-term availability of both LSA spectrum and devices for the end users. The LSA band availability for MNOs in a harmonized manner, globally or at least in Europe, would ensure economies of scale. The MNO should be allowed to deploy LSA only when and where needed, based on market demand, to develop a sound business and investment plans.

Careful consideration of the above design criteria will allow the MNO to determine adequate security, usability and value of the LSA resources, as illustrated in Fig. 3. Availability of spectrum in the long term in a widely harmonized manner is related to the value of LSA, and should be considered already during the negotiation phase. Usability and security of LSA are related to the architecture and design of the LSA system. Usability refers to the ability of MNO to integrate a cell into the LSA band as a part of the heterogeneous network environment and to use it in order to provide predictable QoS to its users. Security, in terms of the interfaces and data transmitted via the LSA system, as well as using minimum information are key criteria evaluated by the MNO.

Enabling LTE/LTE-Advanced Features for LSA Support

In the heterogeneous network context, it is important for the MNO that the introduction of new shared spectrum bands causes minimum modifications to their existing network infrastructure. Several features of the current LTE/LTE-Advanced technology facilitate the introduction of LSA bands. These features may be used for optimizing the use of spectrum resources, maximizing the end user experience, or assisting network planning under the LSA regime. Below, these features are identified and discussed in the context of LSA.

Cell Re-Selection Procedures: User equipment (UE) can re-select the serving cell from different RATs and frequencies based on measurements of the signal strength. When an additional cell in an LSA band is added to the network, this re-selection procedure enables the UE to attach to it. If rapid evacuation of the LSA resource is needed, the MNO can lock the affected LSA cells, and terminals will automatically start a cell re-selection procedure. Additionally, in LTE-Advanced different priorities may be set for different cells [13].

Inter-Frequency Handover: Inter-frequency handover is based on inter-frequency measurements performed by the UE and takes place when the signal strength or quality of the serving cell becomes significantly worse than that of another cell in another frequency band [7]. Inter-frequency handovers between cells in LSA bands and exclusive bands enable optimization of the data rate and user experience. With adequate time reserved for LSA resource evacuation, the MNO may gradually lower the transmit power of the eNodeB (eNB) before delivering the shutdown command (graceful shutdown) in order to allow UEs to conduct inter-frequency handover in a controlled manner and without experiencing connection breaks.

Traffic Steering: Traffic steering is a self-optimization feature referring to the ability to steer traffic to the most suitable cell layer, frequency, and RAT within any network governed by the MNO. Steering is done in order to meet a set of optimization criteria such as network capacity, power consumption, or user experience. Traffic steering is emphasized in LTE networks, which are deployed over multiple frequency layers and coexist with other RATs [7]. The use of LSA resources needs to be supported by traffic steering functions due to their varying availability. However, modifications to the existing procedures might be needed, as discussed in [14].

Carrier Aggregation (CA): CA enables scalable expansion of bandwidth through utilization of radio resources across multiple component carriers [15]. CA provides a wider effective bandwidth to the end user by aggregating component carriers either on the same or different bands. Using CA, a carrier on the LSA band may be used together with a carrier on a licensed band in order to create another cell with almost the same coverage in order to support macro cell capacity, or a cell with smaller coverage to locally improve data rate and throughput. Using CA, an MNO can use LSA resources to provide additional capacity to its users, without the risk of connection break. As a special case of CA termed supplemental downlink (SDL), unpaired spectrum is used for enhancing downlink capacity by bonding the downlink of an FDD channel with the supplemental downlink of a TDD channel. Since LSA is currently considered for TD-LTE bands, SDL might become one of the key enabling technologies for LSA.

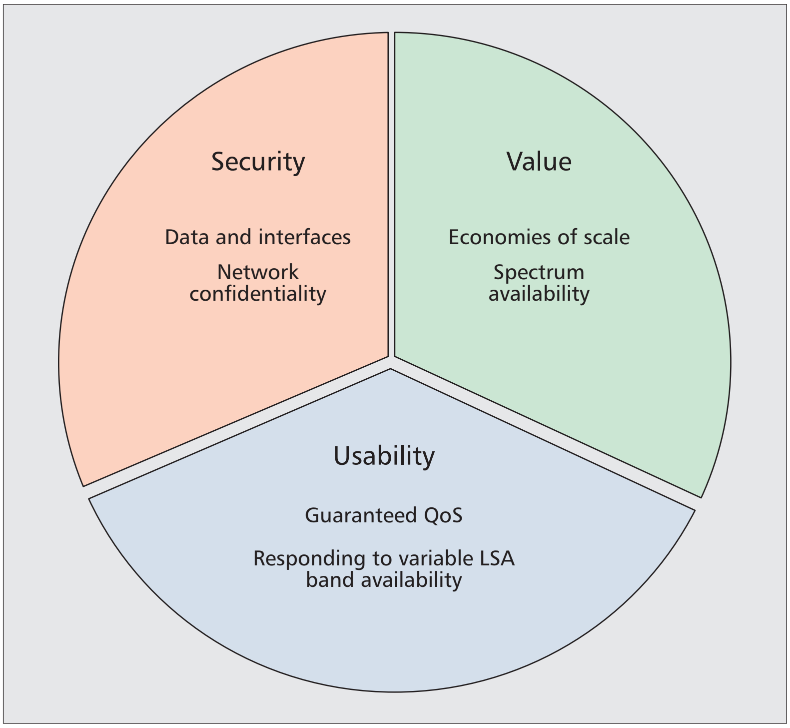


Figure 3. The design criteria for LSA.

Load Balancing: Load balancing is one of the LTE self-optimization features that aims to even out the load generated across the network by moving users from one cell to another through handovers [14]. eNBs on LSA bands can be used as an additional network layer, providing more capacity for wireless broadband users to balance the load. Variable availability of LSA resources leads to considerations on which kind of users (in terms of requested data rates, mobility, etc.) can be best served and are least affected by possible evacuation.

Self-Configuration: Self-configuration denotes the process of bringing a new network element, or part of it, into service with minimum manual intervention [7]. Self-configuration includes both connection establishment and acquiring radio configuration parameters to adapt to the current state of the network deployment. With dynamic radio configuration, the key radio configuration parameters of the new cell and its neighbors can be configured correctly on the fly. Such flexible deployment is particularly useful when considering that the LSA cells may need to be switched off due to evacuation requests, and then re-configured when they are switched on again.

Active Antenna Systems (AAS), Vertical Sectorization, and Beamforming: AAS can perform network coverage and capacity optimization by

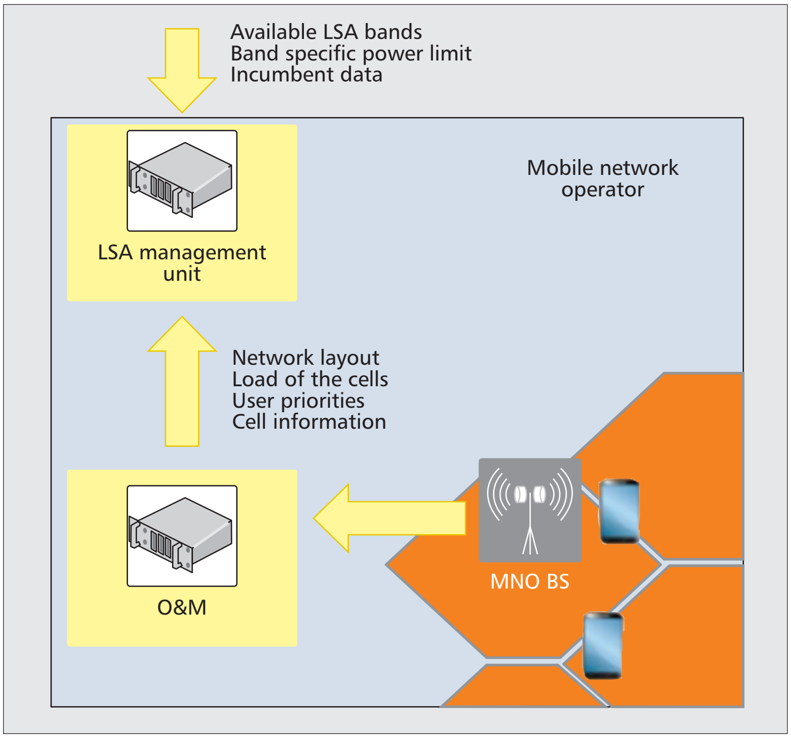


Figure 4. Information required by the LSA management unit.

changing antenna parameters such as tilt, azimuth, or beam shape [7]. In LSA, exclusion zones and potential interference may be reduced by using AAS technologies to isolate accurately and flexibly the zone needed for incumbent protection. Using AAS, the MNO can automatically adjust the size of the cells, in order to better utilize the non-static LSA resource, and to serve non-uniform demand from users. Dynamic steering of beams enables precise capacity distribution where and when needed. Additionally, AAS could be used for sharing radio sites between the incumbent user and the MNO. For example, the incumbent user can use TD-LTE network and radio equipment to deliver its service.

Dual Connectivity: Dual connectivity developed in the 3GPP standardization allows the UE to be connected to two eNBs at the same time. It is mainly intended to be used by small cells under macro cell control. The macro cell acts as an aggregation point with the control channel. The UE could be connected to both the existing LTE FDD cell and the TD-LTE LSA cell at the same time, using dual connectivity.

Use of LSA Resources by the MNO

For an MNO with already deployed mobile networks, the most straightforward way of using LSA spectrum is to create a new resource using macro and/or small cells, and integrate the new resource as part of the layered heterogeneous (2G, 3G, and LTE) network environment. However, the variable LSA resource availability requires fast reconfiguration of the network. The optimization and automation of LSA resource management requires an additional logical management unit (also called the LSA controller in [1]) for network management on top of the existing LTE/LTE-Advanced architecture, as discussed already in [14]. The tasks of the LSA management unit are further discussed in the following section. In order to minimize the implementation cost and complexity of the LSA management unit, the same unit should be used for multiple layers and multi-vendor networks.

Input to LSA Management Unit

To configure the radio and network parameters of the MNO network, the LSA management unit needs to take into account the entire network layout as well as the eNB interactions. The configuration requires knowledge of the MNO's radio access network and potentially business sensitive information. It is therefore foreseen that the LSA management unit needs to fully reside inside the MNO network domain (as also concluded in [14]). Information required by the LSA management unit can be divided into network internal and external information, as depicted in Fig. 4. The external information is provided to the LSA management unit via the LSA database.

The network internal information includes network layout and cell information, such as location, transmit power, direction, height, and tilt angle for each antenna. Furthermore, the cell load in terms of the existing users, their characteristics, and their priorities needs to be taken into account when making load balancing decisions for LSA resources. The network external information includes available LSA resources in a certain geographical area and time frame, as well as information on the radio characteristics of the incumbent user such as power, type, indoor/outdoor deployments, etc. Additionally, it includes the band-specific power limits. A new interface is needed to access the external information. The interface along with the technical information available to the LSA management unit should be considered in the relevant standardization for interoperability and harmonized market [1]. Based on this data and information related to geo-location and LTE signal propagation, the LSA management unit will calculate the interference caused by LTE cells to the incumbent user.

Tasks of the LSA Management Unit

The overall goal of the LSA management unit is to maximize the efficiency of the LSA resource usage, as well as to avoid harmful interference to the incumbent. The different LSA management unit functionalities can be divided into different phases. Those, along with the relevant enabling LTE/LTE-Advanced technologies, are depicted in Fig. 5 and discussed below.

Sharing Framework and Licensing

The first phase of the LSA is the negotiation of the LSA license and the sharing framework between the incumbent(s), administration, and prospective LSA licensees. In this phase, terms and conditions of the LSA license and the sharing framework are stored in the LSA management unit and reported to the O&M as necessary. This information contains, for example, license duration, geographical areas, spectrum bands, considered policies, etc., and it remains stable throughout the LSA license duration. Since the negotiation phase is unique for the LSA concept, supporting techniques should be considered in future standardization.

Network Planning and Configuration Management

In the MNO network planning and configuration phase, the LSA management unit needs to identify cells that are subject to the LSA license. Additionally, network pre-configuration functions may be tested and validated in this phase, in order to minimize the evacuation time and the risk of harmful interference to the incumbent during LSA operations. Self-configuration mechanisms may be used to minimize the manual intervention in the configuration of the network for LSA.

Phase	LSA management unit tasks	Enabling LTE/LTE-A technologies
Sharing framework and licensing	Obtain, store and report information on the sharing framework and license conditions	-
Network planning and configuration management	Identify and configure cells for LSA band	Self-configuration
Activation/ de-activation of the cells on the LSA band	Cell activation/de-activation, interference estimation, maximize/maintain QoS, report on LSA resource usage	Cell re-selection, frequency handover, traffic steering, self-configuration, AAS
Use of the LSA resources	Optimize network use for LSA band, maximize QoS	Cell re-selection, frequency handover, traffic steering, carrier aggregation, load balancing, AAS, CA, dual connectivity

Figure 5. The tasks of the LSA management unit and enabling techniques.

Activation and De-Activation of the Cells on the LSA Band

During the cell activation and de-activation phases, the management unit reports changes of the LSA resource availability to the O&M, as well as the LSA resource usage or executed evacuation to the incumbent user. It also carries out interference estimation in order to optimize the size of the required area for the evacuation and modify the LSA network configurations accordingly, as to allow the incumbent to operate in the area. In case the incumbent activity is highly dynamic, the network needs to be re-configured frequently and the importance of self-configuration mechanisms is emphasized. AAS techniques may be used for minimizing interference and maximizing LSA resource availability in space and time. When a cell is de-activated, existing access procedures defined in 3GPP standardization prevent UEs from transmitting in the LSA band. The LSA management unit makes sure that customer satisfaction and QoS are maintained, regardless of de-activation. The tasks of the LSA management unit and enabling techniquesare maintained, regardless of de-activation. This is supported by existing LTE/LTE-Advanced techniques such as cell re-selection, frequency handover, and traffic steering to other network resources. As the LSA cell is activated, the same LTE/LTE-Advanced techniques will be used for distributing the traffic between cells, thus maximizing the QoS of the users.

Use of the LSA Resources

While the MNO uses the LSA resources, the LSA management unit optimizes the network use and maximizes customer satisfaction in the LSA band, in cooperation with the common network optimization, in a sub-function manner. Network optimization denotes the optimization of the traffic distribution across all bands and networks, and the existing LTE/LTE-Advanced tools include traffic steering, offloading, load balancing, and handover decisions. In addition, the LSA management unit can assist this optimization, e.g. by controlling the user experience by prioritizing and offering different service levels in response to different customer needs. AAS techniques can be used by the MNO to optimize the network, by adjusting the size and position of the cells, in order to distribute capacity where needed. One possible usage strategy of the LSA resource is through CA coupled with a carrier on a licensed band. Using CA, additional capacity can be provided to the users when the LSA resource is available, with minimum impact from evacuation requests. The impact can also be minimized if UEs are connected to both FDD LTE cell on a licensed band and TD-LTE cell on the LSA band at the same time, by using dual connectivity.

As presented in this article, several LTE/LTE-Advanced techniques support and facilitate the introduction of LSA resources in an MNO network. However, there are still issues that need to be considered, especially in terms of interference management and LSA resource evacuation. An incumbent with very dynamic spectrum usage and frequent evacuation requests may pose a challenge to existing techniques. Additionally, the current equipment and resource management procedures set a minimum bound to the evacuation time, below which changes in standardization would be required. In the live implementation of LSA presented in [10], evacuation times of approximately 30 seconds were reported when using existing commercial LTE equipment (eNBs, core network, and O&M) combined with a research platform encompassing the LSA database and management unit. Extending the existing techniques and optimizing the LSA band use in a heterogeneous network environment with a large number of cells will add to the complexity of the required algorithms.

Concluding Remarks

The Licensed Shared Access (LSA) concept offers a new licensing option for mobile network operators (MNOs) to complement exclusively licensed bands. LSA allows the MNO to take into use new shared frequency resources that would not be otherwise accessible due to incumbent usage. These additional resources can be

LSA allows the MNO to take into use new shared frequency resources that would not be otherwise accessible due to incumbent usage. These additional resources can be integrated as part of the existing network system in order to provide additional coverage and capacity.

integrated as part of the existing network system in order to provide additional coverage and capacity. In LSA, the incumbent user maintains higher spectrum usage rights and may therefore request the MNO to terminate transmissions on LSA resources at any time and in any geographical area, even dynamically. The LSA resources provide additional capacity to MNOs, allowing them to respond to the growing traffic demand in a timely manner. In this article we also reviewed the Long Term Evolution (LTE) and LTE-Advanced features that enable and support the integration of LSA cells into an MNO network. We introduced the LSA management unit, an additional logical unit for the optimization and automation of LSA resource management. Additionally, we identified tasks of this unit, in concert with supporting LTE/LTE-Advanced technologies. Considering the detailed information of the MNO network required to perform these tasks, we foresee that this unit should be under the purview of the MNO. The article has focused on the LSA concept from the MNO's perspective; however, techniques and tools are needed to manage the full framework of LSA in a secure and efficient manner. Additionally, sharing rules should be defined to enhance business potential and incentives to all parties involved.

Acknowledgment

The authors would like to acknowledge the CORE+ project consortium: VTT, the University of Oulu, Centria University of Applied Sciences, Nokia, PehuTec, EXFO, Elektrobit, Anite, Rugged Tooling, Finnish Defence Forces, FICORA, and Tekes.

References

[1] "Licensed Shared Access (LSA)," ECC Report 205, Feb. 2014.
[2] "RSPG Opinion on Licensed Shared Access," Radio Spectrum Policy Group, RSPG13-538, Nov. 2013.
[3] "System Requirements for Operation of Mobile Broadband Systems in the 2300 MHz – 2400 MHz band under Licensed Shared Access (LSA)," ETSI TS 103 154, v1.1.1, Oct. 2014.
[4] M. Matinmikko *et al.*, "Cognitive Radio Trial Environment: First Live Authorised Shared Access (ASA) based Spectrum Sharing Demonstration," *IEEE Vehic. Tech. Mag.,* vol. 8, no. 3, Sept. 2013, pp. 30–37.
[5] B. Soret *et al.*, "Multicell Cooperation for LTE-Advanced Heterogeneous Network Scenarios," *IEEE Wireless Commun. Mag.,* vol. 20, no. 1, Feb. 2013, pp. 27–34.
[6] P. Ahokangas *et al.*, "Business Models for Mobile Network Operators in Licensed Shared Access (LSA)," *IEEE Int'l. Symp. Dynamic Spectrum Access Networks*, McLean, VA, 2014, pp 263-270.
[7] S. Hämäläinen *et al.*, *LTE Self-Organizing Networks (SON)*, Wiley 2012.
[8] "Realizing the Full Potential of Government-Held Spectrum to Spur Economic Growth," President's Council of Advisors on Science and Technology (PCAST) Report, July 2012.
[9] T. Lavander *et al.*, "The economic benefits of LSA in 2.3 GHz in Europe," Rep. Plum Consulting, Dec. 2013.
[10] "Technological and Regulatory Options Facilitating Sharing between Wireless Broadband Applications (WBB) and the relevant Incumbent Service/Application in the 2.3 GHz Band," CEPT Report 56, March 2015.
[11] "Technical and Regulatory Solutions for the Sharing Framework between WBB and PMSE within 2300–2400 MHz," CEPT Report 58 (draft), Jan. 2015.
[12] M. Matinmikko *et al.*, "Spectrum Sharing Using Licensed Shared Access (LSA): The Concept and Its Work Flow for LTE-Advanced Networks," *IEEE Wireless Commun. Mag.*, vol. 21, no. 2, Apr. 2014, pp. 72–79.
[13] "User Equipment (UE) Procedures in Idle Mode," 3GPP TS 36.304, v11.5.0, Sept. 2013.
[14] M. Mustonen *et al.*, "Cellular Architecture Enhancement for supporting Licensed Shared Access (LSA) Concept," *IEEE Wireless Commun. Mag.*, vol. 21, no. 3, June 2014, pp. 37–43.
[15] M. Iwamura *et al.*, "Carrier Aggregation Framework in 3GPP LTE-Advanced," *IEEE Commun. Mag.*, vol. 48, no. 8, Aug. 2010, pp. 60–67.

Biographies

MIIA MUSTONEN (miia.mustonen@vtt.fi) received her M.Sc. degree in electrical engineering from the University of Oulu in 2005. She has been working as a research scientist at the VTT Technical Research Centre of Finland in Oulu, Finland, since 2006. She has been actively involved in regulatory activities on IMT-Advanced, cognitive radio systems (CRS), and LSA in both European and international regulatory bodies. Her current research interests include CRS and spectrum sharing models, especially LSA.

MARJA MATINMIKKO (Marja.Matinmikko@vtt.fi) received her M.Sc. degree in industrial engineering and management and D.Sc. degree in telecommunications from the University of Oulu. She is working as a senior scientist and project manager at the VTT Technical Research Centre of Finland. She is leading a national project consortium on Cognitive Radio Trial Environments (CORE) that demonstrated the world's first live LSA trial in Finland in 2013. Her research interests include technical, trialing, regulatory, and business aspects of spectrum sharing.

MARKO PALOLA (marko.palola@vtt.fi) is a senior scientist at the VTT Technical Research Centre of Finland and the key designer of the CORE+ trial environment. He holds a M.Sc. degree in information technology from the University of Oulu, Finland. His research interests include intelligent decision making in networks, LTE SON, and QoS.

SEPPO YRJÖLÄ (seppo.yrjola@nokia.com) is currently a principal innovator at the CTO office of Nokia Networks (NET). He holds a M.Sc. degree in electrical engineering from the University of Oulu and has done his postgraduate studies in the area of telecommunication and radio technology. His current focus is innovating in the area of future radios with cognitive radio business opportunities and potential disruptions as an important aspect to scout.

KARI VEIKKO HORNEMAN (kari.horneman@nokia.com) received his M.Sc.E.E. degree in 1980 from the University of Oulu, Finland. He is currently with Nokia Networks, Oulu, where he is working as a program manager for research programs. His current work is related to spectrum sharing and 5G radio networks. He has been involved in the work on energy efficiency for heterogeneous networks, cognitive radio networks, and relay networks, mainly for LTE, LTE-A, and beyond.

Network Economics in Cognitive Networks

Chunxiao Jiang, Yan Chen, K. J. Ray Liu, and Yong Ren

ABSTRACT

Cellular networks are confronted with unprecedented pressure to accommodate the dramatically increasing demands of mobile Internet. This phenomenon is principally due to the restricted spectrum resource and low-efficiency spectrum management. Recently, cognitive radio technology has been corroborated to significantly improve spectrum utilization efficiency. In order to enhance the spectrum management efficiency in cellular networks, the concept of "cognitive cellular networks" was introduced. It has been realized that the network architecture, algorithms, and protocols cannot be simply designed without considering the socio-economic aspects involved. In this article, we consider the economic issues in cognitive cellular networks from the perspectives of game theoretic modeling and mechanism design. Specifically, both the simultaneous and sequential dynamic spectrum access games are investigated, respectively. The mechanism design related issues (i.e., spectrum auction, pricing, and contract) to ensure desired outcomes are introduced with detailed applications in cognitive cellular networks. The goal of this article is to provide an overview to understand the motivation, problem formulation, methodology, and solutions of the economic issues in cognitive cellular networks.

INTRODUCTION

The last decade has witnessed gigantic development in cellular communication networks, from third generation (3G) — wideband code-division multiple access (W-CDMA), time-division code-division multiple access (TD-CDMA) — 3.5G — Long Term Evolution (LTE), CDMA2000 — to 4G (LTE-Advanced) and ongoing 5G standardization. Meanwhile, mobile applications and mobile social networks have emerged with this trend, leading to a phenomenon of ubiquitous mobile Internet that has penetrated into our daily lives. The faster wireless cellular network access comes mainly at the expense of wider communication bands and more spectrum resource. Until recently, desired improvement of service quality and channel capacity in wireless networks has been severely limited by spectrum resource, which has triggered considerable research activities seeking new communications and networking paradigms.

To mitigate the problem of spectrum resource limitation, cognitive radio technology was proposed in a concept of dynamic spectrum access (DSA), where users can intelligently and efficiently share the spectrum resource [1]. The essence of DSA technology lies in the fact that devices with cognitive capability, called secondary users (SUs), can dynamically search and utilize the licensed spectrum resource not occupied by licensed users, usually called primary users (PUs). Recently, researchers have proposed to introduce DSA into traditional cellular networks, generating the concept of a cognitive cellular network (CCN). As shown in Fig. 1, there are generally two kinds of operation modes in a CCN. One is that the cognitive device-to-device (D2D) SUs or ad hoc SUs attempt to dynamically utilize the licensed spectrum of primary cellular networks, as shown on the left of Fig. 1, while the other is that secondary small cell (e.g., femtocell, picocell) networks dynamically share the licensed spectrum with primary cellular networks but with lower priority, as shown in the right of Fig. 1b.

Currently, most existing works about CCNs have focused on various technical issues, including spectrum sensing to identify the available spectrum resource, interference management between primary cellular networks and secondary small cell networks (e.g., power control, offloading), and dynamic resource allocation [2]. However, the economic issues in CCNs have not been well understood, which is essential when it comes to the practical deployment of CCNs. Past history has shown that implementations of successful technologies not only rely on good engineering solutions, but also need to take the socio-economic aspects into account, which is especially the case for CCNs. In this article, we aim to study the network economic issues in CCNs from the perspectives of game theoretic modeling and mechanism design, to reveal the fundamental problems and corresponding enabling techniques. In the following, we first study the SUs' interactions when they simultane-

Chunxiao Jiang and Yong Ren are with Tsinghua University.

Yan Chen and K. J. Ray Liu are with the University of Maryland.

 0163-6804/15/$25.00

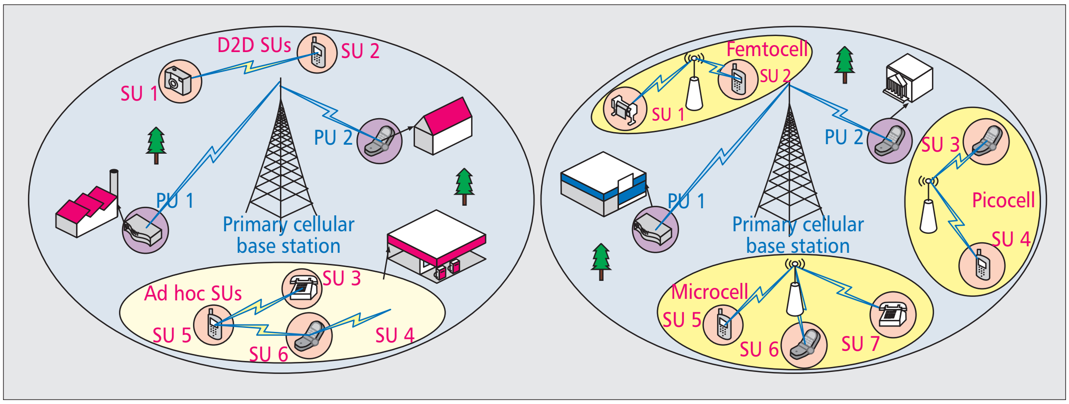


Figure 1. Two kinds of operation modes in CCNs.

Bayesian game is characterized by the uncertain player's type, which embodies any information that is not common knowledge to all players and is relevant to the players' decision making. Although each player is unknown about the exact types of others, he/she has a belief that illustrates the distribution of other players' utility functions.

ously access the primary cellular spectrum. Five simultaneous game models, including Bayesian, repeated, bargaining, coalition, and evolutionary games, are introduced with corresponding scenarios in CCNs. Then we further discuss the sequential DSA games considering two sequential decision structures in CCNs: sequential decision making between PUs and SUs, and sequential decision making among SUs. For the mechanism design issues, spectrum contract, auction, and pricing are introduced, respectively. Overall, this article provides a set of general models and methods for users' socio-economic behaviors and interactions analysis in CCNs.

Simultaneous Dynamic Spectrum Access Games in CCNs

Most existing licensed cellular networks (e.g., LTE, TD-CDMA, WiMAX) are both time/frequency-division-based systems, where all the devices are synchronized with the base station and coordinated through time slots. In such a case, if the SUs, whether ad hoc secondary devices or secondary small cells, intend to dynamically utilize the spectrum of primary cellular networks, they are required to synchronize with the network time slots in order to better control interference. A common operating mode is that the SUs perform spectrum sensing at the beginning of each time slot, and vacate the occupied spectrum at the end of the slot. Under such conditions, the SUs access the primary spectrum simultaneously, which leads to a practical problem of how an SU competes or cooperates with others. Meanwhile, one SU usually has no knowledge of other SUs' private information, such as channel state information and transmission power control policy, making the problem even more challenging.

A simultaneous game is an ideal tool to solve the aforementioned simultaneous spectrum access problem. It refers to a game where all players make a decision simultaneously without knowledge of the actions chosen by other players. In the literature, simultaneous game models have been extensively applied to SUs' spectrum access in CCNs, including Bayesian game [1], repeated game [3], evolutionary game [4], Nash bargaining [5], coalition game [1], and so on. Among those models, the first three are noncooperative game models, while the last two are cooperative. The Bayesian game is characterized by an uncertain player type, which embodies any information that is not common knowledge to all players and is relevant to the players' decision making. Although each player does not know about the exact types of others, he/she has a belief that illustrates the distribution of other players' utility functions. The authors in [1] summarized cognitive radio jamming games under uncertainty, where the transmitter was uninformed of the jammer's exact activities and had to make an expectation over all possibilities. Unlike the one-shot game model, Zhou *et al.* studied the long-run repeated interactions among SUs in a CCN using a repeated game [3]. This game is featured by a repeated decision making and punishment stage, which ensures that all players cannot deviate from the designed Nash equilibrium (NE). In [3], the transmission power level was considered as an SU's strategy, while the utility function was defined as an SU's long-term summation payoff discounted over time. Moreover, a reinforcement learning-based power control algorithm was designed for SUs to converge to the desired NE.

When all SUs in a CCN are fully selfish and uncooperative, NE sometimes becomes extremely inefficient. For a simple instance, in the power control game, all SUs adopting the highest transmission power is an NE [3]. Under such a circumstance, cooperation among SUs in a CCN can improve the system efficiency to a large extent, while the practical problem is how to encourage/stimulate them to be cooperative. A cooperative game can solve this problem well by bargaining or forming coalition among players. In [5], the Nash bargaining solution (NBS) was applied to analyze the cooperation between the PUs (licensed cellular network) and unlicensed ad hoc SUs in a CCN. The NBS can simultaneously satisfy *Pareto optimality, symmetry, fairness, independence of irrelevance alternative,* and *independence of linear transformations*. The authors in [4] derived the NBS by maximizing the product of PUs' utility minus its acceptable minimal utility and SUs' utility minus its minimal utility, where the utility was defined as the achievable throughput. Another typical kind of cooperative game is a coalition game, which focuses on how

Game models	Coop?	Equilibrium	Key features	App in CCN
Bayesian game	No	Bayesian NE	• Players' types • Bayesian expected utility	[1]
Repeated game	No	Repeated NE	• Long-term expected utility • Punishment stage	[3]
Bargaining game	Yes	Nash bargaining solution	• Pareto optimality, fairness • Utility production maximization	[5]
Coalition game	Yes	Nash stable coalitions	Coalition formation algorithm	[1]
Evolutionary game	No	Evolutionarily stable strategy	Replicator dynamics	[4]

Table 1. Comparison of different simultaneous game models.

In the dynamic Chinese restaurant game model, the SUs are assumed to arrive at the primary cellular network by Bernoulli process with certain probability. They sequentially sense and estimate the primary spectrum in a Bayesian manner, and sequentially access the available spectrum based on a multidimensional Markov decision process.

to formulate stable coalitions among a group of players. It has been applied to resource allocation in secure cognitive femtocells [1], where the utility function was defined as the gain in terms of secrecy rate minus the cost of information exchange. A distributed coalition formation algorithm was designed to find stable secure coalitions, and the convergence analysis was also conducted. In Table 1, we summarize and compare all the simultaneous game models.

The NE derived by the aforementioned models can only prevent one SU's deviation, while it is possible that multiple SUs deviate from the NE in order to selfishly attain more utility. When it comes to a small group of SUs' deviations, evolutionary game theory (EGT) can address this problem well. Different from the traditional game models that focus on the property of static NE, EGT emphasizes the evolutionary dynamics and stability of the whole population's strategies, which is called an evolutionarily stable strategy (ESS). In [4], we studied the joint spectrum sensing and access problem in CCNs using EGT. On one hand, when only a few SUs contribute to spectrum sensing, the false alarm probability is relatively high, resulting in a high interference probability and low throughput. On the other hand, when many SUs access the primary channel, the channel becomes very crowded, and little throughput can be obtained by an individual SU. Through analyzing the dynamics of SUs' sensing and access strategies, we derived the ESS from which no one can deviate, as well as a distributed learning algorithm for the SUs to converge to the ESS. From the simulation results in Fig. 2, we can see that the SUs with the designed learning algorithm can quickly abandon the initial undesired strategy, that is, only 10 percent SUs sense, while 90 percent access the spectrum. The system finally converges to different ESSs under different settings of the reward to the SUs who only contribute to sense the spectrum but not access, as shown on the top and in the middle of Fig. 2. In a practical network, the reward can be the credit of the SUs, or a period of network access time free of charge. It can be seen that a higher reward can attract more SUs to participate in spectrum sensing. In addition, the bottom of Fig. 2 shows that when a small group of SUs deviate from the ESS, the system can return back to the ESS quickly after the perturbation.

Sequential Dynamic Spectrum Access Games in CCNs

In a CCN where ad hoc SUs sense and access the licensed cellular spectrum in a distributed manner, different SUs may arrive at or leave the network at different time slots. In such a case, one SU can observe the previous SUs' spectrum access behaviors as well as their released spectrum sensing results, which can be utilized to better understand the availability of the licensed cellular spectrum. Apparently, when considering multiple SUs' access in different time slots or their asynchronous access in one time slot, it becomes a sequential spectrum access problem. Apart from this sequential spectrum access among SUs, the sequential characteristic also exists between the PUs and SUs. In the CCN, the primary cellular networks always have highest priority as leaders, while the SUs, whether ad hoc secondary devices or secondary femtocells, are on the follower side. A common interaction mode is that the primary cellular network first determines its strategies (claiming the maximum tolerable interference level, announcing the amount of spectrum that can be utilized, etc.). Then, according to the PUs' actions, the SUs can make an optimal response on how to access the spectra. To summarize, the sequential dynamic spectrum access problems in CCN have two scenarios:

- PUs first propose the regulations of spectrum utilization, and then SUs decide how to access the spectrum according to the regulations.
- SUs sequentially determine how to sense and access the licensed cellular spectrum.

To address such a sequential spectrum access problems in CCNs, a sequential game is an effective tool to model and analyze the sequential decision making structure between PUs and SUs, and that among SUs. Sequential refers to games where one player makes his/her decision after the others, where an important feature is that later players have some information about the first users' strategies. Due to the sequential

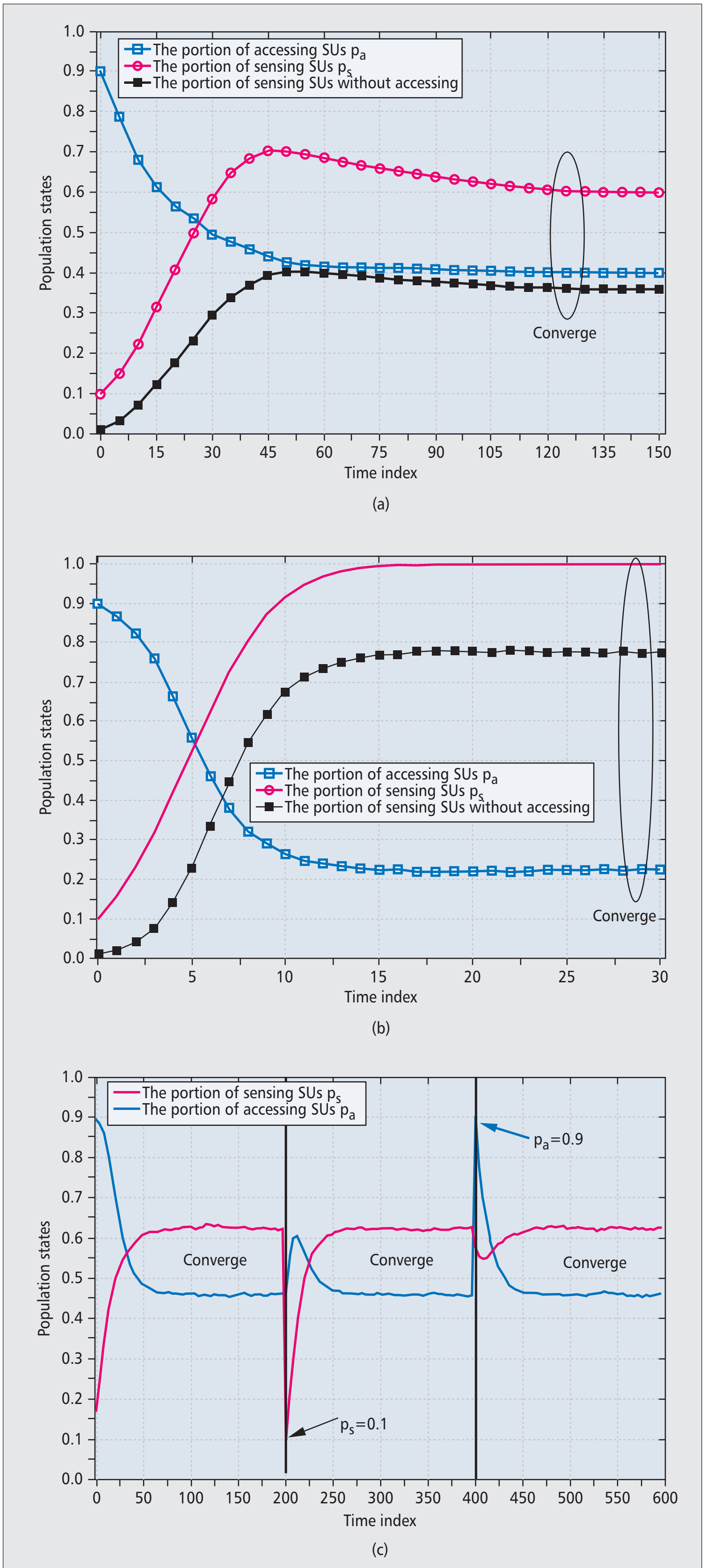


Figure 2. Evolutionarily stable strategy under different reward settings (the reward to the SUs who only contribute to sense the spectrum but not access). Top: reward = 40; middle: reward = 100; bottom: reward = 50.

characteristic, the game is usually represented by an extensive form or decision tree, and the corresponding NE can be found using a backward induction method. For the aforementioned first scenario (i.e., sequential decision making between PUs and SUs), the Stackelberg game has been widely employed to study the PUs' and SUs' best responses. The typical backward induction method to obtain the NE of such a Stackelberg game is first deriving the SUs' spectrum access decision, which is a function of the PUs' spectrum regulation policy, and then maximizing the PUs' utility to determine their optimal spectrum regulation policy, in turn the SUs' best spectrum access decision. In [6], the authors studied a scenario where the SUs act as the relay between the primary cellular base station and the PUs; meanwhile, in return, the base station allocates a portion of time in one time slot for the SUs. The primary base station first determines a fraction of one time slot for the SUs, and then each SU can decide on the payment to the PUs, which can be either money or a period of relay service. When calculating the Stackelberg NE of this game, a backward procedural was analyzed, first maximizing the SUs' utility and then the PUs'.

For the aforementioned second scenario (i.e., the sequential decision making between SUs) similar backward induction method can be utilized for the NE analysis. We studied such a sequential decision making problem in CCN using sequential dynamic Chinese restaurant game in [7], where the spectrum sensing and access analysis were integrated. This game model has two dimensions, one dimension is the ad-hoc SUs sequentially construct its belief on the availability of the primary cellular spectrum by using its spectrum sensing results, as well as the previous SUs' beliefs, which is a social learning perspective. The other dimension is the SUs sequentially access the spectrum according to one's own belief, as well as other SUs' spectrum access decisions. More importantly, the two dimensions are coupled together, since one SU's spectrum access decision is not only determined by its own belief, which is influenced by the previous SUs' belief, but also other SUs' spectrum access decisions due to the negative network externality; that is, the more SUs share a same part of spectrum resource, the less throughput each SU can obtain.

In the dynamic Chinese restaurant game model, the SUs are assumed to arrive at the primary cellular network by a Bernoulli process with certain probability. They sequentially sense and estimate the primary spectrum in a Bayesian manner, and sequentially access the available spectrum based on a multidimensional Markov decision process (M-MDP). A modified value iteration algorithm was also proposed to find the NE of the dynamic sequential spectrum access game. Figure 3 shows the performance evaluation of the proposed game strategy, where the centralized strategy is obtained by exhaustively searching all possible strategy profiles to maximize the social welfare. The myopic strategy is to maximize the immediate utility, that is, to choose the spectrum with the largest immediate reward, and the random strategy is to randomly

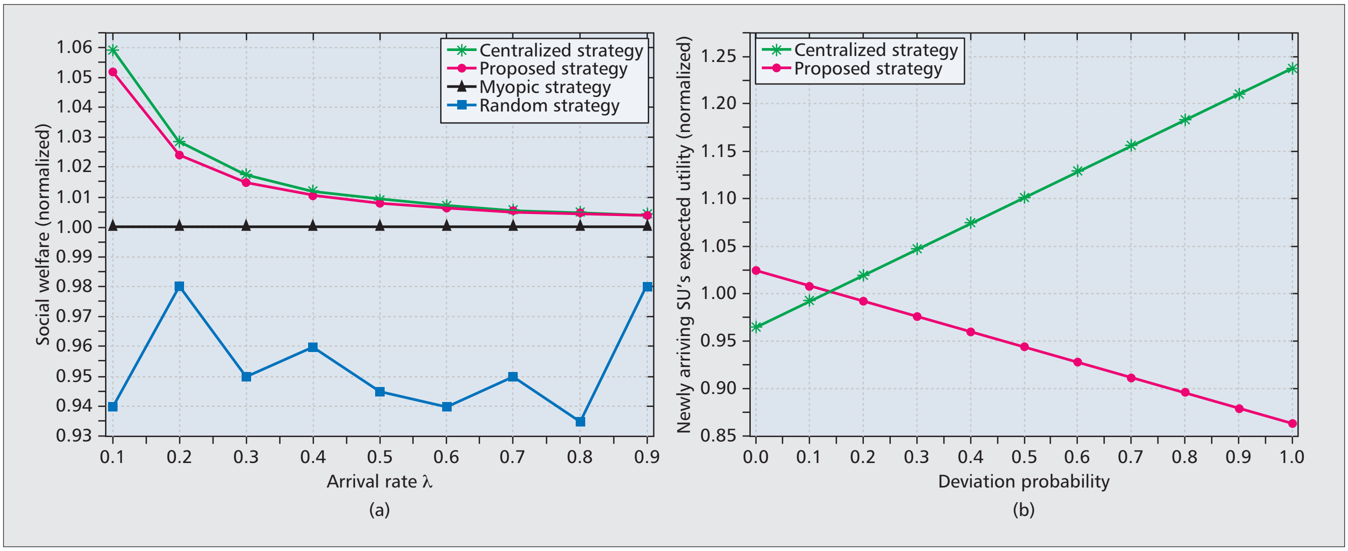


Figure 3. Left: social welfare comparison of different strategies; right: Nash equilibrium verification.

choose a part of primary spectrum to access. We can see that the proposed strategy performs better than both myopic and random ones, but slightly less than the centralized strategy. However, the centralized one cannot avoid the malicious SU's strategy deviation, as shown on the right of Fig. 3, and the complexity is extremely high, sometimes even computationally intractable.

Mechanism Design in CCNs

In a CCN, the primary cellular networks own the licensed spectrum, while the SUs attempt to dynamically utilize the spectrum. In most cases, such dynamic occupation is not cost-free for the SUs. Instead, they have to pay something in return to the PUs (e.g., money or serving as PUs' relays). Thus, the spectrum becomes a special kind of commodity in a CCN, where the cellular network operator can lease the vacant spectrum to the SUs, and the latter purchase the spectrum on demand. Such a spectrum trading process involves many practical problems. On one hand, for the operators of cellular networks, their issues are to price the spectrum to maximize their own profit, and to offer appropriate contracts to attract more SUs. On the other hand, for the SUs, they need to consider how much spectrum to purchase from the primary operator, and how to respond to the offered contract. Meanwhile, how to design a trading mechanism that can ensure that both the primary operator and the SUs attend the spectrum trading rationally, actively, honestly, and efficiently is also an important issue. Apparently, all those practical problems during the spectrum trading process in CCNs are similar to those during a common commodity transaction in real-world economics. Therefore, they can be addressed using the analytical tools in economics.

Mechanism design theory in economics can be applied to solve the aforementioned spectrum trading related problems. Different from the NE analysis under a given game rule, mechanism design is the "inverse" of traditional economic theory, which is typically devoted to a game structure design, as well as the analysis of the performance of a given mechanism. In the literature, some classical mechanism design methods have been adopted to analyze the spectrum trading problems, including contract design [8], auction design [9], and price design [10]. Contract theory studies how economic actors construct contractual arrangements, generally in the presence of asymmetric information; that is, they have little knowledge of the private information of the contract players. For example, in the labor market, employers generally do not completely know employees' private information before employment and need to offer employees a contract with incomplete information. This is quite similar to spectrum trading scenarios, where PUs offer a series of spectrum leasing contracts without SUs' preference information. Duan *et al.* investigated how the PUs should set up optimal/near-optimal contracts in [8], under complete, weakly complete, and strongly complete information of the SUs, respectively, where the information refers to the channel condition of each SU. The contract consists of a set of items representing combinations of the SUs' spectrum accessing time (i.e., reward) and relaying power (i.e., contribution). Through backward induction and maximizing the PUs' utility, that is, the average transmission rate during the entire time period, the author derived the optimal contracts from the perspective of primary cellular networks.

When it comes to the spectrum trading mechanism, auction is an ideal approach to organize the complex interaction between spectrum sellers and spectrum buyers. Generally, an auction mechanism is associated with a winning policy and a charging policy, the performance of which can be evaluated by four key properties:

- *Efficiency*: Resources are distributed to users that value them most.
- *Incentive compatibility*: A user cannot do better by unilaterally misreporting his/her value.

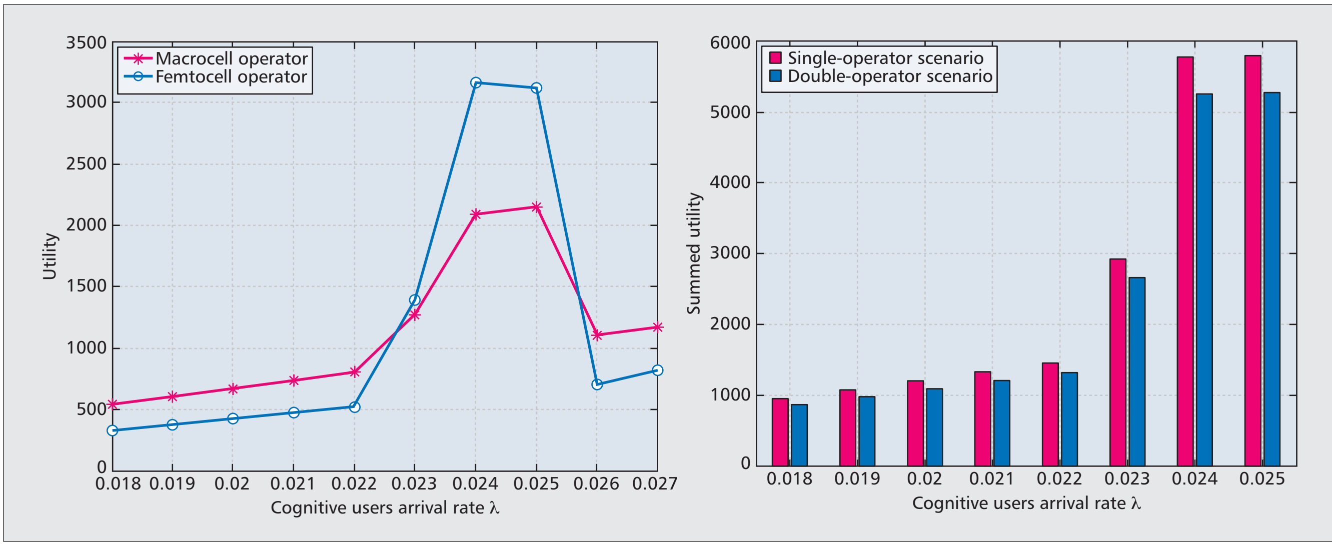


Figure 4. Left: utility comparison between macrocell and femtocell; right: utility comparison between single-operator and double-operator.

- *Individual rationality*: Users always expect non-negative value from the auction.
- *Budget balance*: Auctioneers do not lose money in the auction.

The authors in [9] considered a spectrum double auction model, where the spectrum seller (the cellular network operator) may exaggerate the spectrum value to the auctioneer, while the spectrum buyers (the SUs) may report untruthful channel valuation as bids. The auctioneer collects the spectrum sellers' prices and the spectrum buyers' bids to determine the winners, where the prices and bids refer to the spectrum leasing prices at which sellers are willing to sell and buyers are willing to pay, respectively. A spectrum seller or buyer may submit a different price or bid from its true price or bid, as long as it believes that this is more beneficial. The authors formulated such a double-auction mechanism design problem as an optimization problem: maximizing the auction efficiency defined as the portion of winning buyers under the constraint of incentive compatibility, individual rationality, and budget balance. Meanwhile, the local market phenomenon was also taken into account; that is, a spectrum buyer can trade with any seller whose license area is centered within the same hexagonal cell of the buyer. Therefore, a colored graph representing the spectrum assignment result was also associated with the proposed spectrum double-auction mechanism to coordinate the interference among cells.

Spectrum pricing is another important economic issue in CCNs, where the cellular operator charges the SUs based on their interference level (underlay model) or utilization of vacant spectrum (overlay model). A common method to calculate the optimal price is backward induction, that is, first analyzing the SUs' behaviors given the price, and then maximizing the PUs' revenue to find the optimal pricing policy. In [10], we investigated the spectrum pricing scenario in a heterogenous network, where the secondary femtocells lease spectrum from the primary macrocells. Given the spectrum leasing price, macrocell and femtocell operators set the network access prices independently and non-cooperatively. For the cognitive users, on one hand, accessing femtocells may obtain higher data rate but with higher payment due to additional spectrum leasing costs for the femtocell operator. On the other hand, accessing a macrocell network can lead to lower payment, but users may experience unsatisfactory data rate due to unfavorable locations. Therefore, rational cognitive users make their decisions (i.e., accessing a macrocell or femtocell network) by comparing the corresponding utilities. This pricing problem was formulated as a two-tier pricing model, and the NE prices were derived using the backward induction method under two models, static pricing and dynamic pricing, where "dynamic" means that the network access price is decreasing with the increasing number of users in the network; that is, negative network externality is considered.

The utility of macrocell and femtocell operators (the net profit under the proposed pricing mechanism) are shown on the left of Fig. 4 via simulation. When users' arrival rate is relatively small, the utility of a macrocell operator is larger than that of a femtocell operator, since accessing the uncrowded macrocell with a lower price is preferred by cognitive users. When the arrival rate is in the middle, a femtocell operator's utility becomes higher, which is because with the increasing number of users, accessing a femtocell network can achieve much higher throughput. When the arrival rate is relatively high, the macrocell operator's utility becomes higher again, since both networks become crowded, while the macrocell operator has additional spectrum leasing revenue. A comparison between our work and the scenario where both the macrocell and femtocell are operated by the same operator was also conducted. In such a single-operator scenario, the operator can globally optimize the network access prices of both macrocell and femtocell by maximizing their sum utilities. Apparently, as shown on the right of Fig. 4, the single-operator scenario can perform better than the double-operator scenario due to global optimization, but quite marginally. However, the single-operator scenario can only be applied to the monopoly-based market, which is rarely seen in the real-world market, while the double-operator scenario can be well applied in the competition-based market and provide an equilibrium point for the market, which is more common in the current practical scenario.

Conclusion

The network economics are the soul of CCNs; therefore, it is essential that they be carefully studied and planned for CCNs to thrive. In this article, the network economic issues in CCNs are discussed. We study both the simultaneous and sequential behaviors in CCNs with scenarios and examples to illustrate possible solutions. In addition, mechanism design issues to ensure desired outcomes are also introduced. This article offers an overview of the fundamental issues and key techniques regarding the network economic issues in CCNs, and also points out some possible research directions for future investigations.

Acknowledgment

This research was supported by NSFC China under projects 61371079, 61273214, 61271267 and 91338203, and by a Postdoctoral Science Foundation funded project.

References

[1] A. Attar *et al.*, "A Survey of Security Challenges in Cognitive Radio Networks: Solutions and Future Research Directions," *Proc. IEEE*, vol. 100, no. 12, Dec. 2012, pp. 3172–86.

[2] T. Jiang, H. Wang, and A. V. Vasilakos, "QoE-Driven Channel Allocation Schemes for Multimedia Transmission of Priority-Based Secondary Users over Cognitive Radio Networks," *IEEE JSAC*, vol. 30, no. 7, Aug. 2012, pp. 1215–24.

[3] P. Zhou, Y. Chang, and J. A. Copeland, "Reinforcement Learning for Repeated Power Control Game in Cognitive Radio Networks," *IEEE JSAC*, vol. 30, no. 1, May 2012, pp. 54–69.

[4] C. Jiang *et al.*, "Joint Spectrum Sensing and Access Evolutionary Game in Cognitive Radio Networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, 2013, pp. 2470–83.

[5] A. Attar, M. R. Nakhai, and A. H. Aghvami, "Cognitive Radio Game for Secondary Spectrum Access Problem," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, 2009, pp. 2121–31.

[6] O. Simeone *et al.*, "Spectrum Leasing to Cooperating Secondary Ad Hoc Networks," *IEEE JSAC*, vol. 26, no. 1, 2008, pp. 203–13.

[7] C. Jiang *et al.*, "Dynamic Chinese Restaurant Game: Theory and Application to Cognitive Radio Networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, 2014, pp. 1960–73.

[8] L. Duan, L. Gao, and J. Huang, "Cooperative Spectrum Sharing: A Contract-Based Approach," *IEEE Trans. Mobile Comp.*, vol. 13, no. 1, 2014, pp. 174–87.

[9] W. Wang, B. Liang, and B. Li, "Designing Truthful Spectrum Double Auctions with Local Markets," *IEEE Trans. Mobile Comp.*, vol. 13, no. 1, 2014, pp. 75–88.

[10] C. Jiang *et al.*, "Optimal Pricing Strategy for Operators in Cognitive Femtocell Networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, 2014, pp. 5288–5301.

Biographies

CHUNXIAO JIANG [S'09, M'13] (jchx@tsinghua.edu.cn) received his B.S. (Hons.) degree in information engineering from Beihang University, Beijing, China, in 2008, and his Ph.D. (Hons.) degree from Tsinghua University (THU), Beijing, in 2013. From 2011 to 2013, he visited the Signals and Information Group at the Department of Electrical and Computer Engineering, University of Maryland at College Park, with Prof. K. J. Ray Liu. He currently holds a postdoctoral position with the Department of Electrical Engineering, THU, with Prof. Y. Ren. His research interests include the applications of game theory and queuing theory in wireless communication and networking, and social networks. He was a recipient of the Best Paper Award from IEEE GLOBECOM in 2013, the Beijing Distinguished Graduated Student Award, the Chinese National Fellowship, and the Tsinghua Outstanding Distinguished Doctoral Dissertation in 2013.

YAN CHEN [SM'14] (yan@umd.edu) received his Bachelor's degree from the University of Science and Technology of China in 2004, his M.Phil degree from Hong Kong University of Science and Technology (HKUST) in 2007, and his Ph.D. degree from the University of Maryland at College Park in 2011. His current research interests are in data science, network science, game theory, social learning, and networking, as well as signal processing and wireless communications. He is the recipient of multiple honors and awards including the best paper award from IEEE GLOBECOM in 2013, the Future Faculty Fellowship and Distinguished Dissertation Fellowship Honorable Mention from the Department of Electrical and Computer Engineering in 2010 and 2011, respectively, a Finalist for the Dean's Doctoral Research Award from A. James Clark School of Engineering at the University of Maryland in 2011, and the Chinese Government Award for outstanding students abroad in 2011.

K. J. RAY LIU [F'03] (kjrliu@umd.edu) was named a Distinguished Scholar-Teacher of the University of Maryland at College Park in 2007, where he is the Christine Kim Eminent Professor of Information Technology. He leads the Maryland Signals and Information Group conducting research encompassing broad areas of signal processing and communications with recent focus on future wireless technologies, network science, and information forensics and security. He was a recipient of the IEEE Signal Processing Society 2014 Society Award, IEEE Signal Processing Society 2009 Technical Achievement Award, and best paper awards from various IEEE societies and EURASIP. Recognized by Thomson Reuters as an ISI Highly Cited Researcher, he is a Fellow of AAAS. He is a Director-Elect of the IEEE Board of Directors. He was President of the IEEE Signal Processing Society (2012–2013) where he has served as Vice-President-Publications and on the Board of Governors. He was Editor-in-Chief of *IEEE Signal Processing Magazine*. He has received teaching and research recognition from the University of Maryland, including a university-level Invention of the Year Award and a college-level Poole and Kent Senior Faculty Teaching Award, Outstanding Faculty Research Award, and Outstanding Faculty Service Award, all from A. James Clark School of Engineering.

YONG REN (reny@tsinghua.edu.cn) received his B.S, M.S., and Ph.D. degrees in electronic engineering from Harbin Institute of Technology, China, in 1984, 1987, and 1994, respectively. He worked as a post-doctoral researcher at the Department of Electronics Engineering, THU, from 1995 to 1997. Now he is a professor in the Department of Electronics Engineering and director of the Complexity Engineered Systems Lab (CESL) of THU. He holds 12 patents, and has authored or co-authored more than 100 technical papers on the behavior of computer networks, P2P networks, and cognitive networks. His current research interests include complex systems theory and its applications to the optimization and information sharing of the Internet, space-based information networks, Internet of Things and ubiquitous networks, cognitive networks, and cyber-physical systems.

The authors formulated such a double-auction mechanism design problem as an optimization problem: maximizing the auction efficiency defined as proportion of winning buyers under the constraint of incentive compatibility, individually rationality and budget balance.

Business Modeling for TV White Space Networks

Yuan Luo, Lin Gao, and Jianwei Huang

The authors are with the Chinese University of Hong Kong.

ABSTRACT

Database-assisted TV white space network is a promising paradigm of dynamic spectrum sharing, and can effectively improve spectrum utilization and alleviate spectrum scarcity via centralized control of TV white space databases residing in the cloud. In this article, we discuss business modeling for a database-assisted TV white space network, which is very important for wide commercialization of this promising technology. Motivated by several recent business practices, we propose and study two types of different business models: *spectrum market* and *information market*. In the spectrum market model, spectrum licensees, through spectrum brokers acted by databases, lease the underutilized (licensed) TV channels to unlicensed wireless devices for secondary utilization. In the information market model, databases sell the advanced information regarding (unlicensed) TV channels to unlicensed wireless devices in order to enhance the secondary spectrum utilization performance. We outline the corresponding trading mechanism details for both market models, and evaluate the feasibility and performance of both models through theoretical and numerical studies. Numerical results indicate that both the database operator's profit and the total network profit can be significantly improved under properly designed trading mechanisms.

INTRODUCTION

BACKGROUND

With the explosive growth of mobile smartphones and bandwidth-hungry wireless applications, radio spectrum is becoming increasingly congested. TVwhite space (TVWS) has recently been recognized as a promising new spectrum opportunity for wireless broadband services, due to its low utilization (at most times and in many areas) and brilliant propagation performance [1]. Specifically, TVWS (also called TV channel) refers to the unused or underutilized broadcast television spectrum (in the UHF/VHF frequency band) at a particular time and location. By allowing unlicensed wireless devices (called white space devices, WSDs) reuse the TVWS in a license-exempt and opportunistic manner, we can effectively improve the spectrum efficiency and alleviate today's global spectrum scarcity. As a consequence, in the United States, the Federal Communications Commission (FCC) has unanimously approved the license-exempt use of TVWS to support new wireless applications [1]. Moreover, Ofcom and the Singapore TVWS pilot group have launched several pilots of TVWS technology in the United Kingdom and Singapore, respectively.[1]

To effectively reuse the TVWS spectrum without harming the interests of licensed devices, spectrum regulators have advocated a *database-assisted* TVWS network architecture [1]. In this architecture, unlicensed WSDs obtain the available TV channel information via querying a certified *geolocation* database residing in the cloud instead of sensing the local spectrum environment as in traditional dynamic spectrum sharing systems. To achieve this, the geolocation database needs to house and periodically update information related to network infrastructures of TV licensees as well as their channel occupations. Figure 1 illustrates such a database-assisted TVWS network architecture. To access any TVWS, WSDs first report their locations to a geolocation database (step 1), and then the database computes and returns the available TV channels that WSDs can use in a certain time period (step 2). In this sense, the database-assisted TVWS network is a typical example of the cloud-enabled virtualized network.[2] As illustrated in Fig. 1, each WSD is an infrastructure-based device (e.g., a base station) operated by a secondary operator, and provides *cellular-based* wireless services to its subscribed end users by using the obtained TV channels.[3]

The database-assisted TVWS network has received wide and enthusiastic support, not only from spectrum regulators, but also from standards bodies and industrial organizations.[4] The geolocation database is no doubt the central network entity in such a network. In the United States, the FCC has temporarily certified several major IT companies including Google, Microsoft, and SpectrumBridge as geolocation database operators. Obviously, the long-term and large-scale commercial deployment of such a database-assisted network requires a proper business model that gives the database operators the opportunity to create and capture sufficient value in order to cover their capital expense (CAPEX) and operating expense (OPEX).

[1] For more details, please refer to Ofcom TV White Space Network Pilot (http://stakeholders.ofcom.org.uk/spectrum/tv-white-spaces/white-spaces-pilot/) and Singapore White Space Pilot Group (http://www.whitespaces.sg/projects.php

[2] Specifically, local computation such as location measurement and identification are performed on a WSD, while complex data processing such as TV channel availability computing, channel scheduling, and business modeling (if needed) are implemented in the white space database residing in the cloud. WSDs and the database communicate with each other through the Internet. Such an architecture can reduce the energy consumption of mobile WSDs and improve system performance.

[3] In addition to this cellular-based network scenario, WSDs can also form networks in ad hoc mode. Nevertheless, in this article, we focus on the cellular-based network scenario.

 0163-6804/15/$25.00 © 2015 IEEE

Solutions

Before talking about the business modeling solutions, we first look at the current business practices of SpectrumBridge Inc.,[5] the world's first FCC-certified geolocation database operator in the United States. Specifically, SpectrumBridge offers two different commercial models: *SpecEx* and *White Space Plus*. The formal model, SpecEx, enables TV licensees to lease their licensed but underutilized TV channels to unlicensed WSDs. In this process, the database acts as a spectrum broker to facilitate the trading process (e.g., connecting and matching buyers and sellers). The latter model, White Space Plus, on the other hand, enables the database to sell *information* regarding TV channels to unlicensed WSDs. This is motivated by the fact that the database has more information regarding the quality of TV channels, and such information can potentially be used by unlicensed WSDs to improve their performance. The above discussion actually leads to two different types of business models for TVWS networks, the *spectrum market* and the *information market*, which will be the main focus of this article.

Specifically, the spectrum market and information market target different types of TV channels in TVWS networks: *licensed* and *unlicensed* TV channels. The licensed TV channels are those registered to some TV licensees but underutilized by the licensees. Hence, the licensees can temporarily lease the underutilized (licensed) channels to WSDs for exclusive usage during a short time period. The unlicensed TV channels are those not registered to any TV licensee at a particular location, and hence are a public resource at that location. Unlicensed TV channels are usually assigned by spectrum regulators for public and shared usage among unlicensed WSDs, and not allowed for trading in a spectrum market. Moreover, due to the shared usage by unlicensed WSDs, the quality of unlicensed TV channels is usually not guaranteed. Hence, the database can potentially sell its advanced information regarding the quality of unlicensed TV channels to WSDs.

In this article, we analyze both the spectrum and information markets under *non-competitive* and *competitive* market scenarios. Figure 2 illustrates the taxonomy of TVWS business models in this article. The rest of this article is organized as follows. We begin with an overview of the database-assisted TVWS network as well as the technical and economical issues in this new network. We then provide detailed formulation and analysis of the spectrum and information markets. We further discuss the future challenges and open issues in this area, and conclude the article.

Technology Overview

In this section, we briefly introduce the TVWS network technology as well as the technical and business model design issues in this new type of network. As shown in Fig. 1, one of the most important features of a database-assisted TVWS network is that unlicensed WSDs obtain the TV channel availability information through querying a geolocation database residing in the cloud, rather than directly sensing the current activity levels in the TV channels. Hence, the geolocation database is the central network entity in such a network.

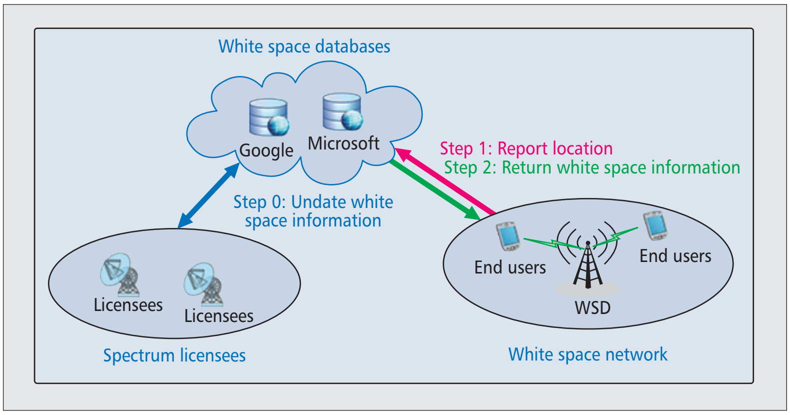


Figure 1. Database-assisted TVWS network architecture.

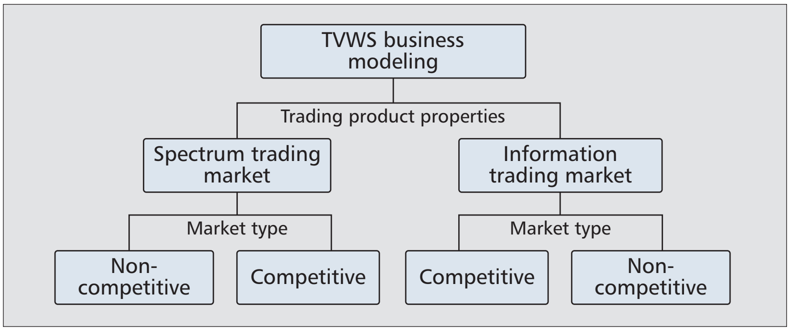


Figure 2. Taxonomy of TVWS network business modeling.

Technical Issues

There are several technical challenges in developing a database-assisted TVWS network, including:

- How can a geolocation database residing in the cloud be designed and managed?
- How does a geolocation database accurately compute the available TV channels in a particular location?
- How can a database-assisted TVWS network be efficiently deployed and optimized?

Many research works have been devoted to address these technical issues. Gurney *et al.* [2] presented an effective geolocation database design and discussed the TVWS determination issue. Murty *et al.* [3] presented and evaluated a TVWS exploration method using a more accurate propagation model with terrain data. Goncalves *et al.* [4] compared the geolocation database assisted approach with the sensing approach. In [5], Feng *et al.* presented the design and implementation of a multicell infrastructure-based TVWS network. Chen *et al.* [6] studied the TVWS network optimization problem. Zhang *et al.* [7] proposed a vehicle-based measurement framework to enhance the performance of a database. In fact, many of the above technical issues have been well solved. As a result, several geolocation databases (e.g., those deployed by Google, Microsoft, and SpectrumBridge) have been successfully developed, based

[4] These include the IEEE 802.22 WRAN standard (http://www.ieee802.org/22/) and the real world white space database systems deployed by Google (http://www.google.org/spectrum/whitespace/), Microsoft (http://whitespaces.msresearch.us/), and Spectrum Bridge (http://www.spectrumbridge.com/

[5] SpectrumBridge Inc., http://www.spectrumbridge.com/

on which some TVWS trial systems have been launched in the United Kingdom, Singapore, and the United States.

Economic Issues

In addition to the technical challenges mentioned above, there are also critical economic challenges in commercially implementing a database-assisted TVWS network. In particular, there is a lack of systematical study on the business modeling of such a network, which has become the main bottleneck of the wide commercialization of this new network. Designing a proper business model for the database-assisted TVWS network requires to consider and answer many challenging questions, including:

- Who will be involved in a TVWS business model, and how will they interact with each other? How can the economic role of each involved network be properly defined?
- What kind of services will be supported in a TVWS network, and how can the services exploit the unique characteristics of licensed and unlicensed TV channels?
- How are the proposed business models analyzed and optimized? How can efficient mechanisms be designed to realize the proposed business models, considering the performance as well as the implementation complexity?

Tackling the above questions is challenging due to the following reasons. First, TV channels are heterogeneous in terms of their properties. That is, some channels are licensed to certain TV licensees, while others are unlicensed. The model related to licensed TV channels must involve their licensees, while the model related to unlicensed TV channels does not have this requirement. Second, database operators are heterogeneous in terms of their interests and advantages. Note that the geolocation databases connect the TV licensees to unlicensed WSDs and play a central role in a TVWS business model. Different database operators may lead to very different business models. Moreover, to balance secondary spectrum utilization and primary licensee protection, spectrum regulators have specified strict technical restrictions on the secondary usage of TV channels. This may bring additional challenges in the design of a feasible business model. Several prior studies [8–11] have considered some economics issues in TVWS networks. However, they only focused on the design of the trading mechanism for the licensed TV channels between the spectrum licensees and unlicensed WSDs, without particularly involving the geolocation databases in the market mechanism design.

Business Models

In this article, we consider two types of different business models motivated by the SpectrumBridge business cases:

- *Spectrum market* for licensed TV channels
- *Information market* for unlicensed TV channels

The Spectrum Market Model

In the spectrum market model, TV licensees temporarily lease their licensed TV channels to unlicensed WSDs for additional revenue. During this process, the database acts as a spectrum broker, purchasing licensed TV channels from TV licensees in advance and reselling the purchased licensed TV channels to WSDs. WSDs further serve their subscribed end users by using the purchased TV channels. Depending on whether WSDs serve overlapping end users, we discuss both *non-competitive* and *competitive* scenarios for the spectrum market model.

Spectrum Market with Non-Competitive WSDs — In the non-competitive scenario, each WSD faces a distinct pool of end users, and there is no competition among WSDs. Hence, we can focus on the interaction between the database and one WSD. The database needs to purchase (reserve) licensed TV channels from the licensees in advance,[6] without knowing the actual demands of WSDs. Therefore, a key problem is *how much should the database reserve for each WSD?* The problem is challenging due to the following factors:

- *Stochastic demand*: Due to the stochastic nature of end users' activities and requirements, the WSD's spectrum demand (from end users) is random, and cannot be accurately predicted in advance. Hence, there is a risk of reservation mismatch (over-reservation or under-reservation).
- *Asymmetric information*: The WSD usually has more information (i.e., less uncertainty) about the end-user demand than the database, due to its proximity to end users. Note that sharing such information with the database can potentially reduce the risk of reservation mismatch. Nevertheless, the WSD may not be willing to share its private information unless it receives sufficient incentives.

In [12], we proposed a *contract-theoretic* spectrum reservation framework to tackle the above challenges. The key idea is as follows. Before reserving , the database announces a reservation contract including a menu of contract items, where each contract item specifies a particular choice of the amount reserved and the corresponding reservation fee; the WSD selects the contract item (i.e., the reservation amount and the corresponding reservation fee) that maximizes its expected profit based on its private demand information. Then the detailed spectrum reservation, trading, and access processes are shown in Fig. 3. Specifically:

1. The database reserves the corresponding amount of TV channels for each reservation period based on the WSD's selection, and charges the WSD the corresponding reservation fee (step 0).
2. The WSD reports the realized actual demand to the database in each access period (step 1).
3. The database sells the corresponding amount of TV channels to the WSD in each access period (step 2).
4. The WSD serves its subscribed end users by using the purchased channels in each access period (step 3).

We studied the optimal contract design systematically in [12]. Specifically, we first characterized the incentive compatibility (IC) and individual rationality (IR) conditions for a feasi-

[6] In order to guarantee the exclusive usage of licensed TV channels by WSDs, the database needs to negotiate with the TV licensees in advance in terms of when and which licensed TV channels can be used by WSDs.

ble contract, which ensure that each WSD discloses its private demand information credibly. Then we further derived the optimal contract that maximizes the database's profit.

Figure 4 presents the performance of the proposed spectrum reservation contract, which significantly outperforms the traditional database-determined reservation scheme (without involving the private information and decision of a WSD) in terms of the total network profit (the first bar group) as well as the database's individual profit (the third bar group). Moreover, the total network profit under the proposed contract is very close to (the gap is less than 3 percent) the centralized benchmark solution, where the database and the WSD make a decision together as an integrated party.

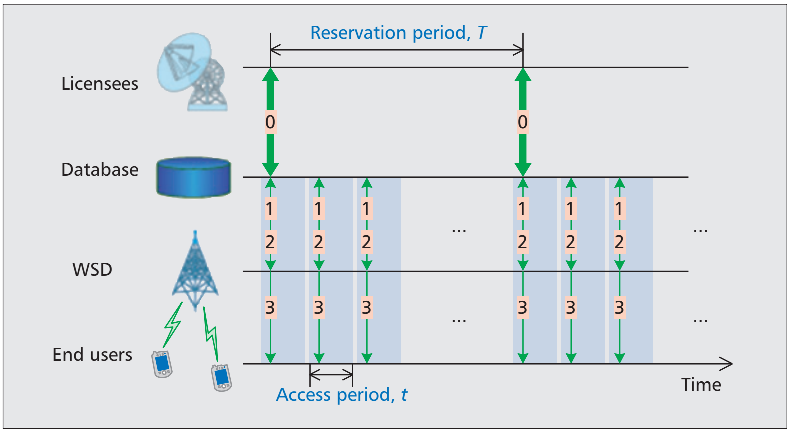


Figure 3. The spectrum reservation, trading, and access process in a spectrum market.

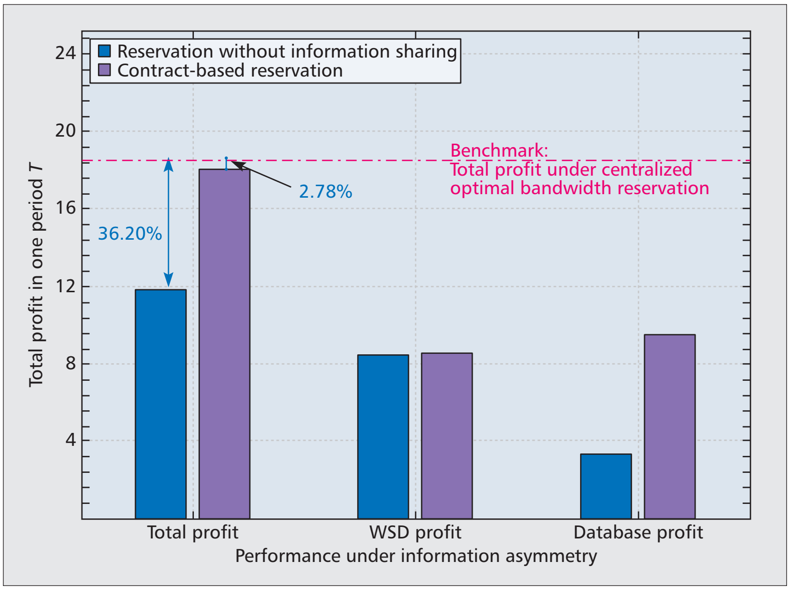


Figure 4. Performance of a spectrum reservation contract.

Spectrum Market with Competitive WSDs — In the competitive scenario, multiple WSDs compete for a common pool of end users. Some new questions arise in this scenario:

- How do WSDs interact with each other as well as with end users?
- What is the equilibrium spectrum reservation level and service price for each WSD, taking the competition of other WSDs into consideration?
- What is the optimal wholesale price for the database to maximize its profit or maximize the network profit?

Addressing these questions is challenging due to the stochasticity of end-user demand. Moreover, when the total reserved licensed TV channels are not enough for satisfying all end-user demand, a WSD may further request unlicensed TV channels to serve the excessive end-user demand at a degraded quality of service (QoS) level. It is important to note that although the unlicensed TV channels can be used by WSDs without any payment to the licensees, WSDs may still incur some cost when accessing the unlicensed TV channels. For example, a WSD needs to consume some time and energy to request unlicensed TV channels from the database. Meanwhile, the database also needs to exert some effort to help WSDs use these unlicensed TV channels and accordingly charges WSDs a certain fee.[7] This further complicates the market analysis. In [13], we proposed a three-stage hierarchical model to study such a competitive spectrum market with both licensed and unlicensed TV channels. As illustrated in Fig. 5, the key processes in this model are as follows:

- In stage I, the database determines the *wholesale* prices of both licensed and unlicensed TV channels.
- In stage II, each WSD determines the reservation level of licensed TV channels and the service price.
- In stage III, each end user demands services from the best WSD based on her preferences (e.g., channel conditions to WSDs) as well as WSDs' service prices. If the total end-user demand is larger than the reserved licensed TV channels, a WSD will further purchase unlicensed TV channels for the excess demand.

We study the above three-stage hierarchical model systematically in [13]. Specifically, we first formulate the WSDs' competition in stages II and III as a non-cooperative game, and characterize the game equilibrium analytically by using super-modular game theory. Then we study two different wholesale pricing solutions for the database in stage I: a welfare maximization solution for a *social planning* database and a profit maximization solution for a *profit seeking* database. For a social planning database (e.g., those managed by non-profit organizations such as government regulators), the database operator's objective is to maximize the total network profit or welfare (i.e., the aggregate profit of WSDs and the database operator). For a profit seeking database (e.g., those managed by third-party businesses such as Google, Microsoft, and SpectrumBridge), the database operator's objective is to maximize its individual profit.

Figure 6 presents the performance of the proposed wholesale pricing solutions, where WM denotes the welfare maximization solution and PM denotes the profit maximization solution. The blue (or red) bar in each bar group denotes the network profit under the WM (or PM)

[7] We also denote the service fee charged by the database as the wholesale price.

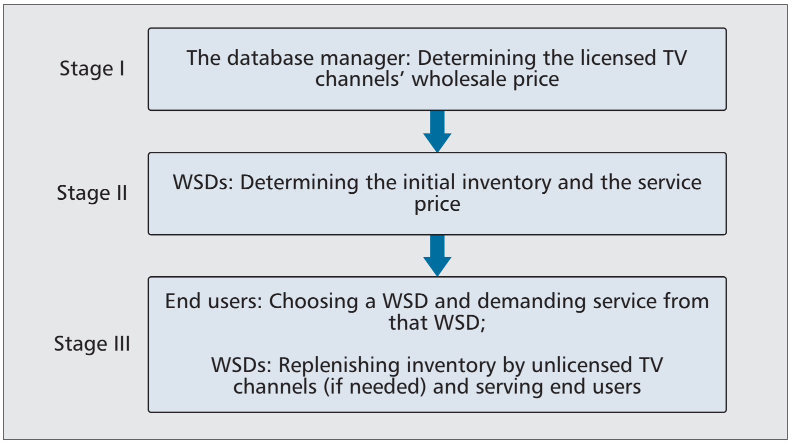


Figure 5. Three-stage hierarchical model.

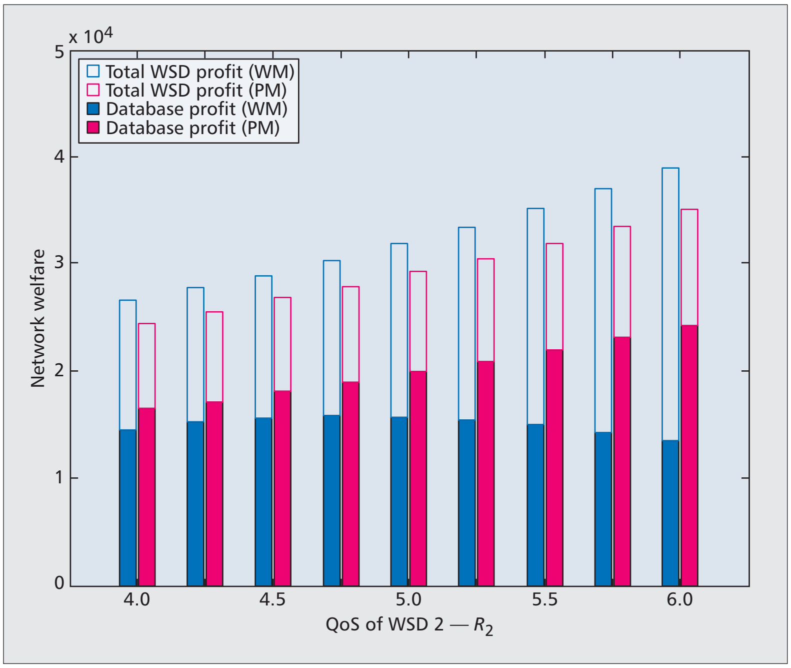


Figure 6. Performance of welfare maximization and profit maximization wholesale pricing solutions.

scheme, where the database operator's profit is denoted by the solid bar, and the total WSDs' profit is denoted by the hollow bar. In this numerical simulation, we consider a market with two competitive WSDs, where the QoS of WSD 1 is fixed at $R_1 = 4$, while the QoS of WSD2 changes from $R_2 = 4.0$ to 6.0. We can see that under both WM and PM schemes, the network profit increases with the QoS provided by WSD 2 (R_2). This is because a higher QoS can potentially attract more end users. Moreover, the network profit under the PM scheme is less than that under the WM scheme. This implies that maximizing the database's individual profit may lead to certain network profit loss from the system perspective.

The Information Market Model

In the information market model, the database sells advanced information regarding the quality of unlicensed TV channels to unlicensed WSDs for profit. This model is motivated by the fact that the geolocation database knows more information regarding the quality of unlicensed TV channels than unlicensed WSDs, and such information can potentially be used by WSDs to improve their performance.

In the following, we formulate and analyze the information market systematically. To do this, we need to answer the following questions explicitly:

- *Information definition*: What kind of information can be traded in the information market? How are the quality and/or quantity of information defined?
- *Information evaluation*: How will WSDs evaluate the information? Will the same information have different values for different WSDs?
- *Market evolution*: How would the information market dynamically evolve along the WSDs' purchasing behaviors? With what type of network image (i.e., positive or negative) will the market appear?
- *Market optimization*: How does the database make the best pricing decision for the information that it sells?

All of the above issues are critical for the formulation and analysis of an information market. Specifically, addressing the first two issues helps us define an information market explicitly; addressing the third issue helps us understand the market evolution; and addressing the last issue helps us design the proper mechanism such that the database can draw desirable profit from the information market.

The Non-Competitive Information Market — In [14], we proposed and studied a non-competitive information market where a single geolocation database sells information to a set of unlicensed WSDs. In the proposed model, the database provides two different kinds of information services:

- *Basic information*: According to the regulator's ruling (e.g., FCC [1]), a geolocation database is mandatory to provide the following basic information for WSDs:
  - –The list of available unlicensed TV channels
  - –The transmission constraints (e.g., maximum transmission power) on each available channel

  Such basic information service provided by the database is free of charge. With this basic information, a WSD can either randomly choose an unlicensed TV channel from the available channel list, or sense all available TV channels in the list (with some sensing cost) in order to figure out the best one.
- *Advanced information*: Beyond basic service, the database can also provide certain advanced information (called advanced service) to make a profit, as long as doing so does not conflict with the mandatory basic service. In [14], we explicitly define the advanced information as *the interference level on each available channel for each WSD*.[8] With this advanced information, a WSD is able to operate on the best

[8] The framework in [14] is rather general, and can apply to other types of advanced information.

available channel (i.e., that with the lowest interference level). Accordingly, the database can charge a subscription fee to every WSD subscribing to its advanced service.

We propose a general framework for evaluating the value of interference information to WSDs. Notice that the interference on a channel (to a particular WSD) may come from nearby TV stations operating on that channel or from nearby WSDs using that channel. The database can (relatively) precisely predict the interference from TV stations, as it maintains a repository of TV licensees. However, it may not be able to predict the precise interference from WSDs, as some WSDs may not inform the database of their choices of channels. Therefore, the overall interference information provided by the database may not be accurate. This will affect the value of information for WSD users, which in turn will affect how the database may price the information.

After characterizing the value of information to WSDs, we can derive the stable market share (i.e., the percentage of WSDs who purchase advanced information from the database operator), called the *market equilibrium*. In contrast to traditional spectrum markets, which are usually congestion-oriented (i.e., the more users purchasing and using spectrum, the lower the value of spectrum for each user), we show that the information market has the appealing property of *positive externality*. That is, the more users purchasing information from the database, the higher the value of the information for each buyer. This is because when more users purchase the information and reveal their channel selections to the database implicitly, the database can predict the interference more accurately.

Due to the positive network externality, the market equilibrium increases with the initial market share. Interestingly, there are several critical points (called *tipping points*) of the initial market share around which a slight change will result in a significant change on the emerging market equilibrium. Finally, based on the market equilibrium analysis, we derive the optimal information pricing plan that maximizes the database profit or revenue.

Figure 7 presents the database's maximum expected profit from the information market, which shows that the database's profit increases with the level of licensee interference (as a high level of licensed interference implies a potentially high value of the database's information for WSDs) and with the WSD sensing cost α (as a WSD with a higher sensing cost has more interest in purchasing the database's information).

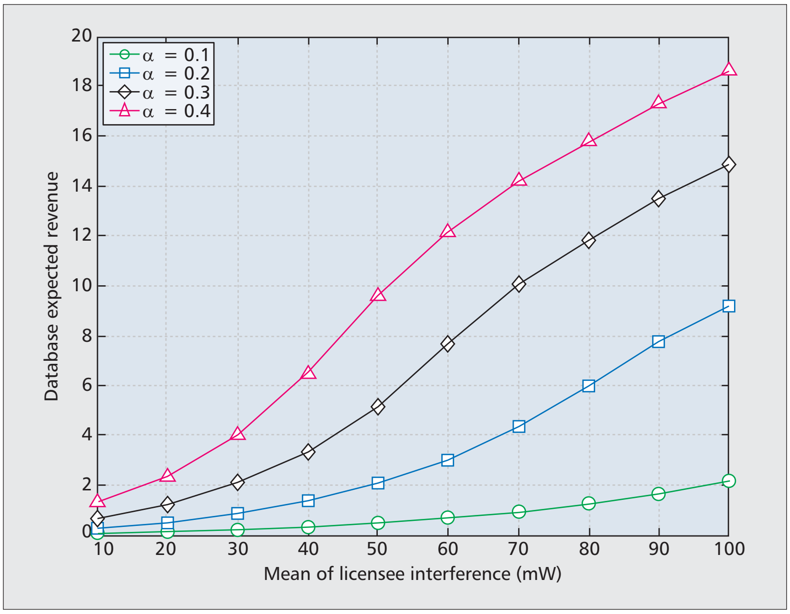


Figure 7. Database profit in a non-competitive information market.

The Competitive Information Market — In [15], we proposed and studied an oligopoly competitive information market, where two databases compete for selling information to WSDs. Comparing with the non-competitive market, the analysis in a competitive market becomes more challenging due to the following reasons:

- The information possessed by different databases may be different, because of either the different knowledge of databases or the different market shares of databases.[9] Hence, WSDs need to jointly consider the quality and price of information before choosing with which database to interact.
- The positive externality of the information market provides a large incentive for each data operator to increase its market share, and hence makes the competition between database operators more intense than a market without positive network externality.

We study the oligopoly price competition in such an information market systematically and characterize the equilibrium analytically in [15]. Our analysis indicates that given the prices of databases, there may be multiple market equilibria, and which one will actually emerge depends on the initial market shares of both databases. We also quantify the impact of the databases' initial market shares and information prices on the market equilibrium. Our results show that

- When the prices of two databases are very different, there is a unique stable market equilibrium independent of the databases' initial market shares, where the lower price database achieves a large market share and the higher price database does not have any subscribers;
- When the prices of two databases are similar, there are two stable market equilibria depending on the databases' initial market shares, where the database with the higher initial market share achieves a large market share at the equilibrium, and the database with the lower initial market share achieves a small market share. Based on this market equilibrium analysis, we further derive the equilibrium of the databases' price competition game.

Our numerical simulations show that the database with a larger initial market share has a significant advantage in the competition, and can achieve most of the market share under the equilibrium.

[9] As mentioned previously, a database's information accuracy increases with its market share.

Our analysis not only shows how different network entities interact with each other in both markets, but also shows how the markets dynamically evolve under different circumstances and what the market equilibria are. These results can serve as an important first step toward the study of the general and large-scale TVWS market.

FUTURE CHALLENGES AND OPEN ISSUES

In spite of recent efforts from industry and academia, there are still many future challenges and open issues in the business modeling of database-assisted TVWS networks.

SPECTRUM MARKET AND INFORMATION MARKET: CONFLICTING OR COMPLEMENTING?

As mentioned previously, the spectrum market mainly targets licensed (and underutilized) TV channels, and the information market mainly targets unlicensed TV channels. In practice, licensed and unlicensed TV channels coexist in most cases. Some WSDs may prefer to lease licensed TV channels from licensees for exclusive use, while other WSDs may prefer to share free unlicensed TV channels with others. Hence, a joint formulation and optimization of both the spectrum and information markets is highly desirable.

AN AD-SPONSORED BUSINESS MODEL

An ad-sponsored business model may be another possible choice for a database operator. In an ad-sponsored model, it is important to understand how the database operators, advertisers, and WSDs interact with each other, and the unique feature of an ad-sponsored TVWS network. Generally, as in many ad-sponsored models, more WSDs making inquiries of the database may provide more business opportunities and incentives for advertisers, but too much advertising may annoy WSDs and drive them away to other databases. It is hence a delicate balance to achieve so as to coordinate the behavior of all involved parties to achieve a win-win situation.

CONCLUSIONS

Business modeling is critical for the practical commercialization of database-assisted TVWS networks. In this article, we outline and analyze two promising business models: spectrum market and information market. These models allow the geolocation databases to exploit the unique characteristics of licensed and unlicensed TV channels. Our analysis not only shows how different network entities interact with each other in both markets, but also shows how the markets dynamically evolve under different circumstances and what the market equilibria are. These results can serve as an important first step toward the study of the general and large-scale TVWS market.

ACKNOWLEDGMENT

This work is supported by the General Research Funds (Project Number CUHK 412713 and 14202814) established under the University Grant Committee of the Hong Kong Special Administrative Region, China.

REFERENCES

[1] FCC 10-174, *Second Memorandum Opinion and Order*, 2010.

[2] D. Gurney *et al.*, "geolocation Database Techniques for Incumbent Protection in the TVWS," *IEEE DySpan*, Chicago, IL, 2008.

[3] R. Murty *et al.*, "SenseLess: A Database-Driven White Spaces Network," *IEEE Trans. Mobile Comp.*, vol. 11, no. 2, Feb. 2011, pp. 189–203.

[4] V. Goncalves and S. Pollin, "The Value of Sensing for TVWSs," *IEEE DySpan*, Washington, DC, 2011.

[5] X. Feng, J. Zhang, and Q. Zhang, "Database-Assisted Multi-AP Network on TVWSs: Architecture, Spectrum Allocation and AP Discovery," *IEEE DySPAN*, Washington, DC, 2011.

[6] X. Chen and J. Huang, "Game Theoretic Analysis of Distributed Spectrum Sharing with Database," IEEE ICDCS, Macau, China, 2012.

[7] T. Zhang, N. Leung, and S. Banerjee, "A Vehicle-Based Measurement Framework for Enhancing Whitesapce Spectrum Databases," ACM Mobicom, Hawaii, 2014.

[8] H. Bogucka *et al.*, "Secondary Spectrum Trading in TVWSs," *IEEE Commun. Mag.*, vol. 50, no. 11, Nov. 2012, pp. 121–29.

[9] S. Liu, *et al.*, "Location Privacy Preserving Dynamic Spectrum Auction in Cognitive Radio Network," *IEEE ICECS*, Abu Dhabi, UAE, Dec. 2013.

[10] M. Parzy, "The Profitability Analysis of the Multi-Band Spectrum Broker," *IEEE PIMRC*, London, U.K., Sept. 2013.

[11] X. Feng, J. Zhang, and J. Zhang, "Hybrid Pricing for TVWS Database," IEEE INFOCOM, Turin, Italy, Apr. 2013.

[12] Y. Luo, L. Gao, and J. Huang, "Spectrum Broker by geolocation Database," *IEEE GLOBECOM*, Anaheim, CA, Dec. 2012.

[13] Y. Luo, L. Gao, and J. Huang, "Price and Inventory Competition in Oligopoly TVWS Markets," *IEEE JSAC*, 2014.

[14] Y. Luo, L. Gao, and J. Huang, "Trade Information, Not Spectrum: A Novel TVWS Information Market Model," *IEEE WiOpt*, Hammamet, Tunisia, May 2014.

[15] Y. Luo, L. Gao, and J. Huang, "Information Market for TVWS," IEEE INFOCOM Wksp. Smart Data Pricing, Toronto, Canada, Apr. 2014.

BIOGRAPHIES

YUAN LUO (ly011@ie.cuhk.edu.hk) is a Ph.D. student in the Department of Information Engineering at the Chinese University of Hong Kong. She received her B.S. degree from Tianjin University, China, and her M.S. degree from Beijing University of Posts and Telecommunications, China, in 2008 and 2011, respectively. Her research interests lie in the field of wireless communications and network economics, with current emphasis on TVWS networks and crowdsourcing. She is the recipient of the Best Paper Award at the IEEE International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks 2014.

LIN GAO (lgao@ie.cuhk.edu.hk) is a postdoctoral researcher in the Department of Information Engineering at the Chinese University of Hong Kong. He received M.S. and Ph.D. degrees in electronic engineering from Shanghai Jiao Tong Univ., China, in 2006 and 2010, respectively. His research interests are in game theory and network economics, in particular, game-theoretic modeling and analysis for cooperative communications, cognitive radio networks, TVWS networks, cellular-WiFi interworking, mobile cloud computing, and mobile crowd sourcing. He is the co-recipient of two Best Paper Awards from the IEEE International Symp. on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks 2013 and 2014.

JIANWEI HUANG [S'01, M'06, SM'11] (jwhuang@ie.cuhk.edu.hk) is an associate professor and director of the Network Communications and Economics Lab (ncel.ie.cuhk.edu.hk) in the Department of Information Engineering at the Chinese University of Hong Kong. He received his Ph.D. degree from Northwestern University in 2005. He is a co-recipient of seven Best Paper Awards, including the IEEE Marconi Prize Paper Award in Wireless Communications in 2011. He has co-authored four books: *Wireless Network Pricing*, *Monotonic Optimization in Communication and Networking Systems*, *Cognitive Mobile Virtual Network Operator Games*, and *Social Cognitive Radio Networks*. He has served as an Editor of *IEEE Transactions on Cognitive Communications and Networking*, the *IEEE Journal on Selected Areas in Communications* Cognitive Radio Series, and *IEEE Transactions on Wireless Communications*. He has served as a Guest Editor of *IEEE Transactions on Smart Grid*, *IEEE Journal on Selected Areas in Communications*, *IEEE Communications Magazine*, and *IEEE Network*. He has served as Associate Editor-in-Chief of *IEEE Communications Society Technology News*, Chair of the IEEE Communications Society Multimedia Communications Technical Committee, and Vice Chair of the IEEE Communications Society Cognitive Network Technical Committee. He is a a Distinguished Lecturer of the IEEE Communications Society.

FULL DUPLEX COMMUNICATIONS

Geoffrey Y. Li

Mehdi Bennis

Guanding Yu

With the rapidly growing demand for mobile applications, future wireless networks are facing considerable challenges in spectral efficiency. To deal with it, many techniques have been developed recently. Among them, in-band full duplex (IBFD), which enables a device to transmit and receive simultaneously at the same frequency, has ignited great interest from both academia and industry. IBFD can potentially double the spectral efficiency in the physical layer and also bring many advantages in high-layer design.

However, to fulfill the minimum performance requirement for the current cellular networks, self-interference cancellation in IBFD communications is still a great challenge to the physical layer design. In addition, some system-level protocols must be redesigned since devices need to send and receive signals at the same frequency simultaneously.

The objective of this Feature Topic is to provide some recent advances and design challenges on IBFD communications. Among a large number of submissions to this Feature Topic, we have carefully selected seven articles that cover various issues from hardware design to system-level protocols.

The first two articles mainly focus on recent advances and challenges related to the hardware design for self-interference cancellation. The article by Mikko Heino *et al.* presents a novel antenna design using resonant wavetraps. It is demonstrated that around 100 dB self-interference cancellation can be achieved by combining the developed antenna design and a digital canceller in a practical system. The article by Leo Laughlin *et al.* discusses how to use electrical balance duplexers to cancel self-interference, where various electrical balance duplexer architectures are compared, and some perspective research opportunities are presented.

The next three articles focus on the medium access control (MAC) layer design of IBFD systems. The article by Karaputugala Madushan Thilina *et al.* first highlights some immediate challenges in both physical-layer and MAC-layer design in IBFD communications, and then provides a thorough survey on the existing full duplex MAC protocols. Some future research directions on this topic are also discussed. The article by Sanjay Goyal *et al.* investigates scheduling in full duplex cellular networks. It is shown that the capacity of the system with IBFD can be almost doubled compared to that with half-duplex by a proposed proportional fairness based scheduler despite self-interference. The energy efficiency of full duplex systems is also addressed in that article. The article by Zhongshan Zhang *et al.* proposes an opportunistic decode-and-forward (DF)-based relay selection scheme for full duplex relay networks, which substantially reduces the outage probability even though there are still some challenges for future research.

The last two articles investigate the applications of IBFD to cognitive radio and device-to-device (D2D) networks. The article by Yun Liao *et al.* introduces a new design approach for full duplex cognitive radio networks, where the secondary user can simultaneously sense and access the vacant spectrum. The article by Li Wang *et al.* provides a comprehensive overview on how to implement full duplex communications in heterogeneous networks and how to integrate full duplex radios in D2D communications. The article also discusses applications of the full duplex techniques in future 5G networks.

Finally, the Guest Editors would like to thank all authors for their submissions even if we are unable to accommodate all of them in the Featured Topic. We greatly appreciate all the reviewers for their timely and professional reviews. We also acknowledge the support and guidance from the Editor-in-Chief of *IEEE Communications Magazine*.

BIOGRAPHIES

GEOFFREY YE LI [F'06] (liye@ece.gatech.edu) is a professor with the School of Electrical and Computer Engineering at Georgia Institute of Technology. He has also held a Cheung Kong Scholar title at the University of Electronic Science and Technology of China since 2006. He was with AT&T Labs — Research for five years before joining Georgia Tech in 2000. His general research interests include wireless communications and statistical signal processing. In these areas, he has published over 300 referred journal and conference papers in addition to 26 granted patents. His publications have been cited around 20,000 times, and he has been listed as one of the World's Most Influential Scientific Minds, also known as a Highly Cited Researcher, by Thomson Reuters. He received the Stephen O. Rice Prize Paper Award in 2010 and the WTC Wireless Recognition Award in 2013 from the IEEE Communications Society, and the James Evans Avant Garde Award in 2013 and the Jack Neubauer Memorial Award in 2014 from the IEEE Vehicular Technology Society.

MEHDI BENNIS received his M.Sc. degree in electrical engineering jointly from EPFL, Switzerland, and the Eurecom Institute, France, in 2002. From 2002 to 2004, he worked as a research engineer at IMRA-EUROPE investigating adaptive equalization algorithms for mobile digital TV. In 2004, he joined the Centre for Wireless Communications (CWC) at the University of Oulu, Finland, as a research scientist. In 2008, he was a visiting researcher at the Alcatel-Lucent Chair on Flexible Radio, SUPELEC. He obtained his Ph.D. in December 2009 on spectrum sharing for future mobile cellular systems. He is the Principal Investigator of the European CELTIC-SHARING project on multidimensional HetNets. His main research interests are in radio resource management, heterogeneous networks, game theory, and machine learning in the context of heterogeneous and small cell networks. He has published more than 50 research papers in international conference proceedings, journals, and book chapters.

GUANDING YU received B.E. and Ph.D. degrees in communication engineering from Zhejiang University, Hangzhou, China, in 2001 and 2006, respectively. After that, he joined Zhejiang University, where he is an associate professor in the Department of Information and Electronic Engineering. Since July 2013, he has also been a visiting professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta. His research interests include energy-efficient wireless communication system design, device-to-device communications, and small cell techniques. He served as General Co-Chair of the CRNet 2010 Workshop, TPC Co-Chair for the Green Communications and Computing Symposium of Chinacom 2013, and a TPC member for various academic conferences. He is serving as a Guest Editor of the *IEEE Communications Magazine* Feature Topic on Full Duplex Communications and a Guest Editor of the *IEEE JSAC* Series on Green Communications and Networking.

Recent Advances in Antenna Design and Interference Cancellation Algorithms for In-Band Full Duplex Relays

Mikko Heino, Dani Korpi, Timo Huusari, Emilio Antonio-Rodríguez, Sathya Venkatasubramanian, Taneli Riihonen, Lauri Anttila, Clemens Icheln, Katsuyuki Haneda, Risto Wichman, and Mikko Valkama

ABSTRACT

In-band full-duplex relays transmit and receive simultaneously at the same center frequency, hence offering enhanced spectral efficiency for relay deployment. In order to deploy such full-duplex relays, it is necessary to efficiently mitigate the inherent self-interference stemming from the strong transmit signal coupling to the sensitive receive chain. In this article, we present novel state-of-the-art antenna solutions as well as digital self-interference cancellation algorithms for compact MIMO full-duplex relays, specifically targeted for reduced-cost deployments in local area networks. The presented antenna design builds on resonant wavetraps and is shown to provide passive isolations on the order of 60–70 dB. We also discuss and present advanced digital cancellation solutions, beyond classical linear processing, specifically tailored against nonlinear distortion of the power amplifier when operating close to saturation. Measured results from a complete demonstrator system, integrating antennas, RF cancellation, and nonlinear digital cancellation, are also presented, evidencing close to 100 dB of overall self-interference suppression. The reported results indicate that building and deploying compact full-duplex MIMO relays is already technologically feasible.

Mikko Heino, Sathya Venkatasubramanian, Taneli Riihonen, Emilio Antonio-Rodríguez, Clemens Icheln, Katsuyuki Haneda, and Risto Wichman are with Aalto University.

Dani Korpi, Timo Huusari, Lauri Anttila, and Mikko Valkama are with Tampere University of Technology.

INTRODUCTION

The industry envisions that upcoming fifth generation (5G) radio communication systems will provide 1000-fold throughput compared to current 4G systems, while in-band full-duplex technology [1] may only double the spectral efficiency at best. This may sound modest given the ambitious targets of 5G systems, but somebody with a can-do attitude would think instead "great, we are already halfway through." Nevertheless, attaining this target requires a combination of several different physical and network layer techniques that without in-band full-duplex technology 5G networks would not reach their full potential.

An in-band full-duplex capable transceiver is able to transmit and receive simultaneously over the same center frequency [1]. With this definition, we can identify three basic communication scenarios of two or three nodes that cover any generic wireless network (e.g., the one shown at the bottom of Fig. 1) when they are freely combined together. In particular, the possible use cases benefiting from full-duplex operation include bidirectional device-to-device transmission between two full-duplex user terminals, a full-duplex access point serving uplink and downlink (possibly half-duplex) users simultaneously in the same radio resource, and two-hop full-duplex wireless relaying (as shown in the middle of Fig. 1). Actually, the latter two use cases are topologically the same with only the conceptual difference that relays always forward their received information, while the uplink user is not likely communicating with the downlink user but with somebody else in a totally different cell. In addition to increasing spectral efficiency, full-duplex capable nodes may decrease latency within multihop links and avoid the guard period that is necessary when switching the direction of the transmission from downlink to uplink in time-division duplex (TDD) systems. At the network level, full-duplex nodes may simultaneously process user plane and control plane signals, decreasing the latency and boosting the operation of the overall system.

In practice, the increase in throughput due to full-duplex operation is limited by the presence of unavoidable self-interference when the transmitted signal couples back to the receiver in the in-band full-duplex transceiver. Even when the transmitted signal is known in digital baseband, it cannot be eliminated completely in the receiver because of RF impairments [2], and a large power difference between the transmitted and received signals. On the other hand, RF cancellation is limited due to the simplified processing capabilities of electronics when compared to digital signal processing. Nevertheless, 70–100 dB overall attenuation levels of self-interference have been reported in the literature [16, 17], and

 0163-6804/15/$25.00

Nevertheless, attaining this target requires a combination of several different physical and network layer techniques and without in-band full-duplex technology 5G networks would not reach their full potential.

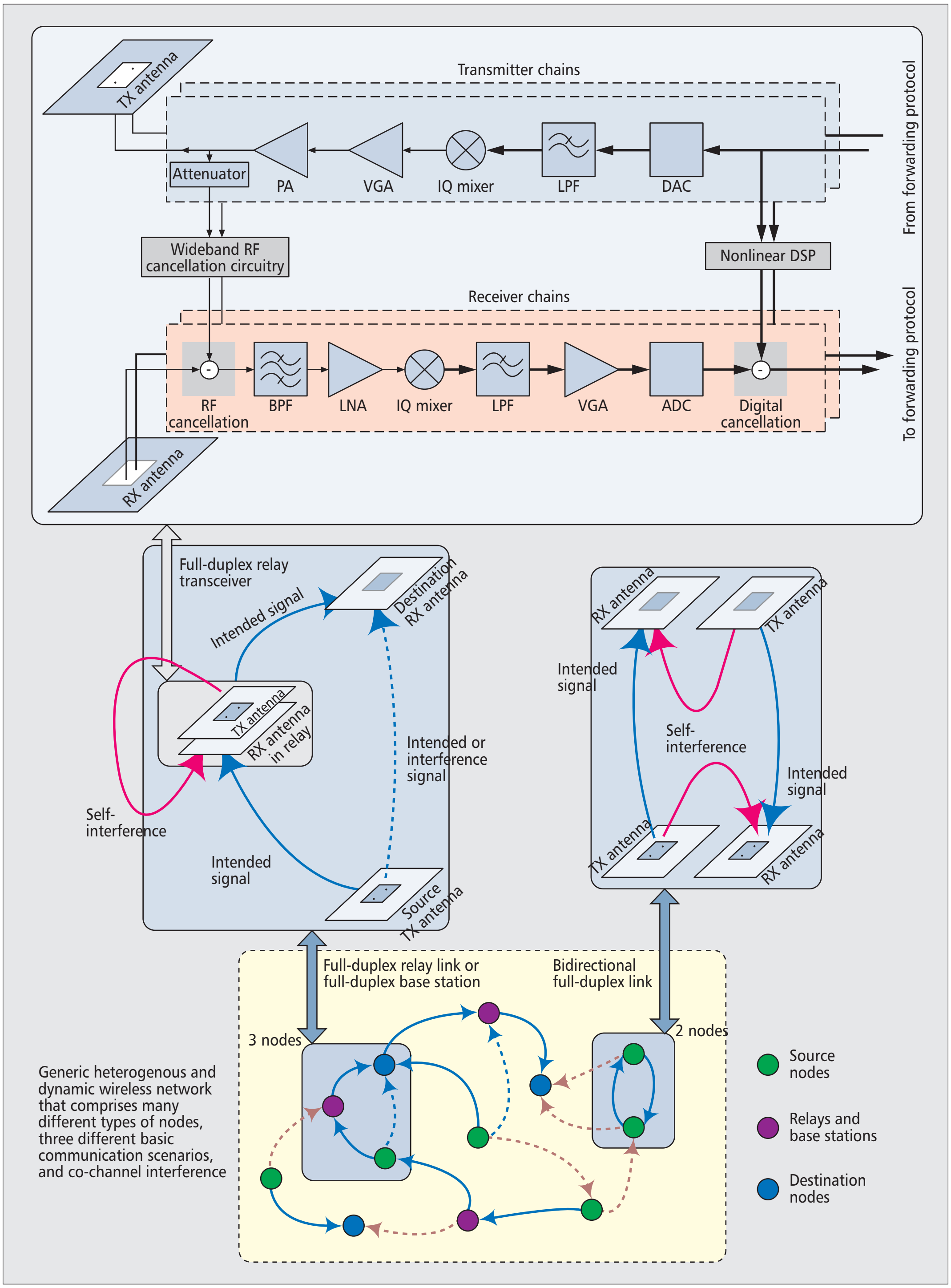


Figure 1. A multi-antenna radio transceiver, and basic two- and three-node full-duplex scenarios occurring in wireless networks.

holistic cancellation techniques are currently subject to intensive research by many groups around the world, including our own consortium. Such attenuation levels are already so high that in local area networks, where transmit powers are limited and nodes may be close to each other, co-channel interference may already be starting to dominate over self-interference, which is our objective as well.

However, even if self-interference is suppressed below receiver noise or ambient co-channel interference, a full-duplex transceiver may outperform its half-duplex counterpart only when there is simultaneous balanced traffic in both uplink and downlink. Obviously, two-way frequency reuse is useless if data mostly flows only in a single direction. In fact, recent statistics from cellular network vendors and operators report commonly up to 8:2 or 9:1 traffic ratios between downlink and uplink, challenging the applicability of full-duplex access points[1, 2]. Ideal full-duplex operation would yield at best only 25

[1] Ericsson, "On the Pulse of the Networked Society," Mobility Report, Nov. 2012, http://www.ericsson.com/res/docs/2012/ericsson-mobility-report-november-2012.pdf

[2] Nokia Solutions and Networks, "TD-LTE Frame Configuration Primer," white paper, http://networks.nokia.com/system/files/document/nsn_td_lte_frame_configuration_wp.pdf

or 11 percent throughput improvement with the above traffic ratios, respectively, in comparison to the wanted 100 percent gain. In contrast, *the traffic in full-duplex relaying is inherently symmetric*, because relays are supposed to always forward the same amount of information as they receive, making wireless relays the most prominent use case for in-band full-duplex operation from the start.

In this article, we explore various sophisticated techniques for self-interference mitigation and cancellation within multi-antenna in-band full-duplex relays of compact size suitable for local area networks. By combining antenna design with novel RF and digital cancellation, we show that the self-interference can be suppressed below noise level even when using regular low-cost components. We report measured results from a complete demonstrator system, integrating advanced antenna designs and analog and digital cancellation, evidencing close to 100 dB of overall isolation between transmitter and receiver chains. Thus, in-band full-duplex relays have the potential to significantly improve the performance in local area networks.

Implementation-Level Research Challenges in In-Band Full-Duplex Transceivers

As already discussed, one of the most crucial issues in wireless single channel full-duplex communications is the inherent transmit signal, or self-interference (SI), which is coupled back to the receiver and acts as a strong source of interference. The SI signal can be as much as 60–100 dB more powerful than the weak received signal of interest, and therefore must be attenuated significantly to allow detection of the actual received signal and enable in-band full-duplex communications in the first place. In theory, this is straightforward as the inherent transmit signal is known within the device, and thus the SI can in principle be perfectly cancelled merely by subtracting the original transmit signal, properly filtered, from the received signal. In practice, however, achieving a sufficient amount of SI cancellation calls for more advanced and elaborate measures since all the sources of distortion in the transmitter and receiver must be accounted for when subtracting the cancellation signal from the received signal.

Passive Isolation and Active Cancellation

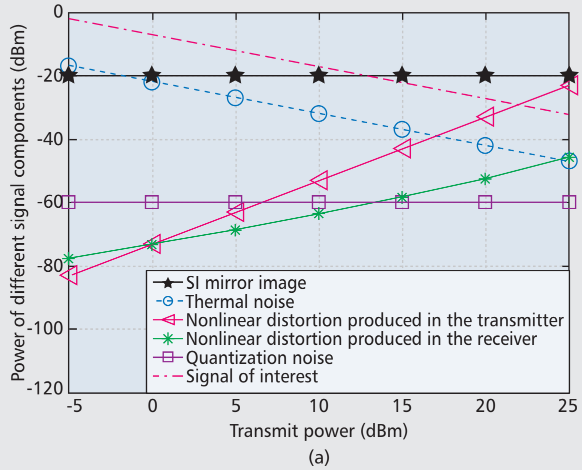
To minimize the amount of the more complicated active cancellation techniques, the passive isolation between the transmitter and receiver should be maximized. This means that the power of the SI leaking to the receiver is smaller in the first place, so not as much active cancellation is needed to attenuate it to a tolerable level. When separate transmit and receive antennas are used, the electromagnetic isolation between the antennas can be improved in a straightforward manner by increasing the spacing between the antennas or using different polarizations, resulting in lower SI power. The main limitation in increasing the spacing is that the transceivers are usually space limited, and hence, we do not have the luxury of increasing the distance between the antennas. Moreover, when multiple-input multiple-output (MIMO) is used to improve the link capacity, both polarizations may be used to provide polarization diversity. Thus, advanced methods are required to improve the electromagnetic isolation between the antennas. These methods include, for instance, the use of band-gap structures as high-impedance surfaces [3] to prevent surface waves between the antennas, inductive loops [4] to produce counter-flowing magnetic fields to reduce the coupling between the antennas, and resonant structures like wavetraps [5] and slots on a ground plane [6] to alter ground plane currents to reduce the coupling. Because these techniques are completely passive in nature, they require no tracking of the SI signal and its possible distortion, while still providing a potentially significant increase in the amount of SI attenuation. In addition to the aforementioned methods to improve the antenna isolation, the port-to-port isolation between the antenna feeds can be improved by using techniques like neutralization [18], which improves the isolation using a cancellation path between two antennas, or connecting lumped elements between antenna feeds [19] to cancel the mutual admittance between the antennas, thereby improving the port-to-port isolation. In this article, one possible method of using resonant wavetraps is discussed and demonstrated to improve the electromagnetic isolation in compact devices, especially relays.

It is, however, typically not possible to mitigate the SI signal perfectly with passive antenna-based techniques, so active SI cancellation is also required. A common solution is to do the active attenuation of the SI signal in two stages: first at the input of the receiver chain, operating at RF frequencies, and then after the analog-to-digital conversion. These SI cancellation stages are usually referred to as RF cancellation and digital cancellation, respectively. RF cancellation is required in order to prevent the complete saturation of the receiver components and the analog-to-digital converter (ADC). Finally, digital cancellation is performed to attenuate the remaining SI signal below the noise floor. Both of these active cancellation methods rely on processing the known transmit signal to produce the cancellation signal.

A block diagram of an in-band MIMO full-duplex relay employing the aforementioned active SI cancellation stages is illustrated in the upper part of Fig. 1. The transmitter and receiver chains have been chosen to follow direct-conversion architecture, which is a typical selection for modern wireless transceivers. This block diagram depicts the transceiver model that will be utilized in the following chapters, thereby acting as a basis for the forthcoming analysis. In this article, the emphasis is on a full-duplex relay, meaning that the transmitters and receivers have separate antennas, which can also be seen in Fig. 1. It should be noted, however, that it is in principle also possible to share an antenna between a transmitter and a receiver, even in an in-band full-duplex transceiver [7].

Achieving a sufficient amount of SI cancellation calls for more advanced and elaborate measures since all the sources of distortion in the transmitter and receiver must be accounted for when subtracting the cancellation signal from the received signal.

> *The most significant impairments include the nonlinear distortion of the SI signal, caused by the different amplifiers in the transmitter and receiver, I/Q imbalance of the transmitter and receiver, phase noise of the local oscillator, and ADC quantization noise.*



Parameter	Value
SNR	15 dB
Bandwidth	12.5 MHz
RX noise figure	4 dB
Antenna isolation	40 dB
RF cancellation	30 dB
TX power amplifier IIP3	13 dBm
I/Q image rejection ratio	30 dB
RX low-noise amplifier IIP3	-9 dBm
Number of bits at the ADC	12

(b)

Figure 2. a) Example power levels of the different signal components at the input of the receiver detector. In this example, it is assumed that the linear SI component can be perfectly suppressed; b) example component parameters and other essential values used in evaluating the power levels.

RF Imperfections

Typically, the performance of the active SI cancellation mechanisms based on linear processing is limited by the non-idealities occurring within the analog/RF circuit of the full-duplex transceiver [2]. This applies especially to the SI cancellation occurring in the digital domain, where only the original transmit waveform is available. For this reason, the most prominent types of such analog/RF circuit non-idealities must be considered in this context, as they have to be regenerated in the cancellation processing to produce a sufficiently accurate cancellation signal. The most significant impairments include the nonlinear distortion of the SI signal caused by the different amplifiers in the transmitter and receiver, the in-phase/quantization (I/Q) imbalance of the transmitter and receiver, phase noise of the local oscillator, and ADC quantization noise. Some of these RF impairments can be modeled on the waveform level, with appropriate behavioral models, within digital cancellation, and therefore do not pose an insurmountable obstacle for in-band full-duplex communications. Good examples of recent work in this field are [2, 4, 8–10, 15, 16]. In particular, the nonlinear distortion of the SI signal can be modeled in the digital domain using a memory polynomial or parallel Hammerstein type of model with carefully chosen coefficients [8, 10]. With such a model, the clean transmit data can be processed to create an accurate SI replica for efficient cancellation. Correspondingly, the effect of I/Q imbalance can be modeled digitally as an additional distortion component, which consists of the complex conjugate of the original baseband transmit signal with certain memory coefficients [9]. Nevertheless, there are also certain sources of distortion that cannot be dealt with afterward, such as thermal noise and quantization errors produced by the ADC. These must already be accounted for in the design process of the transceiver [2].

To illustrate the need for modeling the most dominant impairments at the waveform level, Fig. 2a shows the absolute power levels of the different signal components at the receiver detector input, with respect to the transmit power. The power levels have been calculated for a single-input single-output (SISO) in-band full-duplex transceiver using simplified system calculations, similar to [2], and using realistic RF component and ADC specifications. In this type of analysis, only the power levels of the different signal components are taken into account, which simplifies the calculations considerably while still providing useful insight into the system behavior. In this example, it is assumed that only the linear component of the SI signal can be perfectly attenuated by the different SI cancellation stages. To get the final numerical values for the power levels, typical component parameters from earlier literature and technical specifications have then been used, the most important of which are shown in the table in Fig. 2b for readers' convenience. Based on Fig. 2a, it is evident that at least the transmitter power amplifier induced nonlinearities, alongside the SI mirror image produced by the I/Q imbalances, must be attenuated by the digital cancellation algorithm to enable reliable in-band full-duplex operation and maintain an acceptable signal-to-interference-plus-noise ratio (SINR). Thus, simple linear modeling and processing in the digital domain does not suffice, and more advanced algorithms are required.

In addition to the analog circuit-level impairments, one additional aspect to consider is the time varying nature of the SI propagation channel. This is often overlooked when discussing SI cancellation in full-duplex devices, but in practice it is obviously a key aspect in implementing an in-band full-duplex transceiver. The time varying nature of the SI channel becomes especially crucial in relaying applications where the transmit and receive antennas can be physically separated to achieve higher natural isolation between them. In such a scenario, the SI channel can vary significantly in time. However, even in a single-antenna full-duplex device, the reflections

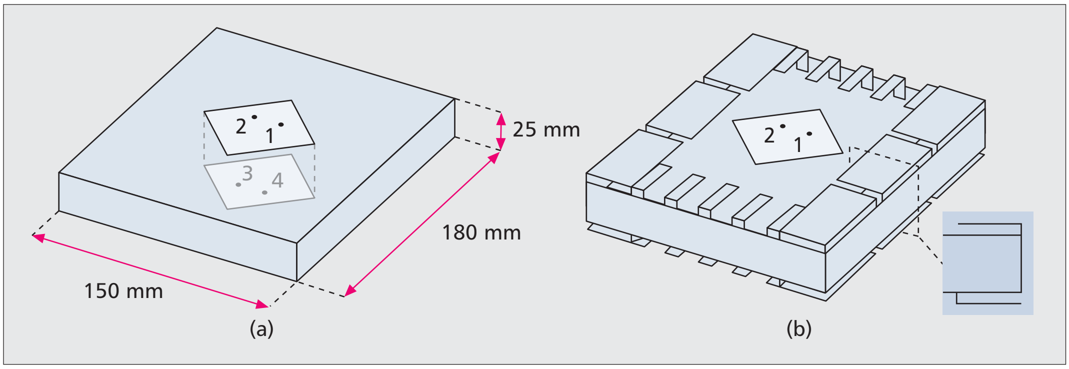


Figure 3. A compact in-band full-duplex relay antenna: a) without wavetraps; b) with wavetraps. In b), wavetraps are facing opposite directions on either side of the relay.

Even though antenna isolation and RF cancellation provide a significant level of SI mitigation before digital processing is applied, the residual SI after ADC conversion may still be strong enough to desensitize the receiver. Therefore, the use of additional cancellation techniques in the digital domain is essential.

from nearby scatterers can change significantly from one moment to the next, thus warranting the use of adaptive SI cancellation solutions.

When considering a complete and functional in-band full-duplex relay, it is more than likely that all of the previously discussed methods will be involved in order to obtain a sufficient amount of SI attenuation. First, the passive isolation between the transmitter and receiver is maximized, after which the SI signal is actively attenuated. In the digital domain, the possible circuit-level impairments are then regenerated with advanced algorithms and cancelled efficiently. Furthermore, these algorithms also have the necessary adaptivity to constantly track the changes in the SI channel parameters. To demonstrate that such a device is quickly becoming a realistic prospect, this article provides some of the latest research results, alongside with measurement data, of these individual key factors, which are necessary to implement a functional in-band full-duplex relay.

Advanced Antenna Designs for Enhanced Transmitter-Receiver Isolation

In this section, the use of resonant wavetraps is discussed as a possible method of improving the electromagnetic isolation in a compact device, such as a back-to-back relay with transmit and receive antennas on either side of a box, as shown in Fig. 3. In Fig. 3a, a closed box of dimensions 180 × 150 mm with dual polarized patch antennas (50 × 50 mm) on the top and bottom is shown. The use of a closed box represents a practical scenario with the electronics contained inside the relay. The patch antenna on the bottom side is co-located with the top side patch antenna, but with the antenna feed locations on the opposite sides of the patch (Fig. 3a). The patch antennas are dual polarized in order to support MIMO. The relay can be used, for example, as an outdoor-indoor relay as discussed in [4]. In order to enable in-band full-duplex operation, the electromagnetic isolation between the antennas on opposite sides of the box has to be improved. Figure 3b shows a schematic of the relay antenna with wavetraps connected to the ground plane.

Wavetraps are resonant structures that can be used to control the surface currents on the ground planes. In this case, a planar quarter-wavelength patch short-circuited at the other end from the ground plane is used as a wavetrap. According to transmission line theory, the input impedance of a short-circuited quarter-wave transmission line is very high. This results in small currents at the open end of the patch with large currents concentrated within the patch. This reduces the surface currents and electromagnetic fields on the other side of the relay box, hence improving the isolation between the upper and lower antennas of the box.

In this work, the patch antennas of the relay are resonant at 2.56 GHz with a –10 dB impedance bandwidth of 139 MHz and a –6 dB bandwidth of 238 MHz. They can be used, for example, for WiMAX or LTE. In the relay structure, ports 1 and 2 (TX) are on the top, and ports 3 and 4 (RX) on the bottom, as shown in Fig. 3a. The wavetraps are then implemented as quarter-wave patches as shown in Fig. 3b with one end of the patch short-circuited to the ground plane. The dimensions and number of wavetraps were optimized with EM-field simulations to obtain the best possible isolation over a sufficient frequency band between the TX and RX antenna ports.

The resonant wavetraps on the longer edge have larger width and are closely spaced, whereas along the shorter edge, the width is smaller and the wavetraps are more sparsely placed. This way, the best minimum isolation defined as the worst isolation among the four different combinations of the RX and TX ports (i.e., S_{31}, S_{41}, S_{32}, and S_{42}) is obtained. The wavetraps on both sides of the relay are similar but are oriented in opposite directions, that is, the open ends of the wavetraps are located close to the antenna on the TX side and away from the antenna on the RX side, as seen in Fig. 3b. The best isolation between the antenna ports was obtained with this particular asymmetry.

Figures 4a and 4b illustrate the simulated average electric field strength at the center cross-section parallel to the shorter edge of the box at 2.56 GHz with and without wavetraps, respectively, when port 1 of the antenna is excited. From Fig. 4a, it can be observed that the transmitted electromagnetic wave diffracts strongly around the corners of the relay enclo-

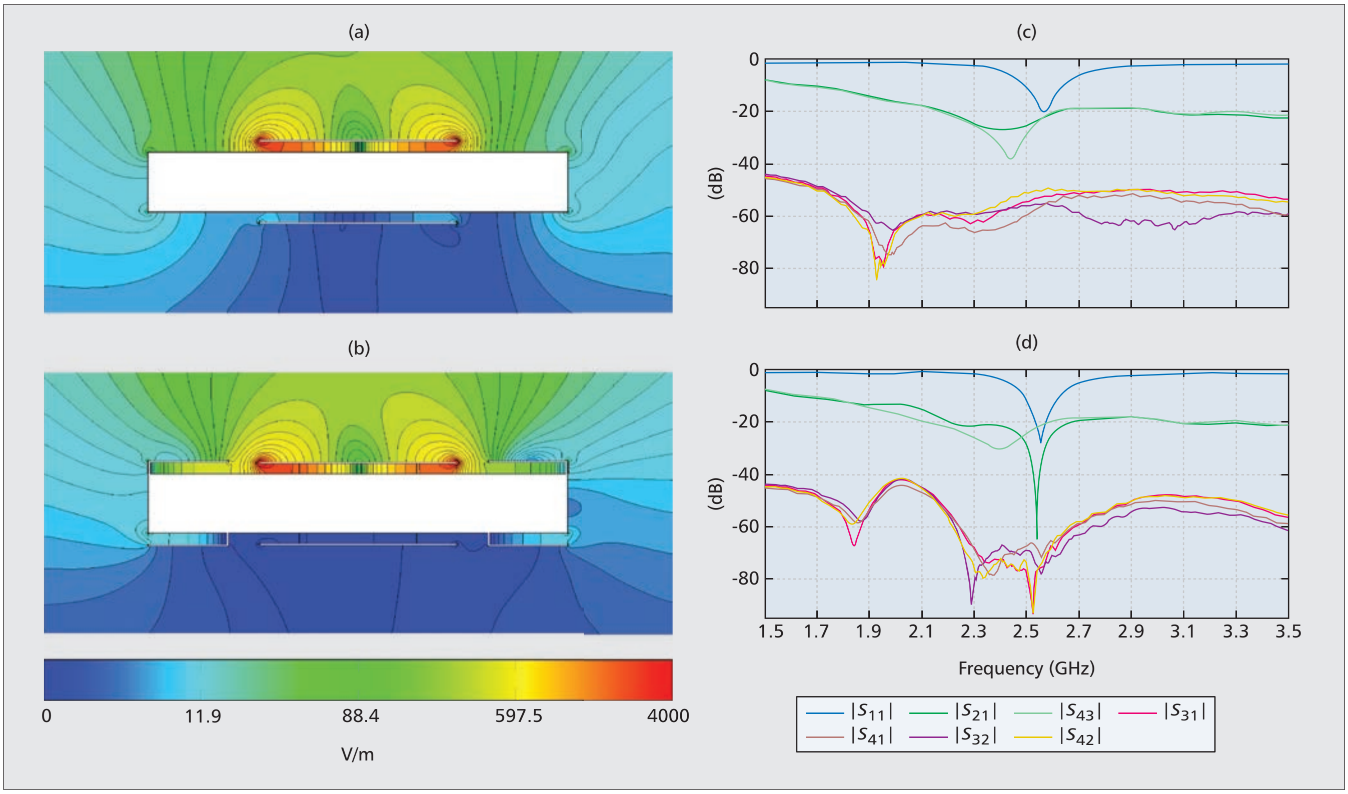


Figure 4. a) Simulated electric field strength without wavetraps; b) with wavetraps; c) measured parameters of the relay without wavetraps; d) with wavetraps.

sure and couples to the receiving side antenna. With the wavetraps, as shown in Fig. 4b, the electric field strength is substantially lower on the receiving side of the relay than without the wavetraps, resulting in clearly improved isolation.

The designed relay antenna was manufactured and then measured in an anechoic chamber. The isolation was evaluated by measuring the S-parameters of the relay antenna as a four-port network. The measured S-parameters are presented in Figs. 4c and 4d, where S11 denotes the antenna input matching, and S21 and S43 the coupling between the antenna ports on the same side of the box. The remaining S-parameters show the isolation between the antenna ports on opposite sides of the box. In the relay without wavetraps, the minimum isolation between RX and TX ports is 50 dB at the resonance frequency of the patch antenna, as shown in Fig. 4c. After adding the wavetraps, the minimum isolation improved by 21 dB to 71 dB at the resonance frequency. In addition, the wavetraps also offer a wideband isolation improvement around the designed frequency, even though they are frequency-sensitive. The isolation across a 167 MHz band around the center frequency is better than 65 dB, which is an improvement of 15 dB over the 50 dB minimum isolation without wavetraps as indicated by the dashed horizontal line in Fig. 4d. Within the given isolation bandwidth, the corresponding antenna port matching is better than –6 dB to ensure enough power radiated by the antennas. The gain of the relay antenna was also measured in the anechoic chamber to determine the impact of the wavetraps on the radiation pattern. The impact of adding wavetraps on the ground plane was minor with a maximum gain of 10.2±0.1 dBi for all ports of the relay without wavetraps and 9.6 dBi for ports 1 and 2 and 10.5±0.1 dBi for ports 3 and 4 for the relay with wavetraps. In conclusion, the wavetraps efficiently concentrate the ground plane currents and reduce the coupling on the other side of the relay box, thereby improving the antenna isolation.

Advanced Self-Interference Cancellation Algorithms

Even though antenna isolation and RF cancellation provide a significant level of SI mitigation before digital processing is applied (~60–80 dB), the residual SI after ADC may still be strong enough to desensitize the receiver [2, 10]. Therefore, the use of additional cancellation techniques in the digital domain is essential in order to reduce the remaining interference below noise levels and guarantee the optimal performance of the system.

Basic Principles

As previously explained, cancellation techniques in the digital domain work in a similar way to their analog counterpart; that is, a digital replica of the SI signal is generated and then subtracted from the received signal. The fidelity of the replica signal will determine the level of residual SI after cancellation and therefore will upper bound the system performance [11]. In order to achieve a nearly interference-free system, the selection of a proper cancellation method is of major importance. When both transmit and receive sides of the full-duplex device behave primarily like a linear filter, cancellation is reduced to designing a filter that identifies the SI channel, since the transmitted signal is known at every time instant. However, a linear cancellation architecture cannot mitigate nonlinear behavior and noise sources such as nonlinear distortion of the power amplifier, I/Q imbalance during modulation/demodulation, phase noise, or quantization noise at the receiver [2, 9, 10, 16]. Some of these impairments can be mitigated by extending the linear architecture to a nonlinear one, as described later in this section.

Typically, digital cancellation is performed in either the time or frequency domain. Cancellation in the frequency domain processes each subcarrier signal individually, which may result in a computationally demanding scheme if the number of subcarriers is large. On the other hand, cancellation in the time domain processes the signal samples independent of the number of subcarriers, but, due to the different interference paths between antennas, requires gauging the delay spread of the SI channel [12]. As shown in Fig. 1, cancellation takes place after baseband demodulation and digital conversion of the received signal, usually being the first operation within the digital pipeline. As a result, the employed signal is sampled above the Nyquist rate, which demands the use of special techniques to deal with arbitrary signal spectra [12]. Here we focus entirely on time-domain cancellation techniques.

Adaptive vs. Block-Based Processing

According to the operation mode of the cancellation scheme, we can distinguish between *online* and *offline cancellation*. *Online cancellation* uses an adaptive approach to obtain the optimal cancellation filter, in which every new sample is employed to iteratively update the cancellation filter. Changes in the environment or in the internal state of the device, such as new coupling paths or reflections, as well as variations in the temperature or transmitted power, can severely modify the SI channel; consequently, an updated estimation of the SI channel is essential to sustain optimal interference levels. An online cancellation scheme can track temporal variations of the device environment while maintaining a constant computational load, although the number of samples required to reach an optimal solution might be high [11]. On the other hand, *offline cancellation* obtains the optimal cancellation filter through a batch operation after receiving several samples. This approach requires fewer samples to reach a solution, but the tracking capabilities are compromised, and the computational load per sample is significantly higher than in online cancellation. A periodic estimation of the SI channel is necessary to maintain good performance levels when considering time-varying scenarios [8]. In general, the number of samples available for cancellation indicates which option to implement. An offline algorithm is preferable when few samples are available, whereas an online algorithm is preferable when a large number of samples is available.

Regardless of the operation mode, the optimal cancellation filter is obtained as the solution of a minimization problem. Assuming that the transmitted signal of the full-duplex device is uncorrelated with the incoming source signal, the optimal cancellation filter minimizes the signal power after cancellation or, equivalently, identifies the SI channel [11]. Consequently, the optimal filter can be obtained using common adaptive techniques like gradient descent or recursive least squares in the online cancellation case, or using, say, conventional least squares in the offline cancellation case. For detailed algorithm descriptions, refer to [10, 11].

Linear vs. Nonlinear Cancellation

In general, the underlying architecture of the digital cancellation filter can be classified into two types: linear and nonlinear cancellation. Whereas *linear cancellation* assumes that the self-interference channel can be modeled as a linear filtering operation, *nonlinear cancellation* takes into account the presence of nonlinear components in the transmission/reception chains. As seen in the system calculations example in Fig. 2a, impairments such as power amplifier (PA) nonlinearity and I/Q imbalance may be significant.

Nonlinear effects like amplifier distortion or mixer nonlinearities can be accurately modeled using polynomial-based systems [8, 10, 16], whereas I/Q imbalance can be modeled using widely linear filters [9], both of which have been extensively studied in the recent literature. In brief, polynomial-based systems model nonlinearities by processing higher-order terms of the input signal (i.e., $\{x_n, x_n|x_n|^2, x_n|x_n|^4, \ldots\}$) when interpreted for complex-valued signals and odd-order nonlinearities [8, 10], whereas widely linear systems model I/Q imbalance by separately filtering the input signal and its complex conjugate (i.e., $\{x_n, x_n^*\}$) [9].

As a consequence of performing the cancellation in the time domain, the self-interference propagation channel from each transmit antenna to each receive antenna is modeled using tapped delay lines, so the overall nonlinear model for a full-duplex device will consist of a linear combination of polynomial basis functions [10]. In general, in a full-duplex MIMO device with N transmit antennas and M receive antennas, $N \times M$ self-interference channels need to be identified. Note that the number of parameters grows linearly with both N and M, and the number of delay taps.

Figure 5a shows a possible nonlinear canceller architecture for the full-duplex device in Fig. 1, assuming two transmit antennas and two receive antennas (i.e., $N = M = 2$), and a nonlinear architecture based on polynomial models. Function $\phi_p(x_n) = x_n|x_n|^{p-1}$ is the polynomial basis function of order p, and $H_p(z)$ is the linear filter associated with $\phi_p(x_n)$. Such nonlinear architecture, where a static nonlinearity is followed by a linear filter, is called a parallel Hammerstein model. Hammerstein models are typically used for modeling PA nonlinearities, and have been found to be excellent behavioral models in terms of the accuracy-complexity trade-off [13]. In general, the number of param-

Online cancellation uses an adaptive approach for obtaining the optimal cancellation filter, in which every new sample is employed to iteratively update the cancellation filter. Offline cancellation obtains the optimal cancellation filter through a batch operation, after receiving several samples.

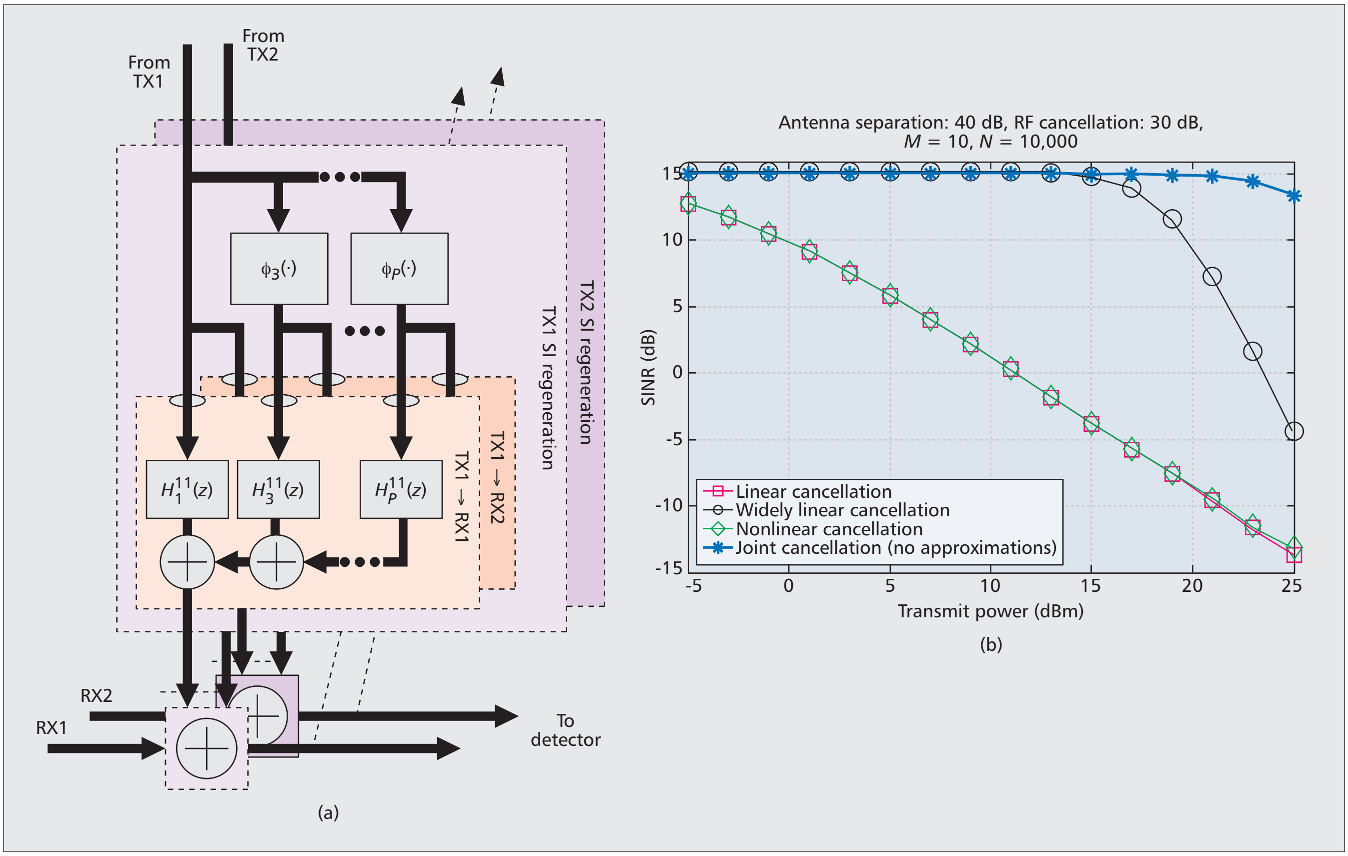


Figure 5. a) Nonlinear adaptive self-interference canceller structure for a 2 × 2 full-duplex MIMO device; b) SINR at canceller output as a function of transmit power in a 2 × 2 MIMO full-duplex device. I/Q modulator mismatch and power amplifier nonlinearity are present. SINR levels with linear, widely linear, and two different nonlinear cancellers are shown. Results from [10] ©IEEE.

eters of a Hammerstein model grows linearly with order p, while in the MIMO case, the increase is relative to $N \times M$ [10]. In [8] the parallel Hammerstein model is successfully used to model the overall nonlinear SI channel comprising a nonlinear PA and a multipath SI channel. While [8] treated the single-antenna case, in [10] this model is extended to the MIMO case, where the I/Q imbalance at the transmit side is also taken into account through the use of additional complex-valued basis functions.

To illustrate the potential of digital cancellation solutions, simulations with the techniques employed in [10] are presented. Figure 5b shows the performance results of different cancellation schemes for a full-duplex device where both transmit and receive signals are 12.5 MHz OFDM signals, and the signal-to-noise ratio at reception, per receiver, is 15 dB. The SI propagation channel between the antennas is a multipath channel, and power amplifier nonlinear distortion and I/Q imbalance during modulation are also included in the system model. Furthermore, an antenna isolation of 40 dB and analog RF cancellation of 30 dB are assumed, similar to the specifications in Fig. 2b. The exact description of the simulation setup can be found in [10]. Concretely, Fig. 5b shows the SINR after digital cancellation for different methods, such as basic *linear cancellation* [11], *nonlinear cancellation* with no I/Q imbalance [8], *widely linear cancellation* with no PA nonlinearity [9], and *nonlinear cancellation* modeling all previous impairments [10], referred to as joint cancellation in the figure. Strictly speaking, the number of basis functions of the joint nonlinear canceller deployed in [10] is 12 times larger than that of the plain linear cancellation scheme, but as shown in [10], only six of these are essential for the implementation. Furthermore, as shorter filters can be used for memory modeling with nonlinear terms, compared to the linear component, the total number of essential parameters to be estimated is still feasible. The nonlinear canceller from [10] is able to obtain SINR results within 1 dB of the optimal case when the transmit power remains below +23 dBm. The widely linear canceller from [9] performs almost optimally when the transmit power is under +15 dBm, while nonlinear cancellation from [8] and plain linear cancellation fail to achieve good performance because of their inability to compensate for the I/Q mismatch, which is the major source of interference in this scenario. Actual RF measurement examples with the parallel Hammerstein-based nonlinear digital canceller in the single antenna case are presented in the next section.

From the simulations results, we can conclude that digital cancellation techniques in combination with antenna isolation and RF cancellation can almost entirely remove the self-interference from the full-duplex device whenever the cancellation architecture is able to accurately model the SI channel. A nonlinear cancellation architecture significantly improves the mitigation levels provided by its linear counterpart when operating with imperfect RF components, while adaptive schemes are able to track variations in the device environment that lead to changes within the SI channel.

Overall Demonstrator System and Performance Evaluations

In order to evaluate the overall SI cancellation performance of a relay transceiver (i.e., a relay antenna with analog/RF and digital SI cancellation), a measurement setup depicted in Fig. 6a is constructed. It consists of a vector signal generator, a PA, the relay antenna, an analog RF canceller, and a spectrum analyzer acting as an I/Q receiver. Digital cancellation processing stages, linear and nonlinear, are implemented on a host computer processor. In the measurements, the aforementioned relay antenna is used, but without the wavetraps, to test the performance limits of the active analog/RF and digital cancellers. The RF center frequency in the measurements is 2.47 GHz.

The analog/RF canceller is a novel design based on a paper by Y.-S. Choi and H. Shirani-Mehr [14]. The canceller consists of two cancellation taps delayed such that the delay of the main SI component lies between the aforementioned canceller taps. The reference signal for these taps is taken from the PA output using a directional coupler. The phase and amplitude of these reference signals are adjusted manually using vector modulators such that a good cancellation level is observed from the spectrum analyzer, indicating that the cancellation signal matches well with the actual SI signal. Cancellation is achieved simply by subtracting the cancellation signal from the RX signal.

The rest of the measurement equipment is standard laboratory equipment, except for the PA, Texas Instruments CC2595, which is a low-cost chip intended for commercial use. It is deliberately chosen in order to demonstrate the performance of the digital nonlinear DSP algorithms under practical circumstances, where the PA distorts the transmit signal heavily. As already discussed, this is inevitable in many applications due to restrictions on the size, cost, and power consumption of the PA. Furthermore, another PA, MiniCircuits ZVE-8G, is also deployed to demonstrate the operation with higher output power and enhanced linearity.

Then Figs. 6b and 6c depict the overall measurement results obtained using the setup of Fig. 6a. The test signal is configured to be a 20 MHz wide orthogonal frequency-division multiplexing (OFDM) signal at 2.47 GHz center frequency with an average transmit power of +20 dBm (TI PA case) or +24 dBm (MiniCircuits PA case), respectively, to test challenging scenarios with fairly high power levels. In Figs. 6b and 6c, the uppermost spectrum is that of the actual transmit signal, measured at the output of the PA. The high distortion levels at the PA output are also evident from the significant spectral regrowth visible outside the actual signal band. This is particularly clear in Fig. 6b, where the used PA is already operating in a highly nonlin-

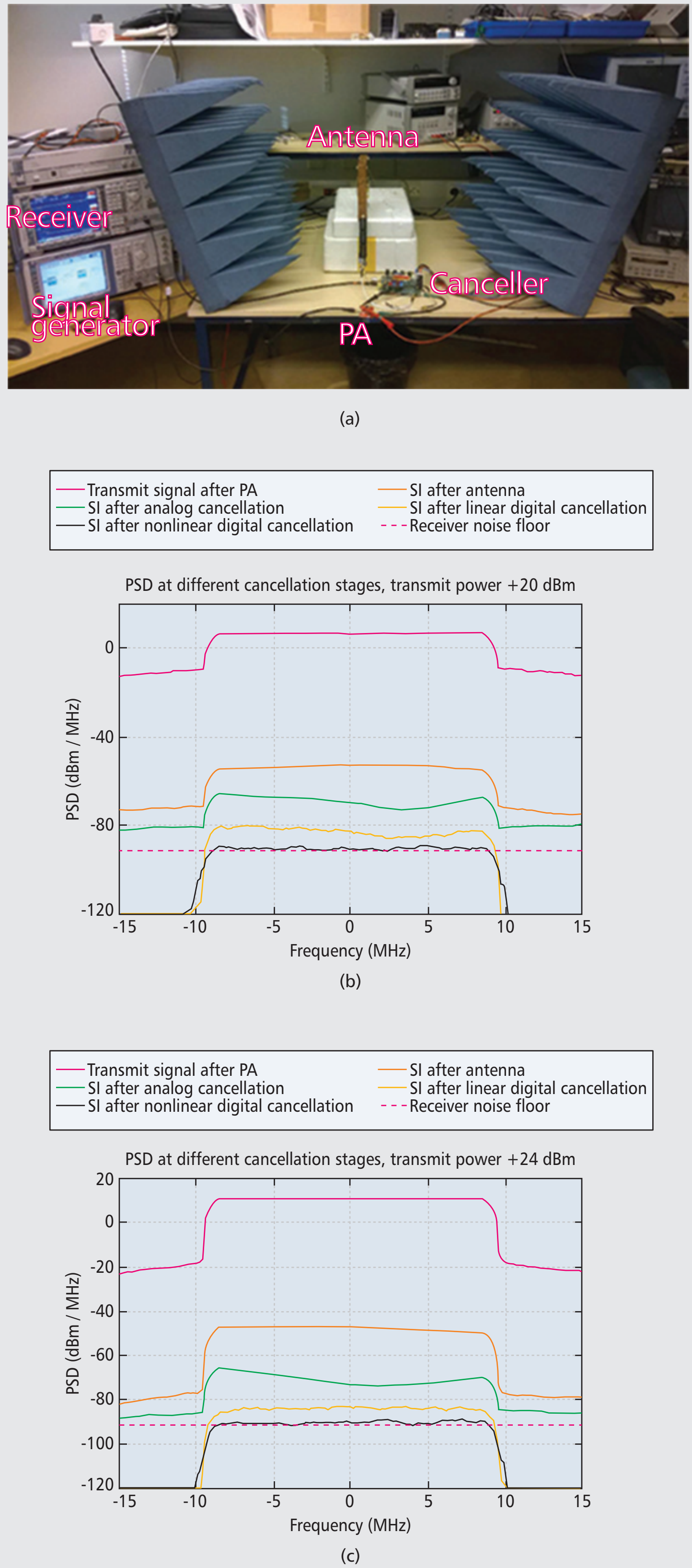


Figure 6. a) Measurement setup and equipment; b) transmit signal and SI signal spectrum at the different cancellation stages for 20 dBm (Texas Instruments CC2595 power amplifier); c) 24 dBm (MiniCircuits ZVE-8G power amplifier) transmit powers. The RF center frequency in the measurements is 2.47 GHz.

The phase and amplitude of these reference signals are adjusted manually using vector modulators such that a good cancellation level is observed from the spectrum analyzer, indicating that the cancellation signal matches well with the actual SI signal. Cancellation is achieved simply by subtracting the cancellation signal from RX signal.

ear region. In Fig. 6c, in turn, the deployed PA is somewhat more linear, but clear nonlinear distortion is also created in this case.

After the PA output, the transmit signal propagates to the receiving antenna, turning into self-interference. From Figs. 6b and 6c it can be observed that the SI signal is attenuated by approximately 60 dB, regardless of the transmit power, when propagating to the receiver, including also 4 dB of losses from the setup. This result is well in line with the observations made in Fig. 4c, where SI attenuations of approximately 55 dB were measured for the antenna without the wavetraps.

The following trace in the figure is the residual SI after analog RF cancellation, which was performed using the aforementioned RF canceller prototype. Depending on the transmit power, the RF canceller is able to attenuate the SI signal by 15 or 22 dB. After this, the data is recorded to digital I and Q samples for offline post-processing with the nonlinear digital cancellation algorithm. In the digital canceller, nonlinear basis functions up to 7th order are considered, each of them having 10 precursor and 10 post-cursor coefficients. The actual parameter estimation is carried out using block-wise least squares, as discussed earlier. From Figs. 6b and 6c it can be observed that the residual SI power after digital nonlinear cancellation is at the IQ receiver noise floor (–91.5 dBm/MHz) with both transmit powers, meaning that the effective interference-plus-noise floor is suppressed by as much as 98 dB in total. This indicates that the total amount of SI cancellation from the antenna, RF canceller, and digital canceller is sufficient to attenuate the SI signal below the receiver noise floor, even with high transmit powers. Also note that with ordinary linear digital cancellation, the power of the residual SI is significantly higher, evidencing that nonlinear modeling is indeed required in the digital cancellation processing of a practical in-band full-duplex transceiver. This is particularly clear in Fig. 6b, where the PA is operating in highly nonlinear region.

Conclusions

In this article, we explore sophisticated techniques for self-interference mitigation and cancellation within multi-antenna in-band full-duplex relays of compact size suitable for local area type networks. A novel antenna design utilizing resonant wavetraps is reported, and it is shown to provide substantially enhanced passive isolation, despite its compact size. We also present novel nonlinear and adaptive digital cancellation algorithms, which can enable enhanced digital self-interference cancellation levels when operating under practical nonlinear RF components. Furthermore, by combining the advanced antenna design with active RF and the proposed digital canceller, we show with actual RF measurements that the self-interference can be suppressed below the receiver noise floor, even when using regular low-cost components. The total measured aggregate self-interference suppression obtained in the measurements was close to 100 dB when operating at the transmit power level of +24 dBm. Thus, dealing with the self-interference problem in in-band full-duplex relays is shown to be technologically feasible, and hence has the potential to significantly improve the performance in, for example, local area type networks.

Acknowledgments

The authors wish to thank Prof. Sergei Tretyakov and Dr. Jari Holopainen, Aalto University, and Mr. Petteri Liikkanen, Tampere University of Technology, for technical guidance and assistance leading to these results. We also wish to thank Dr. Yang-seok Choi, Intel Corporation, for extensive technical consultation and advice related to active RF canceller prototype implementation.

The research work leading to these results was funded by the Academy of Finland (under the project #259915 ”In-band Full-Duplex MIMO Transmission: A Breakthrough to High-Speed Low-Latency Mobile Networks”), the Finnish Funding Agency for Technology and Innovation (Tekes, under the project “Full-Duplex Cognitive Radio”), the Linz Center of Mechatronics (LCM) in the framework of the Austrian COMET-K2 programme, and Emil Aaltonen Foundation. The research was also supported by the Internet of Things program of DIGILE (Finnish Strategic Centre for Science, Technology and Innovation in the field of ICT), funded by Tekes.

References

[1] A. Sabharwal *et al.*, “In-Band Full-Duplex Wireless: Challenges and Opportunities,” *IEEE JSAC*, vol. 32, no. 9, Oct. 2014, pp. 1637–52.
[2] D. Korpi *et al.*, “Full-Duplex Transceiver System Calculations: Analysis of ADC and Linearity Challenges,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, July 2014, pp. 3821–36.
[3] H. Kang and S. Lim, “High Isolation Transmitter and Receiver Antennas Using High-Impedance Surfaces for Repeater Applications,” *J. Electromagnetic Waves and Applications*, vol. 27, no. 18, 2013, pp. 2281–87.
[4] D. Korpi *et al.*, “Advanced Self-Interference Cancellation and Multiantenna Techniques for Full-Duplex Radios,” *2013 Asilomar Conf. Signals, Systems and Computers*, 3–6, Nov. 2013, pp. 2281–87.
[5] P. Lindberg and E. Ojefors, “A Bandwidth Enhancement Technique for Mobile Handset Antennas Using Wavetraps,” *IEEE Trans. Antennas and Propagation*, vol. 54, no. 8, Aug. 2006, pp. 2226–33.
[6] M. Karaboikis *et al.*, “Compact Dual-Printed Inverted-F Antenna Diversity Systems for Portable Wireless Devices,” *IEEE Antennas and Wireless Propagation Letters*, vol. 3, no. 1, pp. 9–14, 2004.
[7] M. E. Knox, “Single Antenna Full Duplex Communications Using a Common Carrier,” *Proc. IEEE 13th Annual Wireless and Microwave Technology Conf.*, Apr. 2012, pp. 1–6.
[8] L. Anttila *et al.*, “Cancellation of Power Amplifier Induced Nonlinear Self-Interference in Full-Duplex Transceivers,” *Proc. 47th Asilomar Conf. Signals, Systems and Computers*, Nov. 2013, pp. 1–6.
[9] D. Korpi *et al.*, “Widely-Linear Digital Self-Interference Cancellation in Direct-Conversion Full-Duplex Transceiver,” *IEEE JSAC*, vol. 32, no. 9, Oct. 2014, pp. 1674–87.
[10] L. Anttila *et al.*, “Modeling and Efficient Cancellation of Nonlinear Self-Interference in MIMO Full-Duplex Transceivers,” *Proc. IEEE GLOBECOM Wksps.*, Dec. 2014, pp. 862–68.
[11] E. Antonio-Rodríguez *et al.*, “Adaptive Self-Interference Cancellation in Wideband Full-Duplex Decode-and-Forward MIMO Relays,” *Proc. IEEE 14th Wksp. Signal Proc. Advances in Wireless Communications*, June 2013, pp. 370–74.
[12] E. Antonio-Rodríguez *et al.*, “SINR Optimization in Wideband Full-Duplex MIMO Relays under Limited Dynamic Range,” *Proc. IEEE 8th Wksp. Sensor Array and Multichannel Signal Proc.*, June 2014, pp. 177–80.

[13] A. Tehrani *et al.*, "A Comparative Analysis of the Complexity/Accuracy Tradeoff in Power Amplifier Behavioral models," *IEEE Trans. Microwave Theory and Techniques*, vol. 58, June 2010, pp. 1510–20.

[14] Y.-S. Choi and H. Shirani-Mehr, "Simultaneous Transmission and Reception: Algorithm, Design and System Level Performance," *IEEE Trans. Wireless Commun.*, vol. 12, no. 12, Dec. 2013, pp. 5992–6010.

[15] V. Syrjälä *et al.*, "Analysis of Oscillator Phase-Noise Effects on Self-Interference Cancellation in Full-Duplex OFDM Radio Transceivers," *IEEE Trans. Wireless Commun.*, vol. 13, no. 6, June 2014, pp. 2977–90.

[16] D. Bharadia, E. McMilin, and S. Katti, "Full Duplex Radios," *Proc. SIGCOMM'13*, Aug. 2013, pp. 375–86.

[17] M. Duarte, C. Dick, and A. Sabharwal, "Experiment-Driven Characterization of Full-Duplex Wireless Systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, Dec. 2012, pp. 4296–4307.

[18] S. Venkatasubramanian *et al.*, "Impact of Neutralization on Isolation in Co-Planar and Back-to-Back Antennas," *Proc. 9th Euro. Conf. Antennas and Propagation*, Lisbon, Portugal, Apr.2015, Bi4–3.

[19] S. Venkatasubramanian *et al.*, "On the Constraints to Isolation Improvement in Multi-Antenna Systems," *Proc. 9th Euro. Conf. Antennas and Propagation*, Lisbon, Portugal, Apr. 2015, C37–5.

Biographies

MIKKO HEINO [S'14] received his M.Sc.(tech.) degree from Aalto University School of Electrical Engineering, Espoo, Finland, in 2014, where he is currently pursuing his D.Sc. degree in radio engineering in the Department of Radio Science and Engineering. His current research interests include antenna isolation improvement methods and antenna design towards in-band full-duplex systems. In 2014, Mikko received the second prize in the 2014 IEEE AP-S Student Design Contest as a part of the Aalto ELEC team.

DANI KORPI [S'14] received his B.Sc. and M.Sc. degrees (both with honors) in communications engineering from Tampere University of Technology (TUT), Finland, in 2012 and 2014, respectively. He is currently a researcher in the Department of Electronics and Communications Engineering at TUT, pursuing his D.Sc. (Tech.) degree in communications engineering. His main research interest is the study and development of single-channel full-duplex radios, with a focus on analyzing the RF impairments.

TIMO HUUSARI [S'15] received his B.Sc. degree (with honors) in electrical engineering from TUT, Finland, in 2012. He is currently a research assistant in the Department of Electronics and Communications Engineering at TUT, pursuing his M.Sc. (Tech.) degree in electrical engineering. His main research interest is the study and development of full-duplex radios, with focus on self-interference cancellation device design and implementation.

EMILIO ANTONIO-RODRÍGUEZ received his B.Sc. and M.Sc. degrees in telecommunication engineering and signal processing applications for communications from the University of Vigo, Spain, in 2006 and 2010, respectively. From 2008 to 2012, he worked as a researcher at the Galician Research and Development Center in Advanced Telecommunications (GRADIANT). Since 2012, he has been pursuing the D.Sc. (Tech.) degree in the Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering, Finland. His research interests are related to adaptive signal processing, interference suppression, and signal processing for communications.

SATHYA N. VENKATASUBRAMANIAN received his B.E degree from Anna University, Chennai, India, in 2010, and received his M.Sc.(tech.) degree in January 2013, majoring in radio science and engineering, from Aalto University School of Electrical Engineering. Since 2013 he has been a doctoral candidate in the Department of Radio Science and Engineering at Aalto University School of Electrical Engineering. His current research interests include antennas and propagation toward in-band full-duplex systems.

TANELI RIIHONEN [S'06, M'14] received his D.Sc. degree in electrical engineering (with distinction) from Aalto University in August 2014. Since September 2005, he has held various research positions in the Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering. He is currently a visiting scientist at Columbia University, New York, New York, and serving on the Editorial Board of *IEEE Communications Letters*. His research activity is focused on physical-layer OFDM(A), multi-antenna, relaying, and full-duplex wireless techniques.

LAURI ANTTILA [S'06, M'11] received his M.Sc. degree and D.Sc. (Tech) degree (with honors) from TUT in 2004 and 2011. Currently, he is a postdoctoral research fellow in the Department of Electronics and Communications Engineering at TUT. His research interests include signal processing for wireless communications, transmitter and receiver linearization techniques, and full-duplex radio systems.

CLEMENS ICHELN received his M.Sc. degree in electrical engineering (Dipl.-Ing.) at Hamburg-Harburg University of Technology, Germany, in 1996, and his Licentiate degree in radio engineering and Doctor of Science in Technology degree from Helsinki University of Technology, Espoo, Finland, in 1999 and 2001, respectively. He is currently working as a university lecturer in the Department of Radio Science and Engineering of Aalto University School of Electrical Engineering. His main research interests are the design of antennas for mobile terminals and compact relays, and their evaluation methods.

KATSUYUKI HANEDA [S'03, M'07] received his D.Eng. degree from the Tokyo Institute of Technology, Japan, in 2007. He is presently an assistant professor in the Aalto University School of Electrical Engineering. His current research activity focuses on high-frequency radios such as millimeter-wave and beyond, wireless for medical and post-disaster scenarios, radio wave propagation prediction, and in-band full-duplex radio technologies.

RISTO WICHMAN received his Ph.D. from TUT in 1995. From 1995 to 2001, he worked at Nokia Research Center, Finland, as a senior research engineer. In 2002, he joined the Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering, where he has been a full professor since 2008. His research interests include signal processing techniques for wireless communication systems.

MIKKO VALKAMA [S'00, M'02] received his M.Sc. and Ph.D. degrees (both with honors) from TUT in 2000 and 2001, respectively. In 2003, he worked as a visiting researcher with the Communications Systems and Signal Processing Institute at San Diego State University, California. Currently, he is a full professor and vice-head of the Department of Electronics and Communications Engineering at TUT. His general research interests include communications signal processing, cognitive radio, full-duplex radio, radio localization, and 5G mobile cellular radio systems.

By combining the advanced antenna design with active RF and proposed digital canceller, we showed with actual RF measurements that the self-interference can be suppressed below the receiver noise floor, even when using regular low-cost components.

Electrical Balance Duplexing for Small Form Factor Realization of In-Band Full Duplex

Leo Laughlin, Mark A. Beach, Kevin A. Morris, and John L. Haine

ABSTRACT

Transceiver architectures utilizing various self-interference suppression techniques have enabled simultaneous transmission and reception at the same frequency. This full-duplex wireless offers the potential for a doubling of spectral efficiency; however, the requirement for high transmit-to-receive isolation presents formidable challenges for the designers of full duplex transceivers. Electrical balance in hybrid junctions has been shown to provide high transmit-to-receive isolation over significant bandwidths. Electrical balance duplexers require just one antenna, and can be implemented on-chip, making this an attractive technology for small form factor devices. However, the transmit-to-receive isolation is sensitive to antenna impedance variation in both the frequency domain and time domain, limiting the isolation bandwidth and requiring dynamic adaptation. Various contributions concerning the implementation and performance of electrical balance duplexers are reviewed and compared, and novel measurements and simulations are presented. Results demonstrate the degradation in duplexer isolation due to imperfect system adaptation in user interaction scenarios, and requirements for the duplexer adaptation system are discussed.

INTRODUCTION

Demand for wireless communication services continues to grow at an astonishing rate. Although in many countries additional spectrum is becoming available (e.g., through the digital television switchover and the forthcoming release of military spectrum), and research into millimeter-wave technology promises to exploit underutilized spectrum at much higher frequencies, this growth in available spectrum pales in comparison to the increasing demands of our connected future. Spectral efficiency will continue to be a key research driver for many years to come.

Radio systems typically require transmit signal powers that are many orders of magnitude higher than the receive signal powers (often by over 100 dB). Due to this basic property, it has long been held that a radio system cannot transmit and receive on the same frequency at the same time, as the higher powered transmit signal would cause *self-interference* at the receiver, thereby obscuring the receive signal and preventing its reception. This limitation arises from the simple fact that wireless signals attenuate quickly with distance, and until now this has prevented the realization of in-band full-duplex (IBFD) radio. In contrast, signals propagating through wires suffer comparatively little attenuation, and IBFD has been commonplace in wired telephony for over 100 years [1].

Today's radio systems achieve duplex operation by simply circumventing this problem. Uplink and downlink channels are separated in time, using time-division duplexing (TDD), or in frequency, using frequency-division duplexing (FDD), thereby avoiding co-channel self-interference. However, since the 1980s [2, 3], and particularly in recent years [4], research has successfully challenged the traditional duplexing paradigm, proposing new system architectures utilizing various techniques to provide high levels of transmit-to-receive isolation, thus allowing simultaneous transmission and reception at the same frequency. This has obvious benefits for spectral efficiency, theoretically providing double the capacity over TDD or FDD systems without increasing bandwidth.

To achieve in-band full duplex operation, the transceiver system must somehow suppress the self-interference to an acceptable level. If the self-interference can be reduced to 6 dB or more below the receiver noise floor, the degradation in signal-to-noise ratio (SNR) due to the self-interference will be less than 1 dB, and the bandwidth efficiency of the two-way radio link is almost doubled. In Wi-Fi systems, this means that around 115 dB of self-interference suppression is required. For Long Term Evolution (LTE), which operates over longer ranges (and hence has a larger gap between transmit power and sensitivity), the requirement is higher, at around 123 dB. As engineers we are, of course, quite familiar with using decibels; however, it is worth reflecting on the magnitude of this number: 123 dB of interference suppression means

Leo Laughlin, Mark A. Beach, and Kevin A. Morris are with the University of Bristol.

John L. Haine is with u-blox A.G.

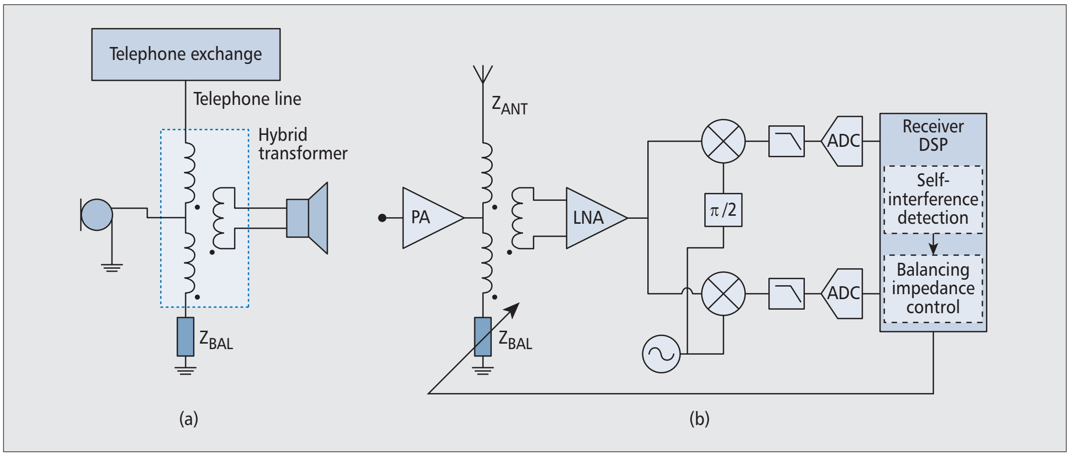


Figure 1. a) An electrical balance duplexer used to isolate the microphone and earpiece in a pre-electronic telephone; b) an electrical balance duplexer in a wireless transceiver including adaptive balancing impedance control.

Electrical balance duplexing is by no means new; it has been used since the early days of wired telephony. In a telephone system, the microphone and earpiece must both be connected to the telephone line, but must be isolated from one another to prevent the users' own speech deafening them to the much weaker incoming audio signal.

that the interference must be reduced by a factor of over 1 trillion. It is little wonder that until recently, in-band full-duplex has been assumed to be impossible.

Self-Interference Suppression

To achieve these high levels of transmit-to-receive (Tx-Rx) isolation, various methods must be employed to suppress the self-interference. Analog and digital cancellation techniques are key enabling technologies for IBFD systems [4], while antenna-based isolation techniques provide passive isolation by exploiting signal directionality at the antenna port [5, 6] or using separate antennas for transmitting and receiving [3, 7]. Antenna separation architectures utilize separate transmit and receive antennas, and can exploit directional antennas and shielding to improve isolation [7]. While this can be extremely effective, the isolation is fundamentally limited by the physical separation of the antennas. For the antenna separations possible in base station applications, high passive isolation of up to 71 dB has been reported [7]; however, for smaller antenna separations, the achievable isolation is significantly lower at around 30 dB [3, 8]. This isolation could also be compromised by interaction with the user's hand [8], and the now common requirement for multiple-input multiple-output (MIMO) would further increase the number of antennas, and hence the size of the device. Consequently, antenna separation architectures may not be an attractive option in smartphone and tablet devices, where size and form factor often take precedence over other design considerations. Similarly, single-antenna architectures utilizing circulators [5] are also unattractive for mobile device applications due to their size, cost, and limited isolation (approximately 15 dB).

In this article we present an *alternative* antenna-based isolation architecture, based on *electrical balance* in hybrid junctions, which can provide high transmit-to-receive isolation, but uses just a single antenna and can be implemented on-chip, making it suitable for use in consumer mobile devices.

Electrical Balance in Hybrid Junctions

Electrical balance duplexing is by no means new; it has been used since the early days of wired telephony [1]. In a telephone system, the microphone and earpiece must both be connected to the telephone line, but must be isolated from one another to prevent the users' own speech deafening them to the much weaker incoming audio signal. This was achieved using a hybrid junction connected as shown in Fig. 1a.

The hybrid junction (also known as a "hybrid coil," "hybrid transformer," "bridge transformer," "magic tee," or simply a "hybrid") is a four-port lossless reciprocal network in which opposite pairs of ports are isolated from one another [9]. Its operation is depicted in Fig. 2b: assuming all ports are terminated with the correctly matched characteristic impedance, a signal arriving at any one of the hybrid's ports is divided between the two adjacent ports, but not coupled to the opposite port; the opposing ports are *conjugate*. This property of "electrical balance" can be exploited to isolate transmitter and receiver circuitry, while allowing use of a shared propagation medium. The wireless duplexing application [6, 10, 11] is entirely analogous to the telephone duplexer, but substitutes the microphone, earpiece, and telephone line with the transmitter, receiver, and antenna, respectively, as shown in Fig. 1b.

In a wireless electrical balance duplexer the transmit current, supplied by the power amplifier (PA), enters the hybrid at the center tap of the primary winding (as shown in Fig. 1b) and is split between two paths flowing in opposite directions, with one component flowing to the antenna (and being transmitted), and the other to the *balancing impedance*. The relative magnitudes of these currents is determined by the transformer tapping ratio, *r*, and by the values of the antenna and balancing impedances. The balancing impedance is adjusted such that these two currents create equal but opposite magnetic fluxes that cancel, and therefore zero current is induced in the secondary winding of the transformer — the receiver is isolated from the transmitter. A signal received at the antenna, Instead of maximizing power transfer between the antenna and the Low Noise Amplifier, it is better to design the duplexer to maximize the voltage gain experienced by the receive signal, and thereby minimize the receiver noise figure.

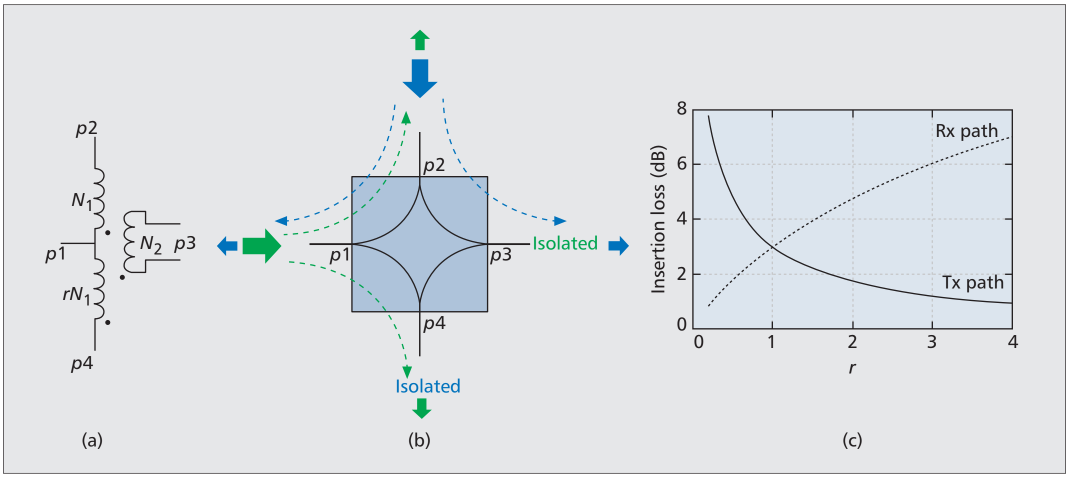


Figure 2. a) A hybrid transformer with tapping ratio, *r*; b) circuit symbol for a generic hybrid junction. The port numberings (p1-p4) correspond to Fig. 2a, and annotations show the signal coupling behavior of a hybrid junction: adjacent ports are coupled and opposite ports are isolated from one another; c) trade-off between Tx and Rx insertion loss, as controlled by the tapping ratio.

however, causes currents to flow through the primary winding in the same direction, thereby coupling it to the receiver winding. There is, of course, some inherent loss associated with the duplexer's operation. Some of the transmit power is dissipated by the balancing impedance, and not all of the power received at the antenna is coupled to the receiver. These losses depend on the tapping ratio, and can be traded off against one another by using a hybrid that is skewed in favor of the transmit or receive path, as shown in Fig. 2c. This relationship between the losses in the transmit and receive paths results from the reciprocity of the hybrid junction. For example, if the circuit were designed to deliver all of the transmit power to the antenna, since the circuit is reciprocal, all power arriving at the antenna would be delivered back to the transmitter (and therefore not to the receiver). Consequently, for the system to function in both transmit and receive modes, there is some unavoidable loss, which for a symmetrical hybrid (i.e when $r = 1$), is 3 dB in both the transmit and receive paths, reducing the transmitter efficiency and receiver sensitivity. Fortunately, however, these losses can be reduced by employing a *noise matched* receiver design.

Noise Matching

It is possible to mitigate the Rx insertion loss of the duplexer by noting that the absolute power of the receive signal is not the parameter that should be maximized; only the SNR of the receive signal is important. Therefore, instead of maximizing power transfer between the antenna and the low noise amplifier (LNA), it is better to design the duplexer to maximize the voltage gain experienced by the receive signal, and thereby minimize the receiver noise figure. This *noise matching* technique is possible because the receiver SNR depends not only on the receive power, but also on the impedance matching between the LNA and the various thermal noise sources in the duplexer circuit. To minimize the noise figure, the LNA input impedance is deliberately mismatched, thus reducing power transfer; however, the reduction in the noise power transferred to the LNA is greater than the reduction in the wanted signal power, and therefore the SNR is increased. In [11], an EB duplexer with noise matched LNA was fabricated and measured. This duplexer design skewed the hybrid to favor the transmission of power between the PA and the antenna, reducing the Tx insertion loss to 2.2 dB. The resulting Rx insertion loss of approximately 4 dB was circumvented through noise matching, thus reducing the LNA noise figure to less than 1 dB (as compared to 3 dB for an impedance matched LNA), and giving a total receiver noise figure of 5 dB. The Tx insertion loss and Rx noise figure of a noise matched EB duplexing transceiver is thus comparable to that of alternative architectures, such as the splitter, circulator, and coupler used for analog cancellation in [5]. A detailed example of noise matching in EB duplexers can be found in [11].

Differential and Common Mode Architectures

The self-interference arriving at the LNA inputs can be divided into two signal modes. Common mode (CM) signal components appear at the LNA inputs with the same phase at each terminal, whereas differential mode (DM) components appear in antiphase. In this system, the LNA is a differential amplifier; consequently, the CM signals are rejected and do not cause any self-interference further down the receive chain. It is therefore the DM isolation of the EB duplexer that determines the level of self-interference suppression. However, CM coupling must also be considered due to the high voltage swings it may present to the LNA input. At all but low transmit powers, these CM Tx signals may be large enough to destroy the LNA circuit, and hence, to enable the transceiver to operate at practical transmit powers, the EB duplexer must isolate the receiver from both DM and CM signals.

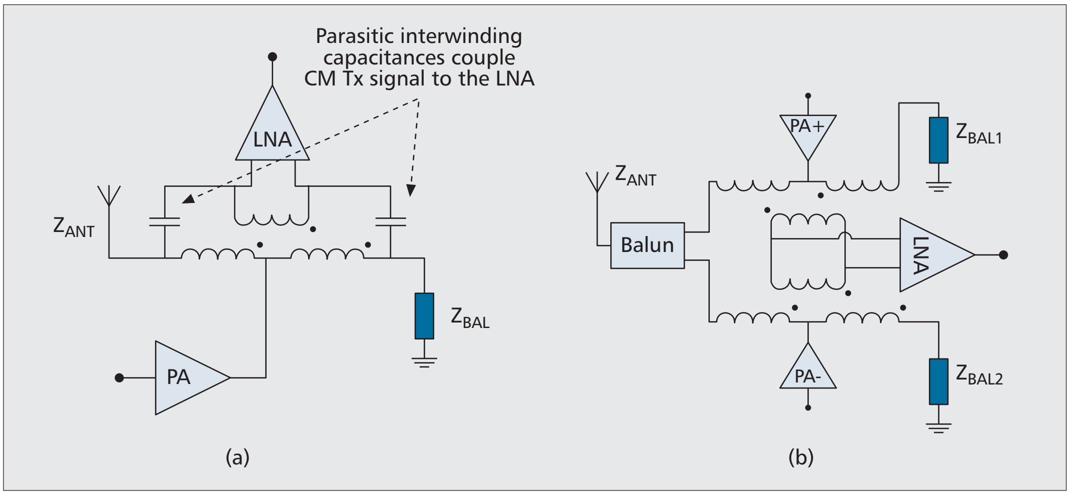


Figure 3. a) A single-ended EB duplexer with parasitic interwinding capacitances causing unwanted CM coupling of Tx signal; b) a fully differential EB duplexer providing high DM and CM isolation.

The self-interference arriving at the LNA inputs can be divided into two signal modes. Common mode signal components appear at the LNA inputs with the same phase at each terminal, whereas differential mode components appear in antiphase.

The duplexer depicted in Fig. 1b theoretically provides infinite CM isolation [11]; however, in a practical circuit, parasitic capacitive coupling between the primary and secondary windings of the transformer (Fig. 3a) results in significant CM coupling between the PA and LNA. The solution to this problem was proposed in [10], where a fully differential EB duplexer is introduced. In this circuit, a differential PA is used to drive two hybrid transformers. The receiver windings of these hybrids are then connected to the LNA in parallel, and since the input signals to the two hybrids are in antiphase, the CM coupling from each cancels out at the LNA. While providing the necessary levels of CM isolation, one drawback of this architecture is the requirement for a Balun, as the additional loss from this component adds to the Rx insertion loss and receiver noise figure.

Limitations of Electrical Balance Duplexing

The ideal hybrid junction provides infinite isolation between conjugate ports when all ports are correctly terminated with matched impedances. However, in practice, the antenna impedance is not an ideal 50 Ω resistor, but exhibits significant variation in the frequency domain, and has a significant reactive part. Furthermore, the antenna impedance may be perturbed by nearby objects, and can therefore also exhibit significant variation in the time domain [12]. Consequently, the balancing impedance must be adaptive in order to track temporal variation in the antenna impedance, requiring a system such as that depicted in Fig. 1b, and the transmit-to-receive isolation will be limited by impedance mismatch between the antenna and balancing impedance.

To obtain isolation across a given bandwidth, the balancing impedance must equal the antenna impedance at all frequencies within that band. However, since the antenna impedance typically varies significantly with frequency, simple balancing networks cannot mimic the antenna impedance effectively over wider bandwidths. For example, the balancing impedance circuit can be adaptively tuned to maximize the transmit-to-receive isolation at the carrier frequency. This will provide extremely high (theoretically infinite) isolation at the carrier frequency, but at other frequencies in the transmit band, the antenna impedance and balancing impedance will not be perfectly matched, and the isolation at these frequencies will therefore be reduced. This limitation in balancing is the dominant mechanism that limits the self-interference suppression achieved by EB duplexing, and since many antennas can exhibit relatively rapid variation in impedance with respect to frequency, this significantly reduces the *isolation bandwidth* of the EB duplexer. Furthermore, since the self-interference signal is spread across the transmit band, maximizing the isolation at the carrier frequency is not necessarily useful if this results in poor isolation in other parts of the transmit band. Instead, the goal is to maximize the receiver SNR, which is achieved by minimizing the mean Tx-Rx gain taken across the whole Tx band [6]. This optimum balancing occurs when the balancing impedance is chosen to minimize the mean square error (i.e., minimum mean square error, MMSE) between the balancing reflection coefficient and the antenna reflection coefficient: *MMSE balancing*. It can be shown [6] that for simple balancing networks, such as the parallel RC circuits used in [10, 11], that this *maximized isolation* is proportional to the variance of the antenna impedance with respect to frequency.

EB Duplexer Performance

Prototype EB duplexers reported in [10, 11] demonstrated high transmit-to-receive isolation over wide bandwidths where the antenna is modelled as a near ideal 50 Ω resistance. While these prototype duplexers have demonstrated that an EB duplexer can be successfully fabricated on-chip with reasonable Tx insertion loss and receiver noise figure, the omission of real antennas from these prototypes means that the observed isolation bandwidth may not be achievable in practice. Figure 4 compares these results with simulations reported in [6] which incorporate a measured antenna reflection coefficient (and therefore model the real antenna perfectly) and with novel measurements from a discreet hardware prototype EB duplexer including antenna presented here.

Hardware EB Duplexer

Figure 4d depicts the hardware EB duplexer in the laboratory. The balancing impedance is implemented using an electromechanical impedance tuner, which provides extremely accurate balancing impedance control, but has a slow tuning speed. The balancing impedance and

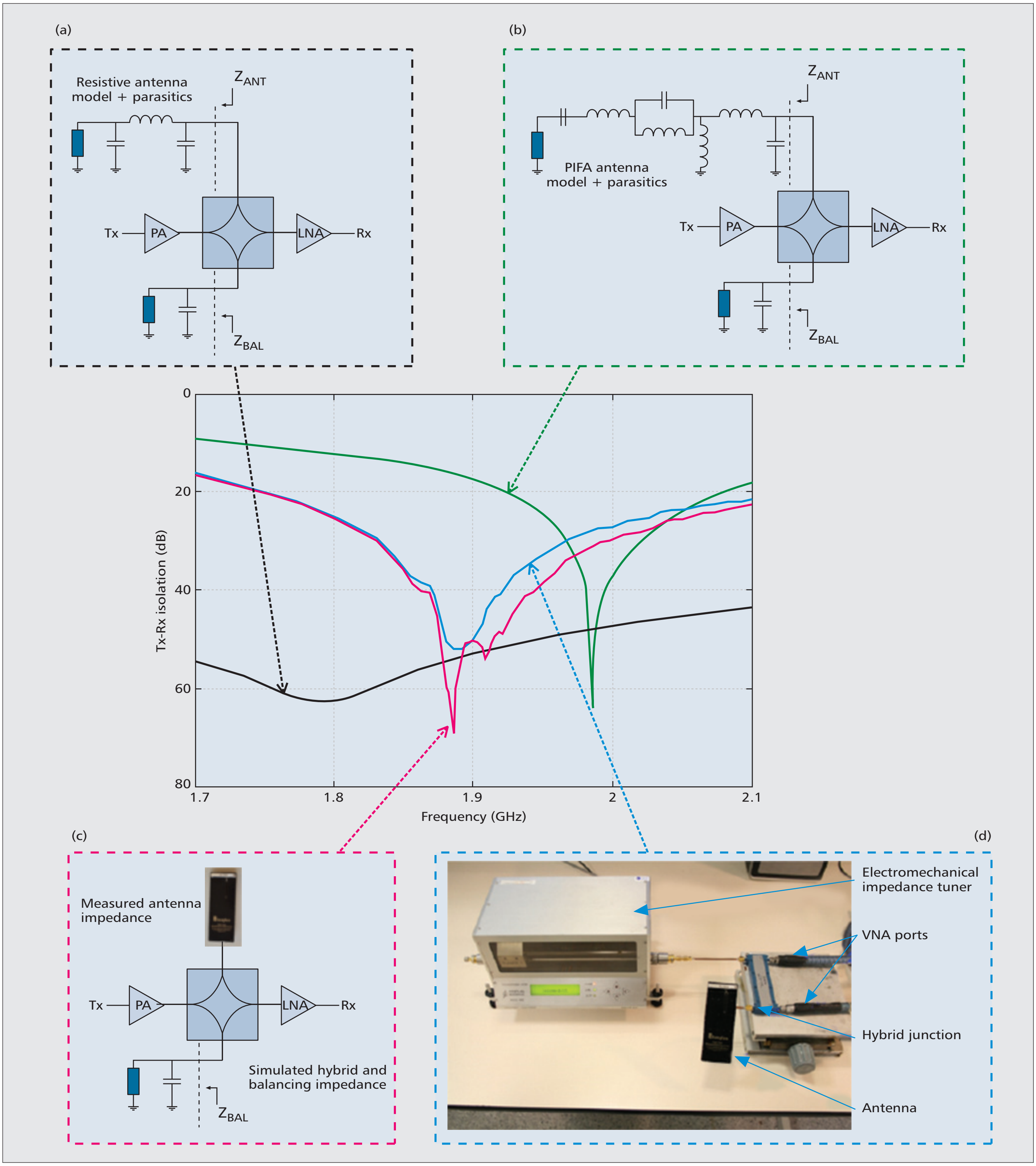


Figure 4. Comparison of reported results for duplexer isolation: a) simulation with resistive antenna model [10]; b) simulation with lumped element PIFA antenna model [13]; c) simulation with measured antenna data [6]; d) measurement of hardware prototype EB duplexer.

antenna are connected to a microstrip hybrid junction, and the lengths of the coaxial transmission lines for the antenna and balancing impedance were selected such that the group delay of the balancing and antenna reflection coefficients are approximately equal (thus minimizing the relative phase variation between the two reflection coefficients). The antenna used is the same antenna that was embedded into the circuit simulations in [6] (Fig. 4c): a Taoglas PAD710 multiband cellular antenna. The transmit-to-receive isolation is measured using a vector network analyzer (VNA), and these measurements also inform a simple gradient descent minimization algorithm running on a PC, which controls the balancing impedance in order to minimize the mean Tx-Rx gain across 20 MHz (using MMSE balancing [6]).

PERFORMANCE COMPARISON

Circuit simulations reported in [10] model the antenna as a 50 Ω resistor with some small parasitic capacitance and inductance, as depicted in Fig. 4a. The results show a *60 dB isolation bandwidth* (the bandwidth over which at least 60 dB of Tx-Rx isolation is achieved) of approximately 50 MHz. Similar isolation bandwidths were achieved by the hardware prototype, also reported in [10], which emulated the antenna on-chip using a similar circuit to the simulated antenna model. The transmit-to-receive isolation observed in this case has a higher peak isolation and a much larger isolation bandwidth compared to the simulations that include a measured antenna and the hardware duplexer presented here. Simulations in [6] report 60 dB isolation bandwidths of less than 10 MHz, and the hardware duplexer did not achieve 60 dB isolation. This clearly demonstrates the significant detrimental impact of the frequency variant antenna impedance. The hardware duplexer shows comparable performance to that predicted by the simulation, although there is some additional reduction in transmit-to-receive isolation due to the increased frequency domain variation in the balancing impedance. Simulations recently presented in [13] use a more realistic printed inverted F antenna (PIFA) model (Fig. 4b); these simulation results are in much closer agreement with measured performance. Furthermore, the isolation achieved by the prototype EB duplexer exceeds that of small form factor antenna separation architectures (~30 dB) and circulators (~15 dB).

DUPLEXER ADAPTATION

As shown, the matching between antenna impedance and the balancing impedance is the critical factor determining the transmit-to-receive isolation, and hence the suitability of this technology for use in full-duplex transceivers. Frequency domain variation in the antenna and balancing impedances has been shown to limit the transmit-to-receive isolation. However, the antenna impedance may also be time-variant due to interaction with objects in the environment. EB duplexing, being a single antenna duplexing technique, is well suited to small form factor device applications; however, it is well known that antennas in handheld devices can exhibit large variation in impedance due to interaction with the user [12]. To maintain Tx-Rx isolation, the balancing impedance must track the antenna impedance as it changes, requiring the adaptive architecture depicted in Fig. 1b. The performance of an EB duplexer is determined by the ability of the balancing impedance to mimic the antenna impedance in the time domain and the frequency domain, and therefore the design of the balancing network is critical to the resulting duplexer performance. Understanding the temporal properties of the antenna impedance, in terms of both the range of variation and the rate of change, is therefore vital to enable the design of an effective balancing network.

Frequency domain variation in the antenna and balancing impedances has been shown to limit the transmit-to-receive isolation. However, the antenna impedance may also be time-variant due to interaction with objects in the environment.

ADAPTATION RANGE

Antenna impedances can vary over a relatively wide range due to environmental effects. Furthermore, full duplex radio is one facet of a general trend toward fully reconfigurable radios, capable of operation at a wide range of frequencies. Antenna impedances may typically exhibit vastly different impedances even at relatively closely spaced operating frequencies, making the range of possible impedances even larger when multiband operation is required. Antenna tuning units (ATUs) [14] provide dynamically controlled impedance matching for the antenna, improving efficiency, and have become common in handheld devices. Including an ATU in the EB transceiver would also have the important benefit of reducing the range of impedances presented to the duplexer, reducing range requirements for the balancing impedance, as depicted in Fig. 5a.

ADAPTATION SPEED

In a practical system, any adaptive balancing impedance would be limited to some finite adaptation rate. As the antenna impedance fluctuates, the difference between the antenna impedance and balancing impedance will increase, thereby reducing the Tx-Rx isolation. The balancing impedance must track these changes at a rate that prevents this error from becoming unacceptably large. In this article, novel results demonstrating the impact of different tracking speeds for the balancing impedance are presented, and requirements for the balancing network adaptation performance are discussed.

As was the case in [6, 15], circuit simulations with measured antenna data were used to investigate the Tx-Rx isolation of an EB duplexer (Fig. 4c). The comparison between this simulation method and measured hardware duplexer presented in Fig. 4 demonstrates the validity of this technique. To observe the changes in antenna impedance typical for a handheld device with an interacting user, the antenna was placed in a plastic mobile phone housing, and multiple antenna reflection coefficient measurements were taken as the "user" held the "device," and emulated browsing and texting movements. Reflection coefficient measurements were taken at 2.5 ms intervals for a period of over 20 s, and these measurements were then incorporated into a circuit simulation that includes an adaptive

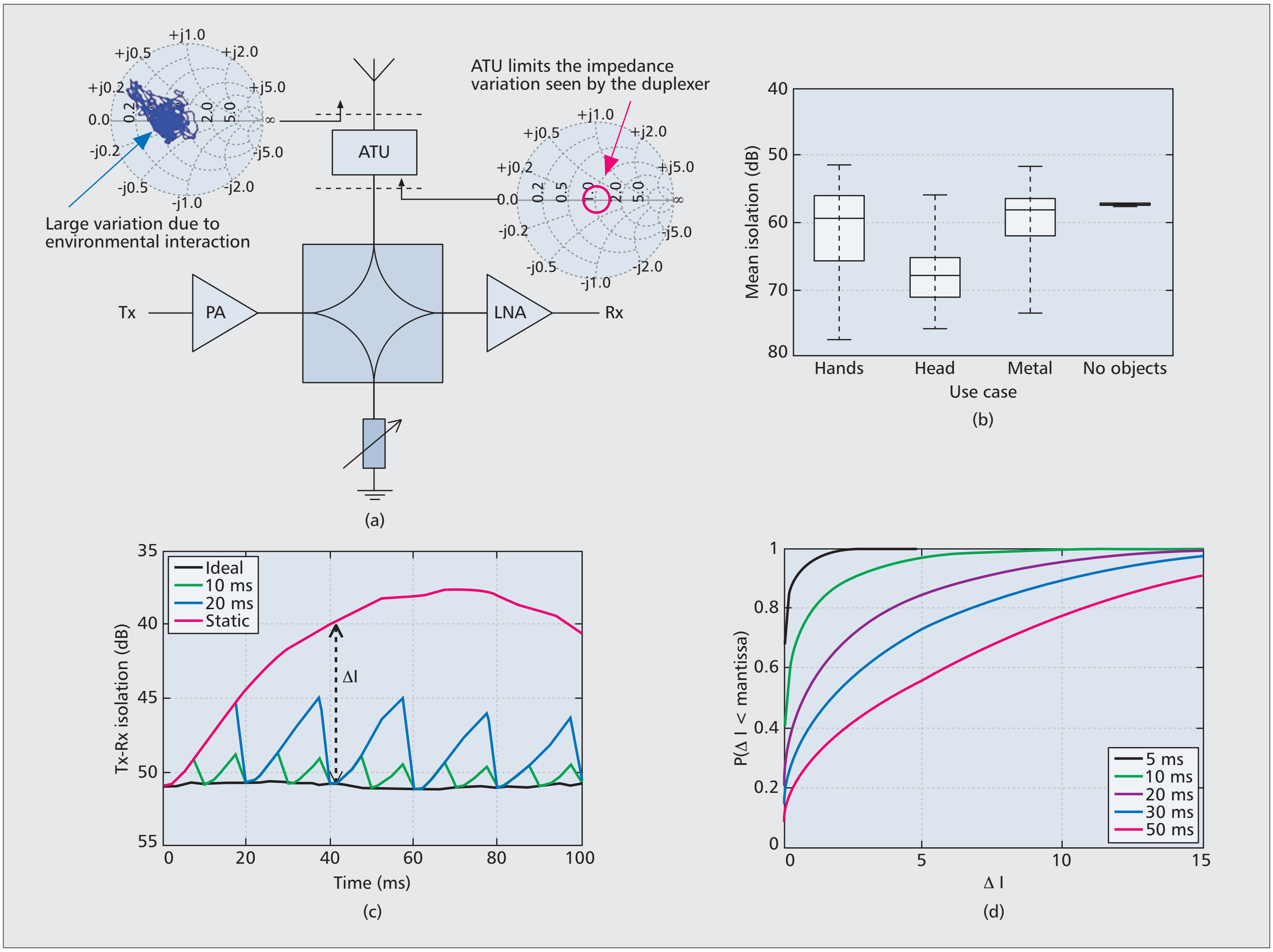


Figure 5. a) An antenna tuning unit (ATU) incorporated into an electrical balance duplexer; b) variation in maximized EB duplexer isolation due to user interaction at 1960 MHz [15]; c) simulated time domain variation in Tx-Rx isolation for four types of adaptive balancing impedance; and d) CDF curves for isolation degradation for five different balancing impedance update rates.

balancing impedance with a variable adaptation rate. The measurements were performed over a 20 MHz bandwidth centred at 1.9 GHz, and the duplexer performance metric used is the mean Tx-Rx isolation across this band. Full details of the circuit simulation are given in [15], where the simulation was used to investigate the variation in Tx-Rx isolation in browsing scenarios and voice call scenarios, and when the device is placed on a metal surface, under the assumption that the adaptive balancing is ideal (i.e., isolation is always maximized). Figure 5b shows results from [15], where the simulated mean Tx-Rx isolation values determined for each antenna measurement are grouped through their quartiles. These results show that provided that the balancing impedance is able to adapt, isolation of over 50 dB was maintained. The simulations in [15] assumed ideal balancing adaptation, such that the duplexer isolation is limited only by frequency domain variation in the antenna and balancing impedances, whereas the simulation results presented here show the additional degradation in Tx-Rx isolation caused by updating the balancing impedance at a finite rate. Figure 5c shows the time domain variation in Tx-Rx isolation for four different simulated balancing networks. The ideal network tracks the antenna impedance perfectly. The "static" balancing network is tuned to maximize the duplexer isolation at $t = 0$ and is then not updated. The degradation due to the untracked variation in the antenna impedance can be clearly seen. After 40 ms the isolation drops to below 40 dB, and was seen to go as low as 19 dB (not shown in the graph). The remaining two curves in Fig. 5c show the Tx-Rx isolation when the balancing impedance is updated at 10 ms and 20 ms. As would be expected, the higher update rate keeps the isolation closer to its maximized value. The additional degradation due to imperfect balancing impedance tracking, ΔI, is the difference in isolation from that of the ideal balancing curve, and this metric is of interest when developing requirements for the impedance tracking.

Adaptation Design Considerations

As is the case with many types of systems, the designer of an EB duplexer may seek to guarantee some level of isolation. To achieve this, the isolation budget of the system would need to include margins to account for any variation in the isolation (in much the same way that radio link budgets include fading margins). Consequently, understanding the variation in isolation is vital to the design of the system. Figure 5d depicts the cumulative distribution function (CDF) of the observed isolation degradation (ΔI) for a range of update rates, and provides useful information to the designer. For example, if only a 3 dB degradation margin can be accepted, for this system the update rate must be at least 5 ms. However, given the variable nature of the isolation, it may be more appropriate to specify the performance statistically. For example, if the criteria is less than 3 dB degradation for 90 percent of the time, a 10 ms update rate would be acceptable. It is also pertinent to note here that this rate of adaptation is relatively slow; adapting to user interaction is unlikely to be a problem for a practical EB duplexer.

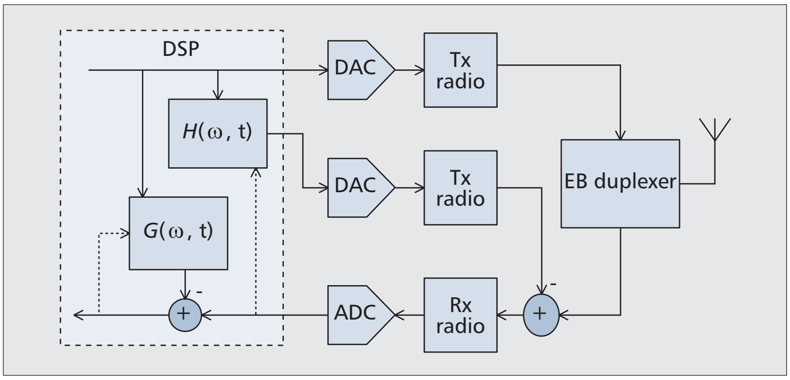


Figure 6. Possible architecture for a single-antenna full duplex radio system incorporating electrical balance, active analog cancellation, and digital baseband cancellation [6].

Research Opportunities

This newly rediscovered technology leaves various research opportunities remaining.

Combining EB with other self-interference suppression techniques. Typically, full duplex radio systems combine various methods to achieve high levels of isolation. For example, a full duplex architecture combining EB duplexing with active analog cancellation and digital baseband cancellation was proposed in [6] and is depicted in Fig. 6. Architectures such as this require much further investigation.

Antenna and balancing impedance co-design. Since the impedance matching between these two components is critical to the isolation, there is potential benefit in designing the two such that their impedances vary in a similar fashion, and this could significantly widen the isolation bandwidth. Conversely, it may be that the antenna type is not known to the designer of the duplexer (typical for RF integrated circuits [ICs], where a single IC design finds its way into a wide range of end products). In this case the duplexer should be designed to cope with an arbitrary antenna impedance.

EB duplexer adaptation in other environments. The novel results presented in this article investigate the EB duplexer adaptation requirements in user interaction scenarios. Other dynamic environments will present different requirements (e.g., vehicular communication).

Balancing impedance implementation. Deep submicron complementary metal oxide semiconductor (CMOS) implementation is possible [10]; however, microelectromechanical (MEMS) tunable components may offer advantages in terms of linearity, power handling, and resolution. Recent publications [13] have also introduced balancing impedance designs which can match the antenna impedance at two frequency points, which is of clear benefit for FDD applications, but should also be investigated to determine the possible improvement in IBFD isolation bandwidth.

Conclusion

Electrical balance duplexing offers high Tx-Rx isolation over wide bandwidths, but requires just one antenna and can be implemented on-chip, making this architecture well suited to small form factor devices. The transmit-to-receive isolation achievable using this technique is determined by how effectively the balancing impedance can mimic the antenna impedance, in both the time and frequency domains. Consequently, the isolation performance is limited by frequency domain variation in the antenna impedance, and the ability of the adaptive balancing impedance to track changes in the antenna impedance. Various contributions reporting simulated and measured duplexer isolation performance have been compared, and it has been shown that frequency domain variation in the antenna impedance is the primary factor determining transmit-to-receive isolation in practical duplexers. Circuit simulations incorporating measured antenna data have been used to investigate the effect of imperfect duplexer adaptation on the Tx-Rx isolation in user interaction scenarios. It was shown that a balancing impedance update rate of 10 ms was adequate to ensure isolation degradation of less than 3 dB for 90 percent of the time. EB duplexing offers significant further research opportunities and full duplex architectures exploiting electrical balance require further investigation.

Acknowledgments

The authors would like to thank Mici McCullagh of u-blox for his technical insight, and Taoglas for generously providing the antenna featured herein. This research is funded by u-blox and the U.K. EPSRC.

References

[1] G. A. Campbell and R. M. Foster, "Maximum Output Networks for Telephone Substation and Repeater Circuits," *Am. Inst. Electron. Eng. Trans.*, vol. XXXIX, no. 1, 1920, pp. 231–90.

[2] C. K. Richardson, "Improvements in or Relating to Transmitters/Receivers," U.K. Patent, GB15775141980.

[3] S. Chen, M. A. Beach, and J. P. McGeehan, "Division-Free Duplex for Wireless Applications," *Electron. Lett.*, vol. 34, no. 2, 1998, pp. 147–48.

[4] A. Sabharwal *et al.*, "In-Band Full-Duplex Wireless: Challenges and Opportunities," *IEEE JSAC*, vol. 32, no. 9, 2014, pp. 1637–52.

[5] D. Bharadia, E. McMilin, and S. Katti, "Full Duplex Radios," *Proc. 2013 ACM SIGCOMM*, 2013.

[6] L. Laughlin *et al.*, "Optimum Single Antenna Full Duplex Using Hybrid Junctions," *IEEE JSAC*, vol. 32, no. 9, Sept. 2014, pp. 1653–61.

[7] E. Everett, A. Sahai, and A. Sabharwal, "Passive Self-Interference Suppression for Full-Duplex Infrastructure Nodes," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, 2014, pp. 680–94.
[8] L. Zhao, L. K. Yeung, and K.-L. Wu, "A Coupled Resonator Decoupling Network for Two-Element Compact Antenna Arrays in Mobile Terminals," *IEEE Trans. Antennas Propag.*, vol. 62, no. 5, May 2014, pp. 2767–76.
[9] H. J. Carlin and A. B. Giordano, *Network Theory: An Introduction to Reciprocal and Non-Reciprocal Circuits*, Prentice-Hall, 1964.
[10] S. H. Abdelhalem, P. S. Gudem, and L. E. Larson, "Hybrid Transformer-Based Tunable Differential Duplexer in a 90-nm CMOS Process," *IEEE Trans. Microwave Theory Tech.*, vol. 61, no. 3, 2013, pp. 1316–26.
[11] M. Mikhemar, H. Darabi, and A. A. Abidi, "A Multiband RF Antenna Duplexer on CMOS: Design and Performance," *IEEE J. Solid-State Circuits,* vol. 48, no. 9, 2013, pp. 2067–77.
[12] K. R. Boyle, Y. Yuan, and L. P. Ligthart, "Analysis of Mobile Phone Antenna Impedance Variations with User Proximity," *IEEE Trans. Antennas Propag.*, vol. 55, no. 2, Feb. 2007, pp. 364–72.
[13] S. H. Abdelhalem, P. S. Gudem, and L. E. Larson, "Tunable CMOS Integrated Duplexer with Antenna Impedance Tracking and High Isolation in the Transmit and Receive Bands," *IEEE Trans. Microwave Theory Tech.*, vol. 62, no. 9, Sept. 2014, pp. 2092–2104.
[14] J. deMingo *et al.*, "An RF Electronically Controlled Impedance Tuning Network Design and Its Application to an Antenna Input Impedance Automatic Matching System," *IEEE Trans. Microwave Theory Tech.*, vol. 52, no. 2, Feb. 2004, pp. 489–97.
[15] L. Laughlin *et al.*, "Performance Variation in Electrical Balance Duplexers Due to User Interaction," *Proc. PIMRC*, 2014.

BIOGRAPHIES

LEO LAUGHLIN received his M.Eng. degree in electronic engineering from the University of York, United Kingdom, in 2011. He is currently working toward his Ph.D. degree at the University of Bristol, United Kingdom. During 2009–2010, he was with Qualcomm, United Kingdom, working on GSM receivers. In 2011, he was with Omnisense Ltd., Cambridge, United Kingdom, working on radio geolocation systems. His research interests are in the areas of in-band full-duplex radio and multiband RF transceiver design.

MARK A. BEACH is a professor of radio systems engineering with the University of Bristol. He is recognized for his research in multi-antenna and RF technologies, and has held the post of head of the Department of Electrical and Electronic Engineering at Bristol. He is now the U.K. representative on EU COST IC1004, a member of the EPSRC Strategic Advisory Team for ICT, and manager of the CDT in Communications at Bristol.

KEVIN A. MORRIS received his B.Eng. and Ph.D. degrees in electronics and communications engineering from the University of Bristol in 1995 and 1999, respectively. He is currently head of the Department of Electrical and Electronic Engineering, University of Bristol. He has authored or co-authored 60 academic papers and is the joint author of five patents. His research interests are in the design of efficient linear broadband power amplifiers for use within future communications systems.

JOHN L. HAINE received his B.Sc. and Ph.D. degrees from the University of Birmingham, United Kingdom, in 1971 and the University of Leeds, United Kingdom, in 1977. He has worked on wireless R&D for a number of companies, including startups. He is currently responsible for RF technology strategy in u-blox AG, working on IoT communications and cellular RF implementation technologies. He is a member of the IET and serves on the Board of Cambridge Wireless.

Medium Access Control Design for Full Duplex Wireless Systems: Challenges and Approaches

Karaputugala Madushan Thilina, Hina Tabassum, Ekram Hossain, and Dong In Kim

ABSTRACT

Recent advances in self-interference cancellation techniques enable in-band full-duplex (FD) transmission in which a wireless node can simultaneously transmit and receive in the same frequency band. However, to fully exploit the benefits of FD technology in a wireless network, in addition to the physical (PHY) layer issues, medium access control (MAC) layer issues such as inter-node collisions, fairness between half-duplex (HD) and FD users, opportunistic selection of different modes of FD transmission, and synchronization issues need to be resolved. To this end, this article first discusses the fundamental concepts, potential benefits, and primary network topologies of FD transmission. We then highlight immediate challenges (both in the PHY and MAC layers) that need to be addressed in designing FD wireless systems. A qualitative comparison among the existing full-duplex MAC (FD-MAC) protocols is then provided. Finally, the primary requirements and research issues for the design of FD-MAC protocols are discussed, and implications of FD technology in cellular wireless networks are highlighted.

Karaputugala Madushan Thilina, Hina Tabassum, Ekram Hossain, and Dong In Kim.

This work was supported in part by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) and in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (2014R1A5A1011478).

INTRODUCTION

Until very recently, the concept of transmission and reception in the same time and frequency domain (referred to as full-duplex (FD) technology) did not seem to be very promising. The primary reason was the overwhelming nature of the so called self-interference (SI), which is generated by the transmitter to its own collocated receiver. SI is a fundamental bottleneck in the progress of FD technology [1]. Fortunately, with the recent advancements in antenna and digital baseband technologies as well as RF interference cancellation techniques, SI can be reduced close to the level of the noise floor in low-power networks, e.g. cognitive radio networks and Wi-Fi networks [2].

At the physical layer (PHY), considering a point-to-point link and perfect SI cancellation, FD transmission offers twice the spectral efficiency of half-duplex (HD) transmission. Due to this attractive feature, FD technology is rapidly extending its applications in different wireless communications scenarios, especially those with low transmission power and distance requirements [3]. For instance, small cell networks, device-to-device (D2D) communications, cognitive radio networks, and multi-hop relaying are potential areas where FD technology can be practically feasible and implementable in the near future. However, to fully exploit the benefits of FD technology, major PHY and medium access control (MAC) layer issues need to be resolved by devising new PHY layer techniques and by modifying the existing MAC layer protocols. These issues include mitigation of residual self-interference, inter-node interference, intercell uplink to downlink/downlink to uplink interference, fairness between HD and FD users, opportunistic selection of different modes of FD transmission, synchronization and time adjustment issues to establish FD transmission, etc. Understanding the role of the PHY and MAC layers is crucial to address these issues.

This article first discusses the fundamental concepts, potential benefits, primary network topologies, and collision domains in FD transmission. Then it highlights the immediate challenges (both in the PHY and MAC layers) that need to be addressed in the design of FD-MAC protocols. A qualitative overview of the existing FD-MAC protocols is then provided. To this end, major issues and approaches for designing FD-MAC are discussed. Finally, implications of FD technology resource and interference management of cellular networks are highlighted.

FUNDAMENTALS OF FULL-DUPLEX WIRELESS SYSTEMS

ANTENNA CONFIGURATIONS FOR FD SYSTEMS

Since FD transmission requires in-band operation of transmitting and receiving RF chains, the conventional duplexers cannot be directly utilized to maintain separation between the two RF transmissions. FD transmission can, however, be realized through the following antenna configurations:

Shared Antenna Configuration [4]: In this configuration, a single antenna can be used for simultaneous in-band transmission and reception through a three-port circulator. Ideally, the circulator prevents the leakage of signals from the transmit RF chain to the receive RF chain. However, in practice, the transmit signal causes SI to the signals received. Moreover, due to hardware limitations and severe interference, multiple circulators cannot be utilized to enable the use of multiple shared antennas.

Separated Antenna Configuration [5, 6]: In this configuration, the total number of antennas is divided into two groups for transmission and reception. This division of spatial resources, however, introduces a trade-off. As such, a fair comparison between HD and FD transmission should consider the exact number of RF chains/antennas required to establish FD transmission. Also, in the separated antenna configuration, it is crucial to first analyze whether the performance gains of a HD multi-antenna transmission are worth sacrificing with the bidirectional (SI affected) FD transmission?

Transmission Modes in Full-Duplex Systems

The fundamental modes of transmission in FD wireless systems are listed below [7].

Half-Duplex: In this mode RF transmission takes place in a single direction between primary transmitter (PT) (i.e. the node that initiates transmission) and primary receiver (PR) (i.e. the node that decodes the signal from PT) (Fig. 1a). Both nodes operate in HD mode if neither PT receives a signal from a secondary transmitter (ST) (i.e. the node that is allowed to transmit simultaneously with PT), or the PR has to transmit to PT or any other node.

Bi-Directional Full-Duplex (BFD): In this mode both PT and ST transmit signals to each other at the same time (Fig. 1b). Note that PT and PR becomes secondary receiver (SR) (i.e. the node that decodes signal from ST) and ST, respectively.

Three Node Full-Duplex (TNFD): In this mode a FD node transmits to one node while receiving from another node (Fig. 1c and Fig. 1d). The transmission modes illustrated in Fig. 1c and Fig. 1d are known as destination-based transmission mode (DBTM) and source-based transmission mode (SBTM), respectively.

- DBTM: In DBTM mode, the destination for PR is another node (i.e. not the PT node) which is located in the vicinity of PR's transmission range. Thus, PR becomes the ST and its receiver becomes SR.
- SBTM: The SBTM mode is enabled when PR does not have data to transmit. However, the PT's neighbor node wants to transmit data to PT. In this case, PT's neighbor node becomes ST and PT becomes SR by activating the FD mode.

As an example, in cellular networks where a base station (BS) may not have data packets for a node transmitting to the BS in the uplink or a node receiving transmission from the BS in the downlink may not have data packets for the BS. In such scenarios, the BFD mode cannot be established. However, in the former case, a BS can exploit the FD opportunity by initiating a downlink transmission to another node (DBTM). In the latter case, a BS may initiate receiving data packets from another node in the uplink (SBTM). Note that both SBTM and DBTM may also occur at the same time, and the mode can thus be referred to as *concurrent SBTM and DBTM*.

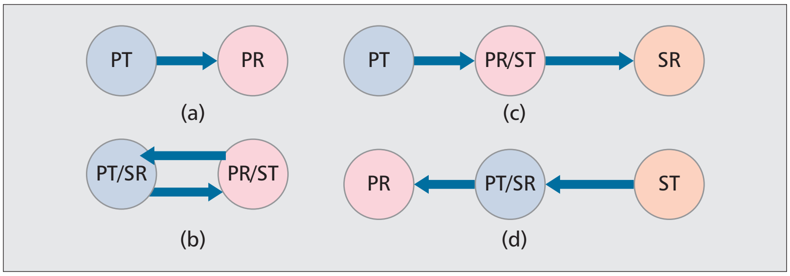


Figure 1. FD transmission modes: a) half-duplex transmission; b) bi-directional FD transmission; c) destination-based FD transmission; d) source-based FD transmission.

Inter-node interference is a crucial factor in TNFD-enabled FD networks depending on the locations and transmit powers of the nodes. Various inter-node interference scenarios are depicted in Fig. 2. The bold circle around the nodes presents the PT's transmission region, whereas the dotted circle denotes the ST's transmission region. Figures 2a and 2b show intra-node collision within a set of nodes operating in DBTM and SBTM, respectively. In DBTM the SR suffers collision from PT if located within the PT's transmission region. In SBTM, the PR suffers collision from ST if located within the ST's transmission region. Similarly, Figs. 2c, 2d, and 2e show the inter-node collisions due to the STs' transmission region overlapping, STs' and PTs' transmission region overlapping, and PTs' transmission region overlapping, respectively.

Fundamental Benefits of FD Technology

In addition to doubling the spectral efficiency, FD technology exhibits several other advantages over conventional HD transmission.

Opportunistic Selection of FD Mode: Radio resources (e.g. resource block (RB) in LTE-A networks) can be opportunistically used for either HD, BFD, or TNFD transmission mode. This mode selection per resource block can be optimized to maximize the overall system utility, considering the prior knowledge of the available data packets and channel/interference conditions of the prospective participating nodes.

Rapid Collision Detection: An FD-enabled wireless device can listen to the RF channel while transmitting to probe the occurrence of other transmissions on the same channel. This would enable fast collision detection, e.g. spectrum sensing in cognitive radio networks.

BFD Mitigates Hidden Terminal[1] Problem [6]: The transmitted signals from both ST and PT can be overheard by their respective neighboring nodes. As a result, the neighboring nodes refrain from initiating transmission until the ongoing BFD transmission is completed. The collisions due to the hidden node problem can thus be eliminated.

[1] Hidden terminals are the wireless nodes that can potentially transmit (and thus introduce interference) during an ongoing transmission. Note that, the hidden nodes cannot be detected by the carrier sensing mechanism.

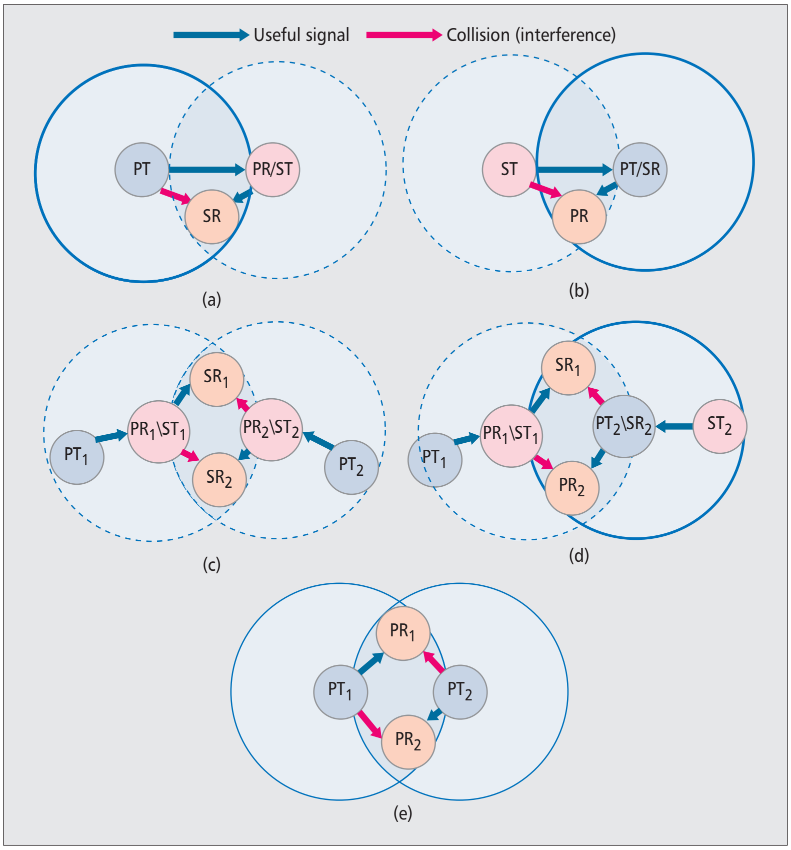


Figure 2. Fundamental inter-node collision possibilities in FD transmission.

Channel Selection: An FD-enabled node can assist nearby nodes in the channel selection process. For instance, if a PR has nothing specific to transmit to PT or any other node, the PR may assist its nearby nodes by informing them about the interference on a channel. This may help the nearby nodes in selecting a channel.

Reduced Latency: FD technology exhibits reduced latency by enabling the reception of feedback signals from the receiver (i.e. channel state information (CSI), control signaling for ARQ and/or network management, ACK, etc.) during transmission.

Secure Transmission: Due to the simultaneous transmission and reception in FD networks, the eavesdroppers receive mixed signals that are difficult to decode.

RF Energy Transfer During Transmission [8]: FD technology enables a wireless node to perform wireless charging while performing an uplink transmission.

Challenges in the Design of FD Wireless Systems

Fundamental Challenges

Residual Self-Interference (SI): SI is caused by the coupling of the transceiver's own transmit signal to the receiver while attempting to receive a signal sent by another wireless node. The key challenge arises from the large power level difference between the transceiver's own transmission and the signal of interest coming from a distant source. To achieve considerable SI cancellation, several antenna, RF, and baseband cancellation techniques [2, 9] are currently under investigation. Even with the several stages of SI cancellation (passive, active analog, and active digital), a certain amount of SI remains in the system, which is referred as residual SI. To achieve the theoretical performance limit of FD transmission, efficient SI cancellation techniques are required that can eliminate residual SI.

Intracell Interference: In a TNFD-based SBTM transmission, a PT communicates with two nodes using HD transmission, i.e. the PR in the downlink and the ST in the uplink. There exists two types of interference in this case: residual SI from the PT to the SR and intra-node interference from the ST to the PR (Fig. 2b). Since the PT and SR are collocated, the SI signal is known and can possibly be canceled at the SR. However, the signal that is the source of the intra-node interference is not known at the PR and thus cannot be canceled on reception. This intra-node interference can significantly degrade the performance of PT-PR transmission depending on the vicinity of the ST with the PR.

Uplink to Downlink and Downlink to Uplink Inter-Cell Interference: In multiuser cellular networks, additional inter-cell interference will be experienced by a node (whether operating in HD, BFD, or TNFD mode) depending on its direction of transmission. For instance, if a node is transmitting in the uplink, it will receive interference from the uplink transmitters of the neighboring cells (conventional) as well as the neighboring BSs who are transmitting in the downlink (additional). This additional interference is therefore referred to as downlink to uplink inter-cell interference. On the other hand, if a node is transmitting in the downlink, it will receive interference from the neighboring BSs (conventional) as well as the uplink transmitters of the neighboring cells, which is referred to as uplink to downlink inter-cell interference. By analyzing these two specific scenarios, it can be concluded that the users who want to operate in HD mode may suffer considerably as the potential gains of FD technology are not applicable to them. Moreover, due to the high transmit power of BSs, downlink to uplink interference becomes more severe than uplink to downlink interference.

Asynchronous TNFD Transmission: While BFD mode may possibly enable a perfectly synchronized transmission between two nodes, it may not be straightforward in TNFD transmission mode, because the emergence of the third node may not occur at the same time when a single-hop HD transmission is initiated. Thus, the primary question arises about the feasibility and performance limit of the asynchronous TNFD systems, i.e. can a new node be added once a HD transmission starts? If yes, then how much gain can be achieved practically?

There can be two possible scenarios, i.e. a PT starts receiving a packet while initiating a transmission, or a PR starts a new transmission while receiving a packet. In the former case, the PT can suppress its own SI while transmitting to the

PR in order to decode new transmission from the ST. Also, correct decoding of new reception requires the PT to estimate the channel between the ST and itself in the presence of the SI. On the other hand, initiating a new transmission at the PR while receiving from the PT is not reliable enough as the process of estimating the channels to establish a canceling signal causes a self-collision at the receiver [6].

Challenges at the MAC Layer

FD networks allow several transmission modes that can result in a new kind of inter-node collisions other than the SI as mentioned previously. As such, minimizing the additional inter-node collisions with traditional MAC protocols such as the carrier-sense multiple access with collision avoidance (CSMA/CA)-based HD-MAC protocols is not straightforward. This section focuses on characterizing the main challenges that need to be considered while designing MAC protocols for FD transmission.

Selecting FD Transmission Modes and Nodes: In FD networks, nodes can operate in any of the transmission modes that are illustrated in Fig. 1. Hence, selecting a set of nodes and an FD transmission mode to maximize the overall utility of FD transmission is crucial. The basic approaches that are currently used in FD-MAC protocols to exploit proper nodes and modes for FD transmission are shared random backoff, header snooping, and request-to-send (RTS)/clear-to-send (CTS) mechanisms. These will be discussed in the next section.

Fairness: Due to the bidirectional communication capabilities in FD networks, the fairness among nodes may degrade by a factor of two compared to HD communication. For instance, assume that there are three sets of potential BFD nodes located parallel to each other. The BFD nodes in the corners can use the same channel without interfering with each other as they are not located in the coverage area of one another. However, the FD set located within the coverage of both FD sets is highly unlikely to start its transmission since both the FD nodes in the middle sense that the channel is busy most of the time. The possibility of this case is double for a FD device located in the coverage of two FD sets. Hence, designing efficient and fair MAC protocols for FD networks is a crucial task. This unfairness is penalized by tuning the channel access probability of nodes as a function of proportion of time in which their transmission has been active [10].

Exploiting FD Opportunities via Buffer: In the existing MAC standards, the head-of-line (HOL) packet of the buffer always gets transmitted irrespective of the buffer length, type of packets, and their respective destinations. For instance, if node PT is capable of performing BFD transmission with node PR but PR has no HOL packet for PT (although some packets are available in the buffer), then PT needs to either wait until the HOL packet of PR is transmitted, or initiate an HD transmission. Later, the PR will need to do an HD transmission to the PT. It can thus be concluded that exploiting the right packets from the buffer that can enable FD transmission opportunities may possibly reduce the overall latency at the cost of delays for the destinations of HOL packets [6].

Residual Hidden Node Problem: In practice, the primary and secondary transmitted packets are offset in time and may have different packet lengths. Therefore, the transmission of all nodes will not end up at the same time. Hence, relying only on FD data transmission (even in BFD mode) does not completely solve the hidden node problem. The hidden node problem in FD transmissions due to asymmetric data packets at the transmitter and the receiver can be referred to as the residual hidden node problem. However, the node that finishes data transmission earlier can resolve this issue by transmitting busy tone signals until the other node completes its transmission [11].

Contention in Asynchronous TNFD Mode: Asynchronous TNFD mode occurs when a transmitter, say the PT, has already started transmission and a new transmitting node emerges who wants to transmit to the PT. In this case, the traditional RTS/CTS mechanism cannot be implemented as the PT is transmitting already. Thus the new node has no way to know about the other nodes who may be trying to transmit to the PT at the same time. Therefore, resolving a contention becomes a challenging task in asynchronous TNFD mode [10].

Deafness in Directional Antennas: If FD systems use directional antennas, then the neighboring nodes around the PT and the ST are unable to detect these transmissions. This scenario is known as the deafness problem in directional transmission [12]. Due to the deafness problem, the neighboring nodes try to access a particular channel assuming that it is available for their transmission. This might end up with a collision in the ongoing transmission. A centralized MAC protocol can therefore be useful in avoiding such collisions as the central coordinator might know the locations of its registered nodes and their corresponding transmission directions.

Existing FD-MAC Protocols

The existing FD-MAC protocols in the literature can be classified based on their operation, handshaking mechanism, transmission mode, time synchronization, and number of channels used for resource allocation, as illustrated in Table 1. In this section we will discuss three key mechanisms that can be utilized in the design of FD-MAC protocols [6]: shared random backoff; header snooping; and collision avoidance with RTS/CTS exchange. In order to execute these mechanisms, the basic structure of the HD packet does not need to be modified significantly, i.e. all the basic fields in the HD packet header (i.e. source/destination address, packet duration, fragmentation, etc.) remain the same. However, to exchange the basic information among the FD nodes, the following unique fields/identifiers are required.

The basic approaches that are currently used in FD-MAC protocols to exploit proper nodes and modes for FD transmission are shared random backoff, header snooping, and request-to-send (RTS)/clear-to-send (CTS) mechanisms.

Scenario	Technique
Network architecture	Centralized [13] or distributed [7, 14, 15]
Control information exchange	RTS/CTS [1, 14] or ACK [5, 9, 10] in data transmission channel, dedicated or dynamically configurable control channel
Spectrum access	Contention [11, 16], time slotted [13], or hybrid
Transmission mode	Bi-directional FD [6, 7] or three node FD [6, 7]
Time adjustment (synchronous or asynchronous)	RTS/CTS [1, 14], ACK [5, 9, 10], or beacon [13]
Channel usage	Single channel [5–7] and multiple channel [11, 16]

Table 1. Classification of FD-MAC protocols.

FD Transmission Mode (FDM) Field: At a given time, the set of nodes selected to initiate FD transmission can operate in either one of the four transmission modes depicted in Fig. 1. To indicate/select/suggest a mode, the FD packet thus needs a separate field with the length of two bits (i.e. '00' for HD-mode, '01' for BFD-mode, '10' for DBTM, and '11' for SBTM).

Full-Duplexing Duration (FDDUR): A two bytes field to indicate the duration of FD transmission is also required. Otherwise, the nodes participating in FD transmission may not synchronize and increase the collision rates.

Complex FD-MAC protocols may require further information exchange among the nodes [6].

Shared Random Backoff (SRB)-Based FD-MAC [6]

In SRB-based protocols, all nodes that have performed handshaking for FD communications and have transmitted at least a data packet delay their transmission for a common duration with the intention of allowing other nodes to utilize the channel. Information about this duration needs to be sent to other nodes in the network by including an additional field in the FD packet header known as the SRB field. SRB-based MAC protocols are well-suited for BFD when participating nodes have many packets to exchange, because the problems in the backoff counter countdown mechanism turn out to be significant in TNFD mode and degrades performance due to the lack of synchronization. SRB protocols are implicitly synchronous due to the common backoff and allow FD nodes to spare channel for other starving nodes.

Illustration: Let us assume that there are two nodes and they want to exchange data between them. Let one node win the contention (PT) and start transmission to the intended receiver (PR). Since the PT has a large amount of data to transmit to the PR, the PT sets the FDM field in the FD header to '01' and the SRB field to a random backoff value. After the PR successfully decodes this packet, it knows that the PT has more packets to transmit and the preferred transmission mode is bidirectional transmission. Since the PR also has many packets to transmit to the PT, the PR sends an ACK to the PT by setting the FDM field in the FD header to '01'. Then, both the PT and PR start their transmissions simultaneously after waiting for the random backoff time that was informed by the PT. This scenario is a *forced shared random backoff* since PR has to follow the backoff time informed by the PT. Otherwise, the PR also can propose a backoff time to the PT by setting the SRB field in the ACK packet. In this scenario, either minimum or maximum backoff time should be selected by both the nodes, and this selection should be set by the MAC protocol a priori. Note that in any SRB-based MAC protocol the first transmission is a half-duplex one.

Header Snooping-Based FD-MAC [6]

In header snooping-based MAC protocols, the primary transmitted packet header is decoded by at least single registered node in the network, excluding the primary receiver. Due to this header snooping, the FD nodes transmit in different time stamps, which results in asynchronous transmission. The header snooping-based MAC protocols are well-suited for applications where nodes operate in SBTM since the PT transmitted packet can be snooped by other nodes located withing the PT's transmission region (Fig. 1d).

Illustration: Let us assume that there are three nodes in the network and one node wins the contention (PT). Then the PT sets its FDM field in the FD packet header to '01' and transmits its packet to the PR indicating its preferred transmission mode as bidirectional transmission. The PR sets its FDM field in the FD ACK header to '00' and it indicates that only HD is possible. This indirectly conveys to the PT that the PR does not have any packet to transmit. Meanwhile, another node (ST) in the network, which wants to transmit packets to the PT, snoops the PT's packets and knows that the PT requests a bidirectional transmission from the PR. However, the ST does not have any idea about the PR's response if the ST is located outside the PR's transmission region. Since the PR does not have any packet to transmit to the PT, the PT sets its FDM field in the FD packet header to '11' and this packet is snooped by the ST. After the ST notices that the FDM field in the PT's packet header is '11,' it starts its transmission simultaneously with the PT. Note that the first packet transmission in the header snooping-based MAC protocol does not utilize full-duplexing.

RTS/CTS-Based FD-MAC [14]

The RTS/CTS mechanism that is used in HD transmission can be utilized even in FD transmission to mitigate the hidden node problem. In FD communication, the FD RTS/CTS headers are obtained by adding a FDM and FDDUR field to the HD RTS/CTS headers. Additionally, another address field is added to the HD CTS header that indicates SR's or ST's address. Let us denote that address field as the STorSRAddress field. Hence, the basic operation of the RTS/CTS-based FD-MAC protocols can be illustrated as follows.

MAC protocol	Mechanism	Application	Benefits	Drawbacks	Transmission modes
Distributed MAC [7]	Header snooping	Cellular, D2D, relay, WLAN	Low overhead in handshaking, considers inter-node interference, select nodes for TNFD based on the signal-to-interference ratio (SIR)	Busy tone signaling required, duration of header snooping period cannot be used for data transmission	BFD, DBTM, SBTM
RCTC MAC [15]	Signature based RTS/CTS	Cellular, D2D, rxelay networks	Low overhead signaling, identify nearby transmissions to maximize system throughput	Vulnerable to collision	BFD, DBTM, SBTM
ContraFlow [10]	Header snooping	Cellular, D2D, relay networks	Does not reserve a channel, fairness is considered	Busy tone signaling, no handshake before transmitting packets	BFD, DBTM, SBTM
Distributed MAC [2, 5]	Header snooping	Cellular, D2D networks	Lower signaling overhead	No handshake before transmitting packets	BFD
Wormhole switching-based MAC [9]	Header snooping	Multi-hop networks	Lower signaling overhead	No handshake before transmitting packets	DBTM
Directional antenna [12]	Header snooping	Multi-hop networks	No ACK frame of contention window	Vulnerable to collision	Multi-hop networks-based on DBTM
RFD-MAC [17]	Header snooping	Relay networks	Asynchronous approach with low signaling overhead	Higher collision period in contention	DBTM
JANUS [13]	Header snooping	Relay networks	Achieve higher throughput by eliminating random backoff, nodes transmit multiple packets with a single set of control packets	Higher collision period in contention	DBTM
SRB-based MAC [6]	SRB	Cellular, D2D	Both the nodes fully utilized data transmission duration	Independent backoff counter count down of different nodes	BFD

Table 2. Qualitative overview of existing FD-MAC protocols.

Illustration: Let us assume there are three nodes and one wins the contention without loss of generality. Then the node that wins the contention (PT) sets its FDM field in the FD RTS header to '01' and transmits a packet to the intended receiver (PR) indicating that its preferred transmission mode is a bidirectional transmission. Let us assume that the PR does not have any packet to transmit to the PT but it has many packets to transmit to another node in its transmission range. Then the PR (ST) sets its FDM field in the FD CTS header to '10' and the STorSRAddress field in the CTS header, which is the address of the node to whom the PR (ST) wants to transmit its data. With the reception of this packet, the PT knows in which transmission mode it is going to operate and waits for another CTS duration. Note that this CTS packet acts as an RTS to the SR. Hence, within this waiting CTS time duration, the SR transmits a CTS packet to the ST (PR) after setting the FDM field to '00.' The nodes around the SR receive this packet and know that a transmission is going on that channel. After the SR's CTS transmission, both the PT and ST start to transmit their data to the intended receivers simultaneously.

Comparative Analysis

A qualitative comparison among the existing FD-MAC protocols is shown in Table 2. The characteristics of each protocol are captured in terms of their underlying mechanism (SRB, header snooping, and RTS/CTS), their application in different transmission modes (BFD, SBTM, and DBTM) and network topologies, benefits, and drawbacks. Note that, keeping in view the unique characteristics of various kinds of FD-MAC protocols, hybrid protocols can also be developed.

Figure 3 depicts the normalized throughput of FD nodes for different FD-MAC protocols when multiple nodes contend for a given channel. As expected, the system performance deteriorates with the increasing number of nodes in the network due to the increasing collisions during contention. This figure reveals that the SRB-based FD-MAC protocol with RTS/CTS outperforms all other approaches because both the nodes (PT and ST) fully utilize the transmission duration from beginning to end for data transmission simultaneously. Surprisingly, both the bi-directional-SRB and SBTM-header snooping methods without RTC/CTS show close per-

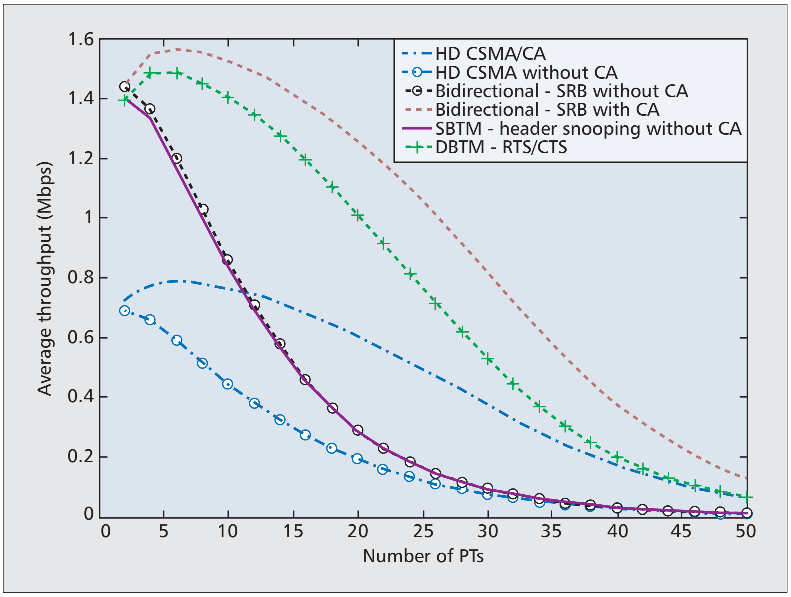


Figure 3. Throughput as a function of the number of PTs for different MAC protocols. Simulation parameters: PHY header = 128 bits, MAC header = 272 bits, RTS = 160 + PHY header, CTS = 128 + PHY header, Short inter-frame space (SIFS) = 28, Payload = 1042 bytes, mini-slot = SIFS + RTS+ SIFS + CTS and contention window size = 16.

formance because the signaling overhead in the header snooping approach is not significant compared to the packet size. However, Fig. 3 clearly reveals that the RTS/CTS-based MAC approach for DBTM outperforms all other approaches (except SRB with CA) even with higher signaling overhead due to the collision avoidance nature of RTS/CTS.

Design Challenges for FD-MAC and Open Research Issues

In this section we highlight design issues in existing FD-MAC and discuss possible solution techniques that can be incorporated with the design of existing FD-MAC. To this end, implications of FD technology on resource and interference management in cellular networks are also discussed.

Design Challenges for FD-MAC Protocols

Node Selection: Existing FD-MAC protocols are designed for a predefined set of nodes (i.e. two or three node set) and thus do not address the challenge of node selection in multi-node FD networks while considering the overall utility of FD transmission and the inter-node collisions. Note that node selection is even more challenging for TNFD modes due to intra-node/inter-node collisions. A straightforward approach is to perform location-aware selection of nodes, i.e. considering the distance from the PT to the SR in DBTM, and considering the distance from the ST to the PR in SBTM. However, this might increase the performance loss of the secondary link due to far-away distances. In this context, optimal location-aware node selection algorithms need to be developed that can minimize intra-node interference without causing excessive degrading of the signal strength of the corresponding link. Further, the inter-node collisions due to the hidden nodes can be alleviated by transmitting RTS/CTS signaling before establishing the FD connection or by transmitting busy tones.

Fairness in the Secondary Link: In existing FD-MAC protocols, the node that wins the contention (e.g. PT) starts its transmission on its best channel without considering the quality of the secondary link. The channel that is selected for primary transmission might be the worst channel for the secondary link. Therefore, designing a MAC protocol that considers both primary and secondary link conditions is highly desirable to improve the overall utility of FD transmission.

Channel Selection: The channel selection in FD-MAC protocols is challenging due to multiple transmission modes and different interference scenarios.

- As mentioned earlier, selecting a common channel becomes more crucial in TNFD mode. The best common channel for TNFD mode can be obtained using a graph based approach. After the channel is selected the process of handshaking among the FD set is also a challenge that needs to further exploit the conventional approaches, e.g. common hopping sequence, control channel, and rendezvous scheme.
- To enhance the fairness between primary and secondary links as mentioned earlier, selecting a common channel that considers the overall utility of the primary and secondary link is crucial and not an easy task due to the signaling overhead. To mitigate this effect, all the participating nodes have to handshake before finalizing their transmit channel. In this context, studying a low overhead FD-MAC protocol is also required.

Backoff Counter Countdown Mechanism: In SRB-based MAC protocols [6], FD-enabled nodes select a common backoff time to start their next transmission simultaneously. However, different nodes independently count down their backoff counters since each node may observe the selected channel's idleness differently due to the heterogeneity of the network. As a result, the nodes that handshake for full-duplexing with a common backoff start their transmission in two different timestamps. This degrades the system performance due to the lack of synchronization in SRB-based MAC protocols.

Control Signaling Overhead: In the existing HD communications, data and control information is exchanged in different frequency bands/time-slots. FD communication is a very good solution to moderate the affirmation issues in HD communication. FD-MAC protocols can potentially exploit the simultaneous transmission and reception to reduce the signaling overhead by discovering the transmission mode and its corresponding nodes, coordination of participating nodes, full-duplexing time slots, and channel identification, etc.

Exploiting FD Opportunities via the Buffer: As discussed earlier, the performance of FD networks can be improved by exploiting FD opportunities via the buffer. However, blindly delaying the HOL packet is not feasible, so a FD-MAC protocol should be developed that optimally decides the duration of the delay for the HOL packet.

Performance Analysis of FD-MAC: As in other MAC protocols, the analysis of FD-MAC protocols is also crucial to understand the fundamental performance limits and to gain design insights. However, analysis of FD-MAC is challenging due to the simultaneous transmission, multiple transmission modes, different packet sizes, busy tones, time adjustments, and signaling overhead.

Implications of FD in Cellular Networks

Cooperation in Cellular Networks: To mitigate uplink to downlink and downlink to uplink inter-cell interference issues, all nearby BSs need to cooperate and optimize the proportion of FD transmission events in a coordinated manner.

Traffic Load-Aware FD Networks: The asymmetry of uplink and downlink data traffic may not directly affirm the significance of FD transmission in cellular networks. Typically, downlink traffic is much more significant compared to the uplink traffic. It is thus of immense importance to optimize the proportion of FD transmissions considering the uplink/downlink traffic load intensities.

Clustering in Small Cell Networks: FD transmission can be applied more frequently in low power small BSs due to the high prospects of SI cancellation. However, uplink to downlink and downlink to uplink inter-cell interference becomes even more critical in densely deployed small cell networks. Efficient clustering methods/techniques need to be adopted that can allow nearby small-cells to coordinate their FD transmissions. For instance, a cluster of small-cells may coordinate the execution of FD transmissions such that all small-cells within that cluster select FD mode in a sequential manner. In this way, the aforementioned interference issues can possibly be minimized. In this regard, inter-cluster cooperation may also be exploited.

Cooperative FD-D2D Transmissions: Device-to-device (D2D) communications allow nearby users to establish a direct link to communicate with each other. While D2D links are typically exploited for HD data transmissions, they may be used for interference mitigation or channel selection if exploited in FD mode. For instance, a D2D transmitter can hear interference signals on the FD receiver and can provide the interference information to its intended receiver along with the data packets. In addition, any downlink cellular user can also exploit the FD-D2D communication to forward interference knowledge to a nearby user. This knowledge can help nearby users in either interference cancellation or channel selection.

Conclusion

Full-duplexing will be one of the main candidate technologies for future wireless communication systems to exploit spectrum, and this will practically enable applications such as cognitive radios. We believe that the maximum gain due to full-duplexing can only be achieved through a smart FD-MAC protocol that jointly addresses the physical layer and MAC layer aspects. In this article we have highlighted the major challenges that need to be considered in designing smart FD-MAC protocols. The possible approaches to solve these challenges have been discussed. Also, the interference management challenges that arise in cellular networks due to the adoption of FD technology have also been discussed.

Full-duplexing will be one of the main candidate technologies for future wireless communication systems to exploit spectrum and this will practically enable applications such as cognitive radios. We believe that the maximum gain due to full-duplexing can only be achieved through a smart FD-MAC protocol that jointly addresses the physical layer and MAC layer aspects.

References

[1] W. Li *et al.*, "System Scenarios and Technical Requirements for Full-Duplex Concept," DUPLO Deliverable D1.1, 2013.

[2] M. Duarte *et al.*, "Design and Characterization of a Full-Duplex Multi-Antenna System for Wi-Fi Networks," *IEEE Trans. Vehic. Tech.*, vol. 63, no. 3, 2014, pp. 1160–77.

[3] M. Duarte, C. Dick, and A. Sabharwal, "Experiment-Driven Characterization of Full-Duplex Wireless Systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, 2012, pp. 4296–307.

[4] D. Bharadia, E. McMilin, and S. Katti, "Full duplex radios," *Proc. ACM SIGCOMM*, 2013.

[5] M. Jain *et al.*, "Practical, Real-Time, Full Duplex Wireless," *Proc. ACM MobiCom*, 2011.

[6] A. Sahai, G. Patel, and A. Sabharwal, "Pushing the Limits of Full Duplex: Design and Real-Time Implementation," Rice University Technical Report TREE1104, 2011.

[7] S. Goyal *et al.*, "A Distributed MAC Protocol for Full Duplex Radio," *Proc. ASILOMAR*, 2013.

[8] X. Kang, C.-K. Ho, and S. Sun, "Full-Duplex Wireless-Powered Communication Network with Energy Causality," arXiv preprint 1404.0471, 2014.

[9] J. I. Choi *et al.*, "Achieving Single Channel, Full Duplex Wireless Communication," *Proc. ACM MobiCom*, 2010.

[10] N. Singh *et al.*, "Efficient and Fair MAC for Wireless Networks with Self-Interference Cancellation," *Proc. IEEE WiOpt*, 2011.

[11] Y. Zhang *et al.*, "FD-MMAC: Combating Multi-Channel Hidden and Exposed Terminals Using a Single Transceiver," *Proc. IEEE INFOCOM*, 2014.

[12] K. Miura and M. Bandai, "Node Architecture and MAC Protocol for Full Duplex Wireless and Directional Antennas," *Proc. PIMRC*, 2012.

[13] P. Math and D. Hoenig, "Janus: A Novel Use of Formal Verification for Targeted Behavioral Equivalence," *Proc. High Level Design Validation and Test Wksp.*, 2008.

[14] W. Cheng, X. Zhang, and H. Zhang, "RTS/CTS Mechanism Based Full-Duplex MAC Protocol for Wireless Networks," *Proc. IEEE Globecom Wksps. (GC Wksps.)*, 2013.

[15] W. Zhou, K. Srinivasan, and P. Sinha, "RCTC: Rapid Concurrent Transmission Coordination in Full Duplex Wireless Networks," *IEEE Int'l. Conf. Network Protocols*, 2013

[16] N. Choi, M. Patel, and S. Venkatesan, "A Full Duplex Multi-Channel MAC Protocol for Multi-Hop Cognitive Radio Networks," *Proc. IEEE CROWNCOM*, 2006.

[17] K. Tamaki *et al.*, "Full Duplex Media Access Control for wireless Multi-Hop Networks," *Proc. IEEE VTC Spring*, 2013.

Biographies

Karaputugala G. Madushan Thilina [S'11] received the B.Sc. degree from the University of Ruhuna, Sri Lanka in 2006, and the M.Eng. degree from the Asian Institute of Technology, Thailand, in 2010. He is currently working toward a Ph.D. degree in the Department of Electrical and Computer Engineering, University of Manitoba, Canada. His primary research interests include the design and analysis of novel medium access control protocols for cognitive radio networks, device-to-device networks, full duplex networks, resource allocation for cooperative/cognitive relay networks, and cooperative spectrum sensing in cognitive radio networks. He is currently a recipient of the University of Manitoba Graduate Fellowship (UMGF). Madushan has served as a technical program committee member for several IEEE conferences.

Hina Tabassum received the B.E. degree in electronic engineering from the NED University of Engineering and Technology (NEDUET), Karachi, Pakistan, in 2004. During her graduate studies she received two gold medals from NEDUET and Siemens for securing the first position among all engineering universities of Karachi. She then worked as a lecturer in NEDUET for two years. In September 2005 she joined the Pakistan Space and Upper Atmosphere Research Commission (SUPARCO), Karachi, Pakistan, and received there the best performance award in 2009. She completed her Masters and Ph.D. degree in communications engineering from NEDUET in 2009 and King Abdullah University of Science and Technology (KAUST), Makkah Province, Saudi Arabia, in May 2013, respectively. Currently she is working as a post-doctoral fellow at the University of Manitoba (UoM), Canada. Her research interests include wireless communications with a focus on interference modeling, spectrum allocation, and power control in heterogeneous networks.

EKRAM HOSSAIN [F'15] is currently a professor in the Department of Electrical and Computer Engineering at the University of Manitoba, Winnipeg, Canada. He received his Ph.D. in electrical engineering from the University of Victoria, Canada, in 2001. He is a Fellow of the IEEE. His current research interests include design, analysis, and optimization of wireless/mobile communications networks, cognitive radio systems, and network economics. He has authored/edited several books in these areas (http://home. cc.umanitoba.ca/~hossaina). He serves as the editor-in-chief of *IEEE Communications Surveys and Tutorials*, and an editor for *IEEE Wireless Communications*. He also currently serves on the IEEE Press Editorial Board. Previously he served as the area editor for *IEEE Transactions on Wireless Communications* in the area of "Resource Management and Multiple Access" from 2009–2011, an editor for the *IEEE Transactions on Mobile Computing* from 2007–2012, and an editor for the *IEEE Journal on Selected Areas in Communications — Cognitive Radio Series* from 2011-2014. Dr. Hossain has won several research awards, including the University of Manitoba Merit Award in 2010 and 2014 (for research and scholarly activities), the 2011 IEEE Communications Society Fred Ellersick Prize Paper Award, and the IEEE Wireless Communications and Networking Conference 2012 (WCNC'12) Best Paper Award. He is a Distinguished Lecturer of the IEEE Communications Society for the term 2012–2015. He is a registered professional engineer in the province of Manitoba, Canada.

DONG IN KIM [S'89, M'91, SM'02] received the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1990. He was a tenured professor with the School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada. Since 2007 he has been with Sungkyunkwan University (SKKU), Suwon, Korea, where he is currently a professor with the College of Information and Communication Engineering. He has served as an editor and a founding area editor of cross-layer design and optimization for *IEEE Transactions on Wireless Communications* from 2002 to 2011. From 2008 to 2011 he served as the co-editor-in-chief for the *Journal of Communications and Networks*. He is currently the founding editor-in-chief for the *IEEE Wireless Communications Letters,* and has been serving as an editor of spread spectrum transmission and access for the *IEEE Transactions on Communications* since 2001. He was the recipient of the Engineering Research Center (ERC) for Wireless Energy Harvesting Communications Award.

Full Duplex Cellular Systems: Will Doubling Interference Prevent Doubling Capacity?

Sanjay Goyal, Pei Liu, Shivendra S Panwar, Robert A. DiFazio, Rui Yang, and Erdem Bala

Sanjay Goyal, Pei Liu, and Shivendra S Panwar, are with New York University Polytechnic School of Engineering

Robert A. DiFazio, Rui Yang, and Erdem Bala are with InterDigital, Inc.

ABSTRACT

Recent advances in antenna and RF circuit design have greatly reduced the crosstalk between the transmitter and receiver circuits on a wireless device, which enable radios to transmit and receive on the same frequency at the same time. Such a full duplex radio has the potential to double the spectral efficiency of a point-to-point radio link. However, the application of such a radio in current cellular systems (3GPP LTE) has not been comprehensively analyzed. This article addresses the fundamental challenges in incorporating full duplex radios in a cellular network to unlock the full potential of full duplex communications. We observe that without carefully planning, full duplex transmission might cause much higher interference in both uplink and downlink, which greatly limits the potential gains. Another challenge is that standard scheduling methods which attempt to achieve the maximum capacity gain lead to a severe loss in energy efficiency. In this article, we identify new tradeoffs in designing full duplex enabled radio networks, and discuss favorable conditions to operate in full duplex mode. New scheduling algorithms and advanced interference cancellation techniques are discussed, which are essential to maximize the capacity gain and energy efficiency. Under this new design, most of the gain is achievable with full duplex enabled base stations, while user equipment still operates in half duplex mode.

INTRODUCTION

Cellular networks have entered a period of unprecedented change and ever increasing importance to the economy and society. To support the exploding demand for video and other high-rate data services, such networks have begun a major shift from being voice-centric, circuit-switched, and centrally optimized for coverage toward being data-centric, packet-switched, and organically deployed for maximum capacity. Cisco's well-known capacity forecasts estimate that mobile data will grow 11-fold in the period 2013–2018 [1]. Such drastic changes are largely due to new trends in usage, with smartphones in particular leading to a proliferation of data-hungry applications. Full duplex (FD) technology, which has the potential to double spectrum efficiency, provides a step toward meeting high demand without requiring new spectrum.

Traditional radio transceivers are generally not able to receive and transmit on the same frequency channel simultaneously because of the crosstalk between the transmitter (Tx) and the receiver (Rx). Given that the intended received signal over the air can be more than a million times weaker than the transmitted signal due to path loss and fading, it is very difficult, if not impossible, to detect the received signal under internal interference from the transmission circuits and antennas. Thus, most of today's communication systems are half duplex (HD) in each channel, while bidirectional transmissions are based on two orthogonal channels, typically using time (time-division duplex or TDD) or frequency (frequency-division duplex or FDD) dimensions, to provide separation between transmit and receive signals.

Recent advances in antenna design and RF circuits [2–6] have made a great leap forward in reducing self-interference (SI) between the Tx and Rx circuits on a common channel. New antenna designs, combined with analog and digital cancellation, are employed to remove most of the SI from the Rx path to allow decoding of the desired received signal. This was demonstrated using multiple antennas positioned for optimum cancellation [2–4] and later for single-antenna systems [5, 6]. Jain *et al.* [3] used a combination of signal inversion and digital cancellation, which achieves 73 dB of cancellation for a 10 MHz orthogonal frequency-division multiplexed (OFDM) signal. In another work, a combination of passive, analog, and digital cancellation achieved a median cancellation of 85 dB, with a minimum of 70 dB and a maximum of 100 dB over a 20 MHz WLAN channel [4]. An antenna feed network was proposed by Knox [5], a prototype of which provided Tx/Rx isolation of 40–45 dB before analog and digital cancellation. Bhara-

0163-6804/15/$25.00 © 2015 IEEE

The high interference in each direction raises several questions regarding the potential performance of FD operation in a cellular system. Even in a single cell, FD gain cannot be achieved without careful selection of the uplink and downlink UEs for simultaneous transmission.

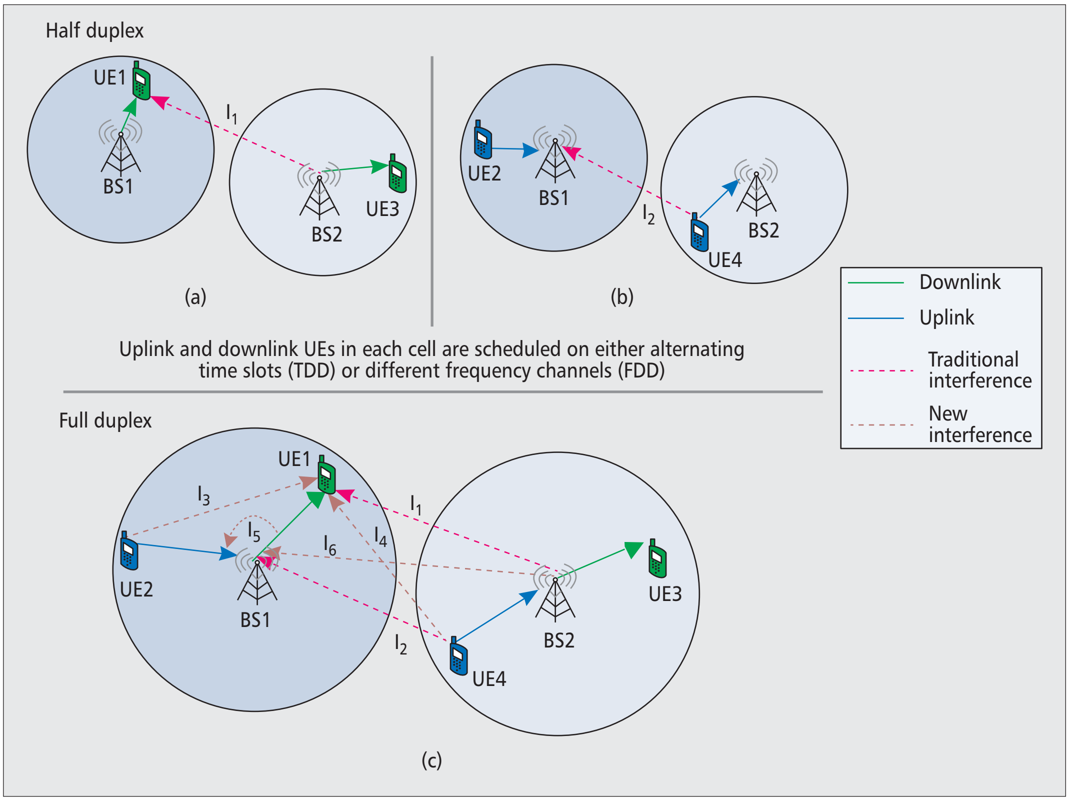


Figure 1. Half duplex and full duplex multi-cell scenarios.

dia *et al.* [6] proposed a single-antenna design that is able to achieve up to 110 dB of cancellation over an 80 MHz bandwidth.

Although extensive efforts have been made in designing and implementing wireless terminals with FD capability, to the best of our knowledge, the impact of FD transmissions in a wireless network in terms of system capacity and energy efficiency has not been extensively analyzed, especially in a cellular environment. FD operation in a cellular small cell system was investigated in the DUPLO project [7]. This work was mostly focused on a joint uplink and downlink beamforming design in a single cell. Alves *et al.* [8] derived average spectral efficiency for a stochastic geometry-based dense small cell environment with both base station (BS) and user equipment (UEs) having FD capability. These analyses do not consider the multi-UE diversity gain that comes through scheduling of the appropriate UEs with power adjustments to mitigate inter-node interference.

Given the premise that sufficient SI cancellation (SIC) is feasible, this article considers access links of a multi-cell system, where the BSs have the capability of FD operation, and UEs are limited to HD operation. Since FD requires a significant change in the hardware with higher cost and power usage, it is more practical to upgrade the infrastructure elements. Hence, our preferred FD cellular system will have the above assumption (i.e., FD BSs and HD UEs) unless specified otherwise.

In the FD cellular system, as shown in Fig. 1, each BS can simultaneously schedule uplink and downlink transmissions on the same resource block. Therefore, each transmission potentially experiences higher interference from within the cell and from neighboring cells compared to the traditional HD cellular systems. The high interference in each direction raises several questions regarding the potential performance of FD operation in a cellular system. Even in a single cell, FD gain cannot be achieved without careful selection of the uplink and downlink UEs for simultaneous transmission [9]. The focus of this article is:

- Identifying the new interference scenario in FD cellular systems
- Finding a suitable cellular environment to deploy FD radios
- Proposing a new proportional fairness-based hybrid scheduling algorithm
- Analyzing and proposing techniques to address energy efficiency

It should be mentioned that in this article, we assume omnidirectional transmission and reception of the signals at all BSs and UEs. Using directional transmission and reception, which may be achieved by beamforming with multiple antennas or other physical (PHY) layer techniques, can certainly change the interference situation mentioned above. Interference mitigation at the PHY layer is out of the scope of this article and will be discussed in our future work.

The rest of the article is organized as follows. We first introduce the FD cellular system. Next, we categorize and discuss the impact of each of the new interference sources during FD operation. The design of a scheduling algorithm that minimizes interference and maximizes FD gains is then described. The impact of FD on energy efficiency is presented, and then we conclude the article.

Full Duplex Operation in a Cellular Environment

Figure 1 illustrates FD operation in a multi-cell environment, and compares it to traditional TDD/FDD operation. In this figure, a two-cell network with two UEs in each cell is considered. UE1 and UE3 are in downlink mode while UE2 and UE4 are in uplink mode. In this section, for illustration purposes, we assume a *synchronous* HD multi-cell deployment in which:

- In a given time slot, all cells schedule either uplink or downlink transmission.
- The number of time slots is divided equally between uplink and downlink transmission.

For simplicity, we only show the impact of interference on users of cell 1. In HD operation, as shown in Fig. 1a, in the downlink, UE1 gets interference (I_1) from BS2, which is transmitting to UE3 at the same time. Similarly, in the uplink, as shown in Fig. 1b, BS1 gets interference (I_2) from the uplink signal of UE4. Figure 1c shows the impact of FD operation on both inter-cell and intra-cell interference, where both BS1 and BS2 transmit in FD mode, and each schedules a downlink and an uplink UE simultaneously. During FD operation, in the downlink, UE1 gets interference (I_1) from BS2 and interference (I_3 and I_4) from the uplink signals of UE2 and UE4. Similarly, in the uplink, BS1 gets interference (I_2) from UE4 and interference (I_6) from the downlink signal of BS2, as well as Tx-to-Rx SI (I_5).

In downlink, the additional interference during FD operation comes from all the active uplink UEs on the same channel, which we call *UE-to-UE interference*. In uplink, additional interference comes from (1) Tx-to-Rx SI, which we simply refer to as self-interference (SI) at its own BS, and (2) neighboring BSs due to simultaneous downlink transmission, which we call *BS-to-BS interference*. In the next section, we discuss the nature of each of these additional interference sources in detail.

Interference During Full Duplex Operation

BS-to-BS Interference

The path loss between BSs is generally much smaller than the path loss between BSs and UEs, especially in an outdoor environment, where BSs are usually installed at higher elevations, and have fewer obstructions and less absorption between each other [10]. When coupled with large transmission power from BSs, this interference becomes very strong. Techniques to mitigate BS-to-BS interference are necessary to realize FD BS deployment, especially in outdoor environments. The issue of BS-to-BS interference is well discussed in the Third Generation Partnership Project (3GPP) documents on dynamic TDD HD deployment [11]. Some methods to mitigate BS-to-BS interference include null forming in the elevation angle at BS antennas [10] and interference management through the cloud radio access network (C-RAN) architecture [12].

Self-Interference

Although there has been significant effort to improve cancellation circuits, the applicability of FD radios in large cells is still questionable. For example, consider a cell with 1 km radius, where the path loss at the cell edge is around 130 dB [11]. Assuming equal per channel transmission power in the uplink and downlink directions, the signal arriving at the BS is 130 dB lower than the signal transmitted. With the best SI cancelling circuit known to date, which is capable of suppressing the crosstalk by 110 dB [6], the received signal-to-interference ratio (SIR) is at most –20 dB. In a typical isolated indoor cell with a radius of about 40 m, as described later, the average received SIR with 110 dB SIC can be as high as 33 dB.

Most of the existing work on FD cellular systems [10, 12] does not consider the effect of SI at BSs. In our view, unless there is a major breakthrough in the cancellation circuit design, FD operations are more practical for small cell deployment, where the smaller coverage area makes it a more suitable environment to deploy FD radios.

UE-to-UE Interference

Since this type of interference depends on the UE locations and their transmission powers, an intelligent coordination mechanism is needed. The goal of the coordination is to select those UEs for simultaneous transmission such that their rate/power allocation would create less interference for each other.

Based on the above discussion, we suggest that:

- A small cell environment is more suited to FD operation.
- Intelligent scheduling of UEs with appropriate rate/power allocation is necessary to extract the capacity gain potential of FD operation.

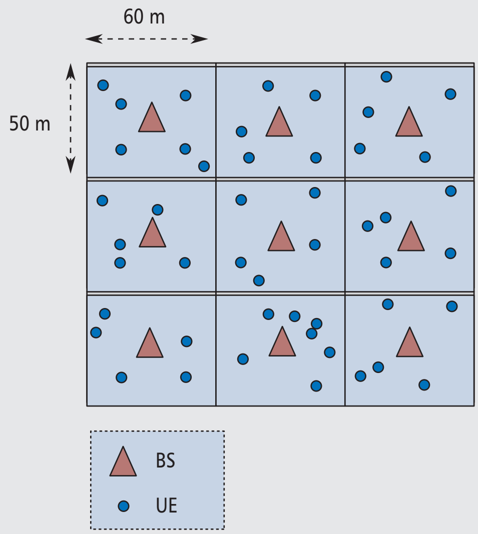
Scheduling in a Full Duplex Cellular System

It is clear from the above discussion that simultaneous uplink and downlink transmission during FD operation comes with additional intra-cell and inter-cell interference. To achieve the potential capacity gain from FD operation, it becomes necessary to intelligently schedule an appropriate combination of downlink and uplink UEs with corresponding transmission rates/powers so that an aggregate network utility can be maximized, while a level of fairness is maintained. Since pure FD transmission may not always be optimal due to the presence of extra interference, we use a hybrid scheduling algorithm where, in a given time slot in a cell, we may schedule only one downlink or uplink UE, or a pair of downlink and uplink UEs.

The scheduling algorithm consists of joint UE selection and rate allocation to maximize the performance of FD operation in a multi-cell system. A proportional fairness-based utility function similar to the one used in [9] is considered, where the objective is to maximize the joint

Since pure FD transmission may not always be optimal due to the presence of extra interference, we use a hybrid scheduling algorithm where, in a given time slot in a cell, we may schedule only one downlink or uplink UE, or a pair of downlink and uplink UEs.

The joint UE selection and rate allocation is a nonlinear non-convex problem with mixed discrete and continuous optimization. Finding a global optimum through an exhaustive search method is computationally difficult, so a suboptimal method is considered.



Parameter	Value
Bandwidth	10 MHz
Number of channels (for each cell)	1
Maximum BS power	24 dBm
Maximum UE power	23 dBm
Thermal noise density	–174 dBm/Hz
Noise figure	BS: 8 dB, UE: 9 dB
Shadowing standard deviation (with no correlation)	LOS: 3 dB NLOS: 4 dB
Path loss within a cell (dB) (R in kilometers)	LOS: 89.5 + 16.9 $\log_{10}$(R), NLOS: 147.4 + 43.3 $\log_{10}$ (R).
Path loss between two cells (R in kilometers)	Max((131.1+42.8 $\log_{10}$ (R)), (147.4 + 43.3 $\log_{10}$ (R)))
Penetration loss	Due to boundary wall of an RRH cell: 20 dB, within a cell: 0 dB

Figure 2. An indoor environment with nine cells.

uplink and downlink utility of the system. We assume a centralized scheduler that has access to global system information (power, channel state information). Scheduling methods with limited or no information exchange among cells will be considered in our future work.

The joint UE selection and rate allocation is a nonlinear non-convex problem with mixed discrete and continuous optimization. Finding a global optimum through an exhaustive search method is computationally difficult, so a suboptimal method is considered. This problem is solved in two steps:

- With maximum power allocated to each UE, we select UEs based on a heuristic greedy method.
- For the selected UEs, geometric programming [13] is used to find the optimum rates/powers such that the aggregate utility in the objective function can be maximized.

The details of the algorithm can be found in [14].

The proposed scheduling method is simulated in indoor remote radio heads (RRHs)/Hotzone cells [15] as shown in Fig. 2, along with a wraparound topology. First, we simulate the system with eight randomly dropped UEs in each cell with full-buffered traffic in both directions. BSs and UEs are assumed to be equipped with single antennas. The simulation parameters are based on 3GPP simulation recommendations [15], and are described in Fig. 2. The channel model used between BSs and UEs is also used between mobile UEs and between BSs for the interference calculations, with the justification that BSs do not have a significant height advantage in the small cell indoor scenario considered. We use the Shannon equation to measure the data rate, where we apply a minimum spectral efficiency of 0.26 b/s/Hz and a maximum spectral efficiency of 6 b/s/Hz to match practical systems.

With these settings, we ran our simulations for different UE drops, each with 1000 time slots, and generated results for both HD and FD systems. For the FD system, different values of BS SIC capability are considered. The residual SI is modeled as Gaussian noise, the power of which equals the difference between the transmit power of the BS and the assumed amount of SIC. For the HD system, results are generated for both synchronous and dynamic TDD operation. In the *dynamic* TDD system, each cell has the flexibility of scheduling its UE in whichever direction provides larger utility at the given time slot. The distributions of average downlink and uplink throughputs for the HD and FD systems are plotted in Figs. 3a and 3b, respectively. The HD system shows a narrow distribution centered near 4 Mb/s in both the uplink and downlink. Since the same kind of channel model is assumed between different nodes, there is not much difference in the interference experienced by a node in synchronous or dynamic TDD systems. Thus, similar results are obtained for these two systems. The FD system shows a wider distribution since the scheduler takes advantage of the random nature of the interference to assign FD operation with an appropriate data rate whenever possible. The throughput gain in both directions increases as the SIC improves. On average, downlink throughput gains of 56, 80, 94, 97, and 98 percent are achieved in the FD system with SIC of 75, 85, 95, 105 dB, and no SI, respectively. Similarly, in the uplink, the corresponding gains are 63, 83, 93, 96, and 97 percent.

We also evaluate the system with non-full

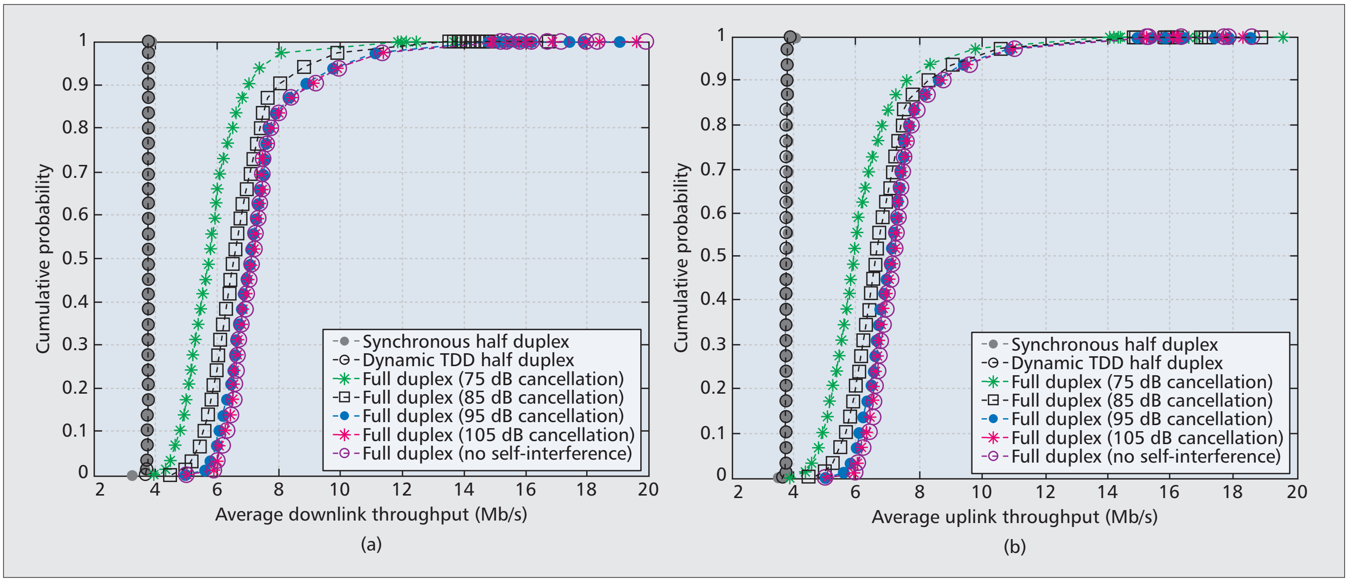


Figure 3. Distribution of average data rates for the half duplex system and full duplex system with different self-interference cancellation capabilities.

buffer FTP traffic [15], where each UE has requests to download/upload files of 1.25 MB. The time interval between completion of a file transmission and an arrival of a new request is exponentially distributed with a mean of 1 s. We calculate the delay for each UE as the total time it experiences from the request arrival to the completion of downloading or uploading a file. Since in an FD system downloading and uploading can take place simultaneously, a significant delay improvement is observed. On average a UE experiences 2.34 s of delay in the downlink and 2.39 s of delay in the uplink in the HD system. In the FD system, in the downlink, this delay reduces to 1.33, 1.05, 0.89, 0.83, and 0.81 s with 75, 85, 95, and 105 dB, and perfect SIC, respectively. Similarly in the uplink, a UE experiences 1.23, 1.01, 0.92, 0.89, and 0.87 s of delay with 75, 85, 95, and 105 dB, and perfect SIC, respectively. Moreover, in the FD system, a UE downloads 48, 69, 83, 90, and 92 percent more files and uploads 56, 75, 86, 88, and 90 percent more files compared to those in the HD system with 75, 85, 95, and 105 dB, and perfect SIC, respectively.

The above results show that the FD system has the potential to increase the capacity of small cells significantly. The related issue of energy efficiency during FD operation to achieve this capacity improvement, which to our knowledge has not been examined before, is presented in the next section.

In this article, we consider symmetric downlink/uplink traffic demands. Asymmetric traffic tends to reduce the need for simultaneous uplink and downlink transmission, which decreases the potential capacity gain of FD operation. However, recent growth in online storage services (Google Drive, Dropbox, iCloud, etc.), and the increasing popularity of photo and video uploading to social networking sites, have increased, and will continue to increase, the uplink traffic volume significantly, making the traffic less asymmetric.

Energy Efficiency of the Full Duplex Cellular System

As energy efficiency becomes more important in cellular system design, it is interesting to examine the energy efficiency of FD operation. Energy efficiency is calculated for each UE by dividing the total throughput by the total power consumed for signal transmission; power consumed in transceiver circuits is not considered. The average per UE downlink and uplink energy efficiency of different systems for the full-buffer traffic case are shown in Table 1 in terabits per joule. Since similar energy efficiency results are obtained for the non-full buffer case, we consider only the full-buffer case in the following discussion. The results show that, compared to HD systems, FD systems experience a significant penalty in energy efficiency. The extra energy is used to combat the additional interference created during FD transmission. Also note that the goal of the scheduler presented above is capacity maximization while maintaining fairness; it does not try to optimize energy efficiency.

In general, as SIC gets better, with the reduction in SI, UE-to-UE and BS-to-BS interference increases due to the larger number of FD transmissions scheduled. Because of this trade-off phenomenon, the impact of SIC on energy efficiency is difficult to explicitly formulate. Our simulations show that uplink and downlink energy efficiency increases as SIC improves. The improvement in SIC provides a significant reduction in interference on the uplink transmission compared to the increase in interference due to the higher number of FD transmissions. This trade-off results in reduced uplink transmit power, which in turn reduces the additional interference on the downlink transmission.

Since the main reason for the lower energy efficiency of the FD system is the additional power needed to combat the extra interference, two solutions can be proposed to alleviate this issue. The first solution is to cancel or mitigate

	Half duplex (synchronous, dynamic TDD	Full duplex (75 dB cancellation)	Full duplex (85 dB cancellation)	Full duplex (95 dB cancellation)	Full duplex (105 dB cancellation)	Full duplex (no self-interference)
Downlink	(3.74, 4.40)	0.045	0.097	0.227	0.326	0.434
Uplink	(4.91, 4.91)	0.017	0.151	0.734	1.360	1.971

Table 1. Average energy efficiency (terabits per joule) for downlink and uplink directions.

		75 dB cancellation	85 dB cancellation	95 dB cancellation	105 dB cancellation	No self-interference
Average throughput gain	Downlink	44%	77%	90%	99%	100%
	Uplink	43%	77%	90%	99%	100%
Average energy efficiency (Tb/J)	Downlink	0.51	2.18	1.59	3.04	3.98
	Uplink	0.31	1.60	0.86	2.66	4.08

Figure 4. Average throughput gain and energy efficiency (terabits per joule) with full duplex UEs.

		75 dB cancellation	85 dB cancellation	95 dB cancellation	105 dB cancellation	No self-interference
Average throughput gain	Downlink	44%	72%	91%	95%	96%
	Uplink	50%	69%	85%	89%	91%
Average energy efficiency (Tb/J)	Downlink	2.02	1.01	0.80	0.75	0.73
	Uplink	1.46	2.47	3.23	3.58	3.61

Figure 5. Average throughput gain and energy efficiency (terabits per joule) with a power allocation method that includes a penalty for higher power consumption.

the additional interference using techniques such as beamforming and sectorization. In this particular small cell indoor scenario, where most of the inter-cell interference is mitigated by penetration loss between the cells, intra-cell interference plays a dominant role during FD operation. If we assume that sufficient SIC is available for the small cell scenario (e.g., 105 dB), allowing FD operation on the UEs (FD UEs) may remove UE-to-UE intra-cell interference. In this case, the BS and one UE in each cell will simultaneously transmit in both uplink and downlink directions. Thus, a downlink UE will not experience intra-cell interference from an uplink UE in the same cell. Our simulation results, shown in Fig. 4, indicate that average throughput gains in such an FD system are 44, 77, 90, 99, and 100 percent in the downlink, and 43, 77, 90, 99, and 100 percent in the uplink for 75, 85, 95, and 105 dB, SIC and no SI, respectively. For the lower SIC case of 75 dB, although the energy efficiency is higher than in the previous case of HD UEs, the throughput is lower. As cancellation improves, there is not much difference in the average throughput from the previous case, but energy efficiency improves significantly. For 105 dB SIC, in the downlink, 3.04 Tb/J is achieved compared to 0.326 Tb/J, and in the uplink, 2.66 Tb/J is achieved as compared to 1.36 Tb/J. These results show that in the higher SIC scenario, it is beneficial to have FD UEs, especially in a small indoor environment. In this case, energy efficiency does not vary monotonically with SIC because of the trade-off mentioned earlier in this section.

A second solution to improve energy efficiency is to keep using HD UEs but implement a more intelligent scheduling algorithm in which, during the rate/power allocation step, a utility function incorporating the cost of using high power is considered, that is,

System Utility = Total System Throughput – Weighted Total Selected UEs' Power Allocation (1)

In the above expression, the first term, *Total System Throughput*, is still calculated based on proportional fairness. The second term, *Weighted Total Selected UEs' Power Allocation*, is computed by applying a penalty for UEs using high power. The level of penalty varies for different UEs. For example, UEs farther from the cell center should have a lower penalty than UEs closer to the center. The details of the algorithm can be found in [14].

As shown in Fig. 5, this type of scheduling causes a modest loss in the throughput gain compared to the original scheduling algorithm; however, a significant improvement in energy efficiency is observed. As an example, an energy efficiency of 2.02 Tb/J is achieved compared to 0.045 Tb/J in the downlink with 75 dB SIC.

Conclusion

Recent progress in antenna and RF circuit design to achieve cancellation of SI between Tx and Rx channels of a wireless device has made it possible to build FD radios. This article is an effort to analyze the system-wide impact of FD transmission in a cellular environment. With current state-of-the-art interference cancellation circuits, we believe it is more feasible to operate in FD mode in small cells. In this article, we focus on an indoor small cell environment. In ongoing work [14], we show that FD is also useful in a sparse outdoor small cell environment. New sources of interference are identified in FD transmission, and we note that if the uplink and downlink transmissions are not coordinated, the added interference greatly reduces the potential gain of FD operation. Due to the hardware/cost requirement to build FD radios, we suggest a network with FD BSs and HD UEs. With our proposed proportional fairness-based scheduler that jointly optimizes UEs and selects rates, we show that the system capacity almost doubles. Another challenge of FD operation is to reduce the energy consumption, since significant energy is consumed to combat the strong interference that is new to FD operation. In our view, energy efficiency could be improved by either enabling FD UEs or using a modified scheduling algorithm that penalizes using high power during FD operation. We conclude that FD is a promising technology with the potential to significantly improve cellular system throughput.

In our view, energy efficiency could be improved either by enabling FD UEs, or using a modified scheduling algorithm that penalizes using high power during FD operation. We conclude that FD is a promising technology with the potential of significantly improving cellular system throughput.

References

[1] "Cisco Visual Networking Index: Forecast and Methodology 2013–2018," Cisco white paper, www.cisco.com, 10 June 2014.
[2] A. K. Khandani, "Methods for Spatial Multiplexing of Wireless Two-Way Channels," U.S. Patent No. 7,817,641, 19 Oct. 2010.
[3] M. Jain *et al.*, "Practical, Real-Time, Full Duplex Wireless," *17th Annual Int'l. Conf. Mobile Computing and Networking*, 2011, pp. 301–12.
[4] M. Duarte *et al.*, "Design and Characterization of a Full Duplex Multi-Antenna System for WiFi Networks," *IEEE Trans. Vehic. Tech.*, vol. 63, no. 3, Mar. 2014, pp. 1160–77.
[5] M. Knox, "Single Antenna Full Duplex Communications Using A Common Carrier," *IEEE 13th Annual Wireless and Microwave Tech. Conf.*, Cocoa Beach, FL, 15–17 Apr. 2012, pp. 1–6.
[6] D. Bharadia, E. McMilin, and S. Katti, "Full Duplex Radios," *ACM SIGCOMM*, Hong Kong, 12–16 Aug. 2013, pp. 375–86.
[7] Duplo: http://www.fp7-duplo.eu/.
[8] H. Alves *et al.*, "On the Average Spectral Efficiency of Interference-Limited Full-Duplex Networks," *IEEE 9th Int'l. Conf. Cognitive Radio Oriented Wireless Networks and Commun.*, Oulu, Finland, June 2014.
[9] S. Goyal *et al.*, "Improving Small Cell Capacity with Common-Carrier Full Duplex Radios," *IEEE ICC*, Sydney, Australia, June 2014.
[10] Y.-S. Choi and H. Shirani-Mehr, "Simultaneous Transmission and Reception: Algorithm, Design and System Level Performance," *IEEE Trans. Wireless Commun.*, vol. 12, no. 12, Oct. 2013, pp. 5992–6010.
[11] 3GPP TR 36.828, "Further Enhancements to LTE Time Division Duplex (TDD) for Downlink-Uplink (DL-UL) Interference Management and Traffic Adaptation," v. 11.0.0, www.3gpp.org, June 2012.
[12] O. Simeone, E. Erkip, and S. Shamai, "Full-Duplex Cloud Radio Access Networks: An Information-Theoretic Viewpoint," *IEEE Wireless Commun. Lett.*, vol. 3, no. 4, Aug. 2014, pp. 413–16.
[13] M. Chiang *et al.*, "Power Control by Geometric Programming," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, July 2007, pp. 2640–51.
[14] S. Goyal *et al.*, "Full Duplex Operation for Small Cells," CoRR, vol. abs/1412.8708, 2014, http://arxiv.org/abs/1412.8708.
[15] 3GPP TR 36.814, "Further Advancements for E-UTRA Physical Layer Aspects," v. 9.0.0, www.3gpp.org, May 2010.

Biographies

Sanjay Goyal (sanjay.goyal@nyu.edu) received his B.Tech. degree in communication and computer engineering from the LNM Institute of Information Technology, India, in 2009, and his M.S. degree in electrical engineering from NYU Polytechnic School of Engineering, New York, in 2012. Currently, he is a Ph.D. candidate in the ECE Department at NYU Polytechnic School of Engineering. His research interests are in designing and analyzing wireless network protocols with full duplex communication, especially for the MAC layer.

Pei Liu is a research assistant professor in the ECE Department at NYU Polytechnic School of Engineering. He received his Ph.D. degree in electrical and computer engineering from NYU Polytechnic School of Engineering in 2007. He received his B.S. and M.S. degrees in electrical engineering from Xi'an Jiaotong University, China, in 1997 and 2000, respectively. His research interests are in designing and analyzing wireless network protocols with an emphasis on cross-layer optimization, especially for the PHY and MAC layers.

Shivendra S. Panwar [F] is a professor in the ECE Department at NYU Polytechnic School of Engineering. He is the director of the New York State Center for Advanced Technology in Telecommunications (CATT), the faculty director of the New York City Media Lab, and a member of NYU Wireless. He co-authored *TCP/IP Essentials: A Lab Based Approach* (Cambridge University Press). He was a winner of the IEEE Communication Society's Leonard Abraham Prize for 2004.

Robert A. DiFazio [SM] is chief engineer for InterDigital Labs specializing in next generation cellular, millimeter-wave, small cells, machine-to-machine communications, and spectrum sharing. He was previously at BAE Systems working on a variety of radio and navigation systems. He has a Ph.D. from NYU Polytechnic School of Engineering and serves on Industry Advisory Committees for NYU-Polytechnic and the New York Institute of Technology. He holds numerous issued and pending US patents.

Rui Yang received his M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1987 and 1992, respectively. He has 17 years of experience in the research and development of wireless communication systems. Since he joined InterDigital Communications in 2000, he has led several product development and research projects. He is currently a principal engineer at InterDigital Labs. His interests include digital signal processing and air interface design for cellular and WLAN networks. He has received more than 15 patent awards in those areas.

Erdem Bala received his B.S. and M.S. degrees from Bogazici University, Istanbul, Turkey, and his Ph.D. degree from the University of Delaware, all in electrical engineering. He is currently a research engineer at InterDigital Labs conducting research on future radio access networks. Previously, he was an intern at Mitsubishi Research Labs and a software engineer at Nortel Networks. He has numerous issued and pending patents, and has received the InterDigital Innovation Award and the Nortel Networks Pride Award.

Full Duplex Techniques for 5G Networks: Self-Interference Cancellation, Protocol Design, and Relay Selection

Zhongshan Zhang, Xiaomeng Chai, Keping Long, Athanasios V. Vasilakos, and Lajos Hanzo

Zhongshan Zhang, Xiaomeng Chai, and Keping Long are with the University of Science and Technology Beijing.

Athanasios V. Vasilakos is with the University of Western Macedonia.

Lajos Hanzo is with the University of Southampton.

This work was supported by the key project of the National Natural Science Foundation of China (No. 61431001), the 863 project No. 2014AA01A701, Program for New Century Excellent Talents in University (NECT-12-0774), the open research fund of National Mobile Communications Research Laboratory Southeast University (No. 2013D12), Fundamental Research Funds for the Central Universities, and the Foundation of Beijing Engineering and Technology Research Center for Convergence Networks.

ABSTRACT

The wireless research community aspires to conceive full duplex operation by supporting concurrent transmission and reception in a single time/frequency channel for the sake of improving the attainable spectral efficiency by a factor of two as compared to the family of conventional half duplex wireless systems. The main challenge encountered in implementing FD wireless devices is that of finding techniques for mitigating the performance degradation imposed by self-interference. In this article, we investigate the potential FD techniques, including passive suppression, active analog cancellation, and active digital cancellation, and highlight their pros and cons. Furthermore, the troubles of FD medium access control protocol design are discussed for addressing the problems such as the resultant end-to-end delay and network congestion. Additionally, an opportunistic decode-and-forward-based relay selection scheme is analyzed in underlay cognitive networks communicating over independent and identically distributed Rayleigh and Nakagami-*m* fading channels in the context of FD relaying. We demonstrate that the outage probability of multi-relay cooperative communication links can be substantially reduced. Finally, we discuss the challenges imposed by the aforementioned techniques and a range of critical issues associated with practical FD implementations. It is shown that numerous open challenges, such as efficient SI suppression, high-performance FD MAC-layer protocol design, low power consumption, and hybrid FD/HD designs, have to be tackled before successfully implementing FD-based systems.

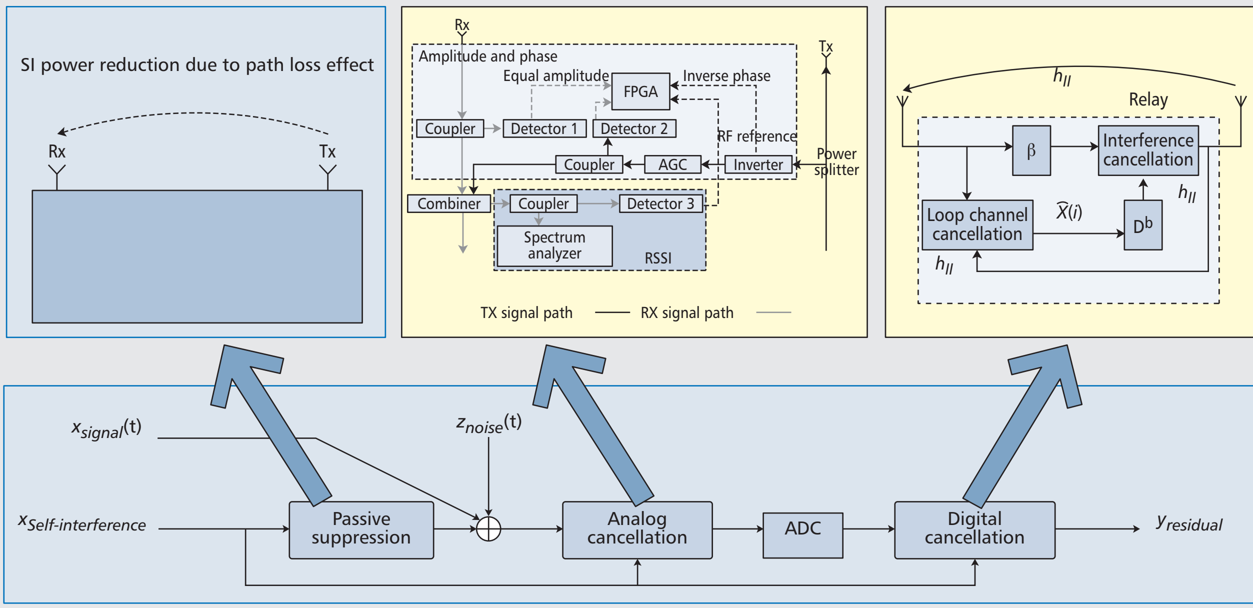
INTRODUCTION

The spectral efficiency (SE) of networks has to be further improved in order to deliver ever increasing data rates. However, the operational wireless communication systems usually rely on half duplex (HD) operations, leading to erosion of resource exploitation. The promise of radical full duplex (FD) operation, on the other hand, improves the achievable SE of wireless communication systems by always transmitting and receiving in the entire bandwidth.

The main driving force behind the advances in FD communications is the promise of nearly doubled channel capacity compared to conventional HD communications, thus offering the potential to complement and sustain the evolution of the fifth generation (5G) technologies toward denser heterogeneous networks with flexible relaying modes [1]. Recently, a range of theoretical and practical aspects of FD communications have been investigated by quantifying the performance gains of FD modes (FDMs) [2], which exhibits advantages over the half-duplex mode (HDM) in terms of either having increased throughput or reduced outage probability (OP), albeit achieved at the cost of increased complexity. Furthermore, recent advances in FD communications have increased both the attainable throughput and the diversity orders of wireless communication systems. Once increased hardware/software complexity is tolerated to facilitate more sophisticated signal processing, it would be possible for an FD device to reduce the bit error rate (BER). In addition, the packet loss ratio (PLR) of FDM may also be reduced, provided that a larger buffer size is provided by FD devices.

However, as a downside, the FD gain is eroded by self-interference (SI) due to the large power difference between the power imposed by a device's own transmissions and the low-power received signal arriving from a remote transmit antenna. Excessive SI may even result in reduced capacity for FD systems that falls below that of HD systems. Consensus reached by both industry and academia show that it is critical to perform efficient SI suppression/cancellation in implementing radical FD communication systems. Apart from the aforementioned physical-layer issues, the conception of FD medium access control (MAC) protocols requires substantial further research. Experience indicates that FD schemes may not always outperform their HD counterparts, and hybrid schemes that switch between HDM and FDM can also be developed for adap-

0163-6804/15/$25.00 © 2015 IEEE



Algorithm	Transmit power	Center frequency	Bandwidth	Antenna distances	Antenna separations	Cancellation capability	Full duplex gain
Compact/separated antennas	5 dBm	2.6 GHz	100 MHz	20 mm		48 dB	
				5 m		70 dB	
Antenna cancellation		2.4 GHz	100 MHz	λ/2		60 dB	1.84
ZigZag		2.4 GHz					1.25
Antenna separation	–5 dBm ~15 dBm	2.4 GHz	625 kHz	20 cm		39 dB	
				40 cm		45 dB	
Directional diversity	12 dBm	2.4 GHz	20 MHz	10 m	≥45°		1.6~1.9
				15 m	≥90°		≥1.4

Figure 1. Practical implementable SI suppression algorithms and their performance comparison.

tively exploiting the radio resources, while at the same time maximizing the SE [3]. Again, an FD scheme may not always outperform its HD counterpart, requiring a hybrid HD/FD scheme to be implemented to gain an advantage over either of the individual schemes. In this article, we survey/compare different FD techniques. Some existing SI cancellation techniques such as passive suppression, active analog, and digital cancellation are discussed. Furthermore, the critical issues associated with FD-based MAC-layer protocols are also studied. Finally, the choice of the optimal relay selection scheme conceived for FDM is elaborated on, followed by a variety of new directions and open problems. The main contributions of this article include
- Surveying the critical issues related to FD transmissions from a physical-layer perspective relying on SI suppression
- Giving cognizance to the MAC-layer protocols
- Proposing an FD-based opportunistic decode-and-forward (DF)-based relay selection scheme in the context of underlay cognitive networks and analyzing the OP of the multi-relay cooperative communication links
- Outlining several challenges associated with FDM-based device/system realizations
- Discussing both the advantages and drawbacks of various FD techniques, while identifying their challenges and new directions

The remainder of this article is organized as follows. The classification of both passive and active SI suppression is detailed next. Typical FD MAC-layer protocols, such as the FD-MAC technique [4], are then discussed, followed by several critical issues related to the associated practical implementation and commercial realizations. We then propose an opportunistic FDM relay selection scheme, followed by a range of open challenges and the future directions of FD communications. Finally, our conclusions are provided.

Self-Interference Cancellation

Existing studies [5, 6] showed that it is critical to accurately measure and suppress the SI in FD communication. For instance, as revealed in [7], the SI power as well as spatial reuse may substantially reduce the FD gain over the HDM in terms of the network-level capacity, rendering it well below 2 in common cases. However, if the SI level at the input of the FD relay (i.e., after performing SI suppression) can be at least 3 dB lower than the noise level, the remaining SI may not seriously degrade the end-to-end throughput

Since the analog SI cancellation methods are never perfect, the residual SI after analog cancellation should be further reduced with the aid of digital cancellation. In the existing digital-cancellation protocols, ZigZag exhibits a significant advantage in terms of the achievable FD gains.

[8]. SI cancellation techniques are usually classified into passive and active suppressions, as shown in Fig. 1.

PASSIVE SI SUPPRESSION

Passive SI suppression is defined as the signal-power attenuation imposed by the path loss due to the physical separation between the transmit and receive antennas of the same device. Typical passive SI suppression techniques include:

Directional SI suppression: In this technique, the main radiation lobes of the transmit/receive antennas of an FD device have minimal intersection, enabling the SI to be partially suppressed prior to the receiver's RF front-end.

Antenna separation and SI cancellation: Increasing the path loss between the transmit/receive antennas constitutes an effective approach to attenuate the SI power, in which method a higher antenna separation implies better SI suppression performance. When relying on antenna separation, the natural isolation may also exploit the surrounding buildings or the beneficial inclusion of a shielding plate, provided that strict restrictions imposed on the device size can be satisfied.

ACTIVE SELF-INTERFERENCE SUPPRESSION

In [9], active SI suppression methods were shown experimentally to be capable of facilitating FD communication at ranges up to 6 m and at transmit powers typical of WiFi devices, revealing that the interference level can be reduced by 50 dB and 40 dB under static and dynamically fading interference channel scenarios, respectively, if an RF SI canceller is combined with a baseband canceller. The family of active suppression techniques can be subdivided into analog cancellation, digital cancellation, and combined analog/digital cancellation, as discussed below.

Analog Cancellation — In analog cancellation, the family of time-domain (TD) cancellation algorithms such as training-based methods can be employed by both single-input single-output (SISO) and multiple-input multiple-output (MIMO) based techniques, where the latter may perform SI suppression by exploiting the spatial diversity achieved by the associated multiple transmit and/or receive antennas.

- Classic TD training-based methods can be beneficially utilized for estimating the SI leakage, while facilitating reliable SI cancellation. Asymmetric complex signals, in which the inputs are chosen to be complex but not circularly symmetric, can also be utilized for mitigating the SI in single-antenna-aided FDM relays under DF relaying. The optimum SI cancellation weight vectors can be exploited by increasing the signal-to-noise ratio (SNR) of the source → relay and relay → destination links, thus beneficially improving the attainable throughput of FD relaying channels.
- The increased degree of freedom (DoF) offered by the spatial domain (SD) antenna arrays of MIMO systems may be utilized to provide a range of new solutions for SI cancellation. In MIMO aided FD systems, relays are capable of operating in either the antenna-partitioning-based mode (i.e., all antennas operating in the FDM but partitioned into transmit and receive antenna sets) or antenna-sharing-based mode (i.e., allowing antennas to be utilized more efficiently by exploiting the increased dimensions of MIMO channels and/or by relying on time-division duplexing, TDD, aided channel reciprocity).

Digital Cancellation — Since analog SI cancellation methods are never perfect, the residual SI after analog cancellation should be further reduced with the aid of digital cancellation. Of the existing digital cancellation protocols, ZigZag [10] exhibits a significant advantage in terms of the achievable FD gains. Note that ZigZag imposes no change on the conventional IEEE 802.11 MAC protocols when there is no collision, thus maintaining the same throughput as if the colliding packets were scheduled a priori in separate time slots in the presence of transmission collisions. It has been observed that 10 percent of the transmitter-receiver pairs of a wireless network often experience severe packet loss due to packet collisions imposed by statistical channel multiplexing. The asynchronous nature of successive collisions can be successfully exploited in ZigZag to address the problem of high packet loss rate (PLR). By using ZigZag, the average PLR at hidden terminals was shown to be reduced from 72.6 to about 0.7 percent, while improving the average throughput by 25.2 percent compared to the conventional IEEE 802.11 standards.

Performance Comparison — The SI suppression capabilities of some typical algorithms are characterized in Fig. 1. Although numerous sophisticated techniques have been proposed for performing SI cancellation in FD devices, both advantages and disadvantages are exhibited in the context of each approach, as shown in Table 1.

OPEN RESEARCH ISSUES IN SI SUPPRESSION

Although passive SI suppression techniques are capable of attenuating the SI power in proportion to the path loss, enabling a higher antenna separation usually requires a larger or even infeasible device size. More detrimentally, increasing the antenna separation implies a degradation of the SI channel estimation. Furthermore, numerous additional challenges have to be addressed in the context of the existing active SI suppression techniques. For instance, the achievable SI cancellation capability may be limited by relying on standalone analog or digital cancellation. It is thus rather critical to effectively balance the roles of the analog- and digital-domain functions in the overall SI cancellation, carefully revealing the overall benefits of combined analog/digital cancellation. In the following, a number of possible solutions to the above-mentioned challenges should be proposed.

Antenna configuration for practical size-limited FD devices: In passive SI suppression schemes, the best antenna configuration in terms of the attainable SI suppression can be achieved upon installing the transmit and receive antennas at the opposite sides of the device to create sufficient separation, requiring the device size to be large enough.

Combination of active and passive SI sup-

Category		Algorithm	$T_x \times R_x$	Advantage	Disadvantage
Passive suppression		Directional diversity Antenna separation		1) SI attenuated due to path loss 2) Decreases inter-device interference 3) Improves power efficiency 4) More separation implies a better attenuation of SI signal	1) Performance depends highly on AS and beam pattern 2) AS is restricted by variant factors such as device size and interference channel estimation accuracy 3) Restricted applications to SISO
Active suppression	Analog cancellation	Antenna cancellation	2×1	1) Easy to implement 2) High cancellation capability 3) Robust in narrowband systems	1) Broadband-induced loss 2) Degrades the received signal 3) Limited transmit power 4) Requires fixed AS
		Pre-nulling	$M \times 1$	1) Simple to implement 2) No influence on receiver BER 3) Stringent requirements on antenna isolation are required	1) SI channel estimation is required 2) Designed specifically for flat-fading channels
		AFC	$1 \times M$	1) Low complexity 2) Needs no training sequence 3) No delay insertion in the relay 4) Compensates for multipath propagation	The second-order statistical information of the source signal is required to be exploited by the filter design
		Pre-coding/decoding	$M \times M$	1) Better than pre-nulling 2) Enables advanced optimization 3) Capacity optimization	1) Requires SI estimation 2) Requires SVD of SI channel matrix
		Block diagonization	$M \times M$	1) Outpeforms ZF beamforming 2) Precoding with adaptive power allocation to optimize the sum rate	1) CSI is required by the base station 2) SVD is required 3) Power allocation satisfies KKT conditions
		ZF filters	$M \times M$	1) High capacity for a high SNR 2) Multiple spatial streams are supported in the MIMO relay	1) Perfomance loss in low-SNR regions 2) SVD is required
		Optimal Eigenbeam-forming	$M \times M$	Power of the residual SI is minimized	1) Beam selection matrices are calculated 2) SVD is required
		Maximum SIR	$M \times M$	1) Improves the useful signal 2) Suppresses both SI and noise	1) High complexity in deriving the optimum matrices 2) Channel attenuation highly impacts the performace
		MMSE filtering	$M \times M$	1) Improves the useful signal 2) Suppresses both SI and noise	High complexity
		TAS	$M \times M$	1) Has a low complexity 2) Avoids losses in low-SNR regions 3) Adaptivity to varying SIRs	1) High-dimensional MIMO complicates the best subset selection 2) Unique solution for the best subset selection is not always achievable
	Digital cancellation			1) Residual SI after analog cancellation can be eliminated in digital domain 2) Modulation independence 3) Addresses hidden terminal problem 4) High collision-combating capability	1) Quantization noise cannot be reduced 2) Becomes unneccessary if preceded by a powerful analog cancellation 3) Limited cancellation capability

Table 1. Performance comparison among variant SI suppression algorithms.

Theoretically, an FD system having an infinite dynamic range and perfect channel estimation can perfectly eliminate the SI signal. However, the hardware limitations, including transmit/receive signal quantization, non-linearities, in-phase and quadrature mismatch, etc. all might erode the practical implementations of FD systems.

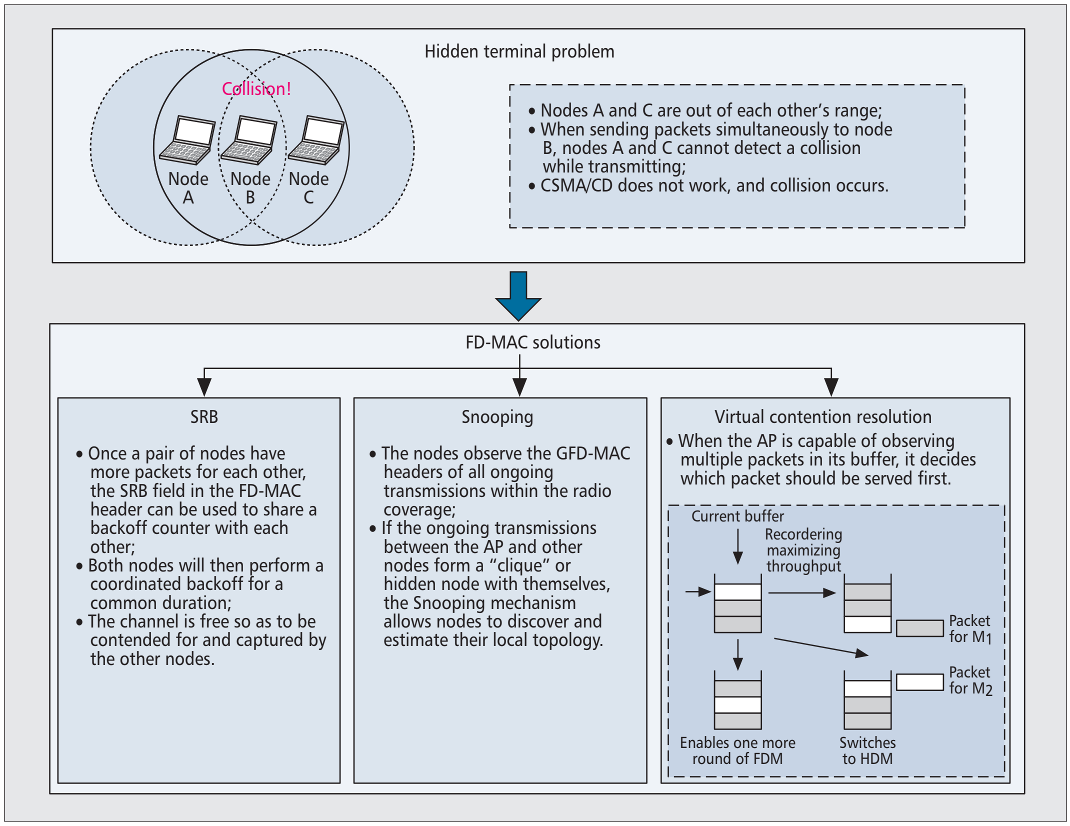


Figure 2. Mechanisms of FD-MAC protocol [4] — shared random backoff (SRB), snooping, and virtual contention resolution — can be employed for addressing the problem of the hidden terminal.

pressions: Since none of the individual cancellation techniques is capable of satisfying the system requirements in terms of the attainable SI cancellation capability, a high-capability cancellation scheme by combining the active and passive methods is necessarily developed.

Low-complexity spatial-domain suppression approaches: Many of the existing spatial domain SI suppression methods relying on complex matrix computations may significantly erode the FD gains owing to their infeasibility. Therefore, low-complexity algorithms conceived for high-dimensional MIMO channels are capable of dramatically improving the SI cancellation capability at a reasonable hardware/software cost.

Transmit power control for improving SI suppression: A higher transmit power will definitely lead to a lower SI channel estimation error, but the absolute level of the residual SI power may still increase for a high SI power; however, the ratio between the residual error and the overall SI might be reduced.

MAC Layer Protocol Design For Full Duplex Systems

Apart from the aforementioned physical-layer solutions, FD research opportunities have also been explored in the context of efficient MAC protocols for addressing the challenges of long end-to-end delays of network congestion and the hidden terminal problems. For instance, in [4], a new MAC protocol referred to as FD-MAC was developed and implemented for infrastructure-based WiFi-like networks to provide opportunities for all the accessed nodes while trying to maximize the overall network throughput and maintaining fairness to all users simultaneously. In order to satisfy the above-mentioned requirements, three mechanisms, shared random backoff (SRB), snooping, and virtual contention resolution, can be employed, as illustrated in Fig. 2. FD-MAC is capable of guaranteeing seamless wireless access while maximizing the FD gains. Experimental results showed that FD-MAC achieves a throughput gain of up to 70 percent over its comparable HD counterpart [4].

FD Realizations in Practical Systems

Although very few FD realizations have been implemented in commercial systems due to the technical and/or economic challenges, a substantial amount of related research has already been undertaken by addressing several challenges in this context, discussed below.

Hardware Limitations

In [11], the performance of co-channel FDM-based MIMO nodes was analyzed in the context of modeling their realistic hardware characteristics. Theoretically, an FD system having an infinite dynamic range and perfect channel estimation can perfectly eliminate the SI signal. However, the hardware limitations, including transmit/receive signal quantization, nonlinearities, in-phase and quadrature (I/Q) mismatch, and so on, all might erode the practical implementations of FD systems.

RECEIVER COMBINING

Apart from the impairments imposed by SI signals and the above-mentioned hardware limitations, another challenge comes from the fact that FD-based systems might not be capable of invoking some sophisticated combining schemes such as maximum ratio combining (MRC) unless the source node and the FD-based relay are perfectly phase-synchronized. In order to address this challenge, a co-phasing scheme can be employed in the direct and relay links, facilitating a significant coherent combining gain at the destination.

HYBRID HD/FD RELAYING

Note that FDM may not necessarily always outperform HDM in terms of throughput or channel outage probability, particularly when the FD devices suffer from high residual SI power. A hybrid HD/FD scheme, which facilitates switching between HDM and FDM, may thus be expected to outperform either HDM or FDM alone.

Scheduling for hybrid schemes: In [12], a time-domain scheduling scheme was proposed for performing a hybrid of full and half duplex relaying (FHDR) while formulating the objective function as a nonlinear programming problem. The solution of hybrid FHDR can be analytically derived by solving the above-mentioned nonlinear programming problem. Furthermore, proportional fairness in terms of all the users' end-to-end throughput can be achieved in hybrid FHDR. As compared to an equal opportunity scheduling scheme, hybrid FHDR is capable of achieving a superior performance in terms of its sum rate without sacrificing fairness among users.

Opportunistic hybrid scheme: Opportunistic duplex-mode resource allocation is motivated by resolving the fundamental trade-off between the achievable SE and the attained SI suppression capability. Explicit conditions, under which a specific duplex mode is preferred over the other, can be provided [3], enabling opportunistic hybrid FD/HD relaying to offer significant performance gains over the conventional system design that is confined to either of its constituent modes. Furthermore, the benefits of the trade-off between the FDMs and HDMs depend heavily on the employment of transmit-power adaptation, potentially making FDMs more attractive.

Hybrid schemes in cognitive radio networks: A significant performance improvement can be attained in cognitive radio networks by developing a hybrid FDM/HDM scheme based on the classic zero-forcing criterion, provided that the multiple-antenna-based secondary transmitters have FD capabilities. The hybrid scheme has been shown in [13] to achieve almost three times the cognitive user rates provided by the HDM with the aid of the same RF chains.

FULL DUPLEX RELAY SELECTION

Cooperative relaying has been identified as a promising solution for effectively combating the shadowing effects to extend the radio coverage and significantly improve the channel capacity simultaneously [14]. Numerous relaying protocols, such as amplify-and-forward (AF), DF, and compress-and-forward (CF), can be employed for efficient relaying as a means of guarding against severe signal fading. Theoretically, the more relays the cooperative communication systems are equipped with, the higher the DoF provided by the relaying channels, hence promising improved performance quantified in terms of channel capacity and/or link reliability. In a multi-relay-aided cooperative communication system, activating more relays tends to attain a better DoF, because the system becomes capable of combining a higher number of independently fading signals associated with multiple relays.

However, usually orthogonal channels created in terms of carrier frequencies, time slots, or spreading codes among relays are allocated in multi-relay systems in order to mitigate the inter-relay interference, consequently eroding the increased DoF benefits due to their increased spectrum demand. In order to mitigate the above-mentioned penalty, the method of relay selection relying on channel state information (CSI) feedback has been regarded as one of the most promising solutions. The optimal relay selection scheme, in which the specific candidate relay having the "best" channel[1] is activated, while deactivating the other relays, is shown to be an ideal way of optimizing the diversity order in a cost-efficient manner.

Relay selection techniques invoked in HDM-based systems have been widely studied, where the attainable benefits accrue from the fact that the system usually activates relays roaming about halfway between the source and destination. As distinguished from conventional HDM-based relay selection, in which time domain orthogonal channels must be allocated to the source → relay and relay → destination phases, FDM-based relay selection algorithms may provide a higher performance gain in terms of their outage probability and/or channel capacity due to their essential capability of concurrently transmitting and receiving in a single time/frequency slot [15]. **In order to minimize the negative impact of the SI signal on the performance of FDM-based systems and optimize the signal-to-interference-plus-noise ratio (SINR) of the source → relay → destination link simultaneously, the relay having the lowest SI power among all the candidate relays can be activated.**

However, FDM-based relay-selection policies have not been widely explored, let alone evaluating their impact on the achievable system performance. Furthermore, the channel capacity of FDM-based relay selection schemes may be significantly eroded by SI power.

In this section, opportunistic DF-based relay selection schemes in underlay cognitive networks communicating over independent and identically distributed (i.i.d.) Rayleigh and Nakagami-*m* fading channels are considered in the context of FDM relaying. The principle of the proposed FDM relay selection scheme is described in Fig. 3.

Relay selection under Nakagami-*m* channels: When $N = 4$ and $\lambda = 10$, as shown in Fig. 4, a smaller $\overline{\gamma}_{LI}$ implies a reduction in residual SI power. The OP thus becomes a monotonically increasing function of $\overline{\gamma}_{LI}$, where $\overline{\gamma}_{LI} = 0$ corre-

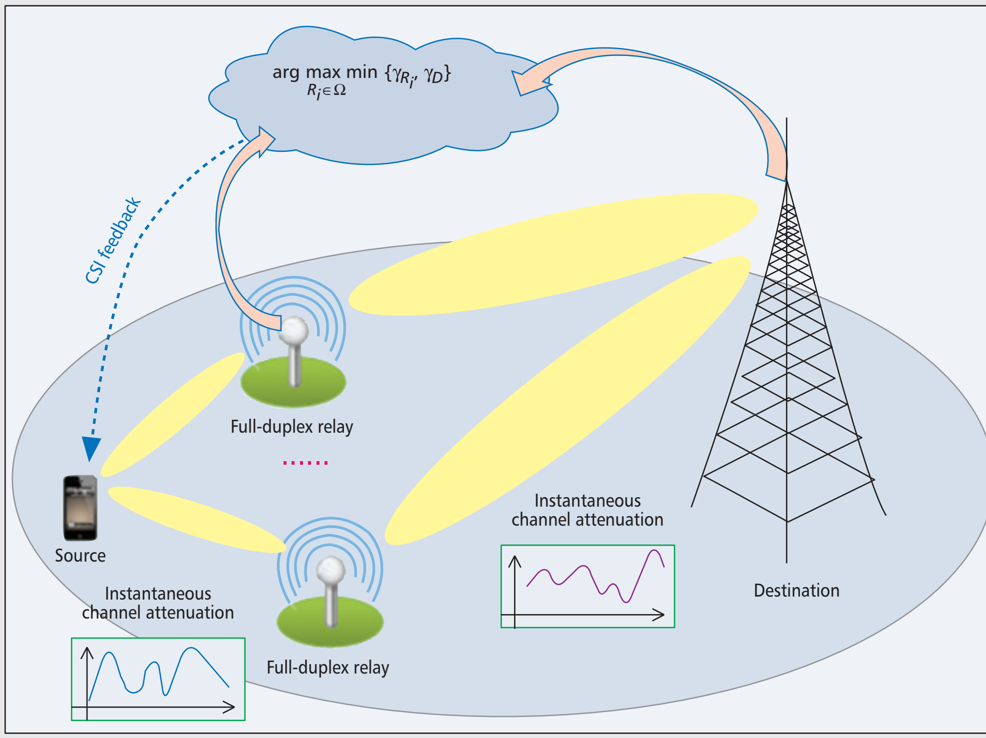
FDM may not necessarily always outperform HDM in terms of throughput or channel outage probability, particularly when the FD devices suffer from a high residual SI power. A hybrid HD/FD scheme, which facilitates switching between HDM and FDM, may thus be expected to outperform either HDM or FDM alone.

[1] Note that in cooperative relaying systems, the "best channel" can be defined in terms of the quality of the source → relay link, relay → destination link, or even the concatenated source → relay → destination link, subject to the practical CSI feedback.

Although FD techniques are capable of significantly improving both the achievable SE and the network throughput compared to the classic HD approach, both efficient SI suppression and FD-based MAC-layer protocols are highly required. Numerous open challenges are still to be tackled before successfully implementing FD devices.

sponds to perfect SI cancellation at the relay. Both passive SI suppression and active analog/digital domain SI cancellations can be invoked to reduce the residual SI power. However, for all realistic scenarios with $\overline{\gamma}_{LI} > 0$, the SI cancellation would be imperfect, resulting in non-zero residual SI power in the FDM devices. Compared to FDM, HDM is capable of reducing the interference imposed on the primary users, especially when the SI level is higher. For instance, HDM could outperform the FDM when $\overline{\gamma}_{SR} > 15$ dB.

Relay selection under Rayleigh fading channel: When $N = 4$, $\lambda = 10$, and $\overline{\gamma}_{LI} = 5$ dB, we have assumed that all channel coefficients suffer from i.i.d. Nakagami-m fading. The OP performance of the proposed relay selection scheme is shown in Fig. 5. Note that the Nakagami fading parameters $M = (m_{SP}, m_{RP}, mSR, mRD, m_{LI})$ of the different cases significantly impact the attainable diversity order of the cooperative network. Furthermore, the outage performance will be more severely impacted by m_{SR}. As illustrated in Fig. 5, the optimum operating SNR of the pro-



Preliminary settings:							
S	Source node	R_i	the i-th relay	D	Destination	γ_{R_i}	SINR of R_i
γ_{LI}	Variance of residual loop-interference/SI-to-noise ratio	N	Number of candidate relays	γ_D	SINR of D	P_R	Power allocated to relays
h_{SP}	Interference imposed on primary users by S	h_{R_iP}	Interference imposed on primary users by R_i	P_S	Power allocated to S	λ	Preset interference-tolerant threshold in primary users
DF protocol is performed in FDM based relays				$S \rightarrow D$ link suffers from deep fading and is not activated			

Algorithm:

1 Calculate the average γ_{LI}, i.e. $\overline{\gamma_{LI}}$; Calculate γ_{R_i} and γ_D based on $\overline{\gamma_{LI}}$; Feedback γ_{Ri}, and γ_D to S;

2 Determine the candidate relay set $\Omega = \{R_i | P_S \bullet |h_{SP}|^2 + P_R \bullet |h_{R_iP}|^2 \leq \lambda,\ 1 \leq i \leq N\}$;

3 Determine the best relay subject to $\arg \max_{Ri \in \Omega} \min \{\gamma_{R_i}, \gamma_D\}$; Making the other relays keep silent;

4 If INDIVIDUAL POWER CONSTRAINS

5 Calculate P_S and P_R separately in the Control Unit

6 Else If SUM POWER CONSTRAINS

7 Calculate (P_S + P_R) in the Control Unit

8 End Else If

9 End If

10 Allocate P_S and P_R; Perform Signal Transmission and Forward;

11 Return

Figure 3. Opportunistic FDM-based relay selection in underlay cognitive networks communicating over i.i.d. fading channels.

posed opportunistic relay selection scheme will always be attainable in the range of (10 dB, 15 dB), while the attainable diversity order of underlay-based cognitive networks may be severely impacted by the fading parameter of the source $\rightarrow$ relay link.

Remaining Challenges snd Potential Future Research

Although FD techniques are capable of significantly improving both the achievable SE and the network throughput compared to the classic HD approach, both efficient SI suppression and FD-based MAC-layer protocols are highly required. Numerous open challenges are still to be tackled before successfully implementing FD devices. In this section, general design guidelines for FD wireless communication systems are offered based on the aforementioned discussions.

Remaining Challenges

From the discussion above, some open challenges associated with FD technology have to be tackled.

FD-based device complexity issues: Carrying out powerful SI cancellation increases both the cost and complexity of FD-based devices, mainly because complex matrix computations have to be performed at the transceiver. Furthermore, the hardware limitations will also constrain the performance gain of FDM.

FD-based MAC-layer protocol design: Apart from the physical-layer solutions discussed above, a properly designed FD MAC-layer protocol, which should be backward-compatible with the existing HD-based MAC-layer protocols, is highly required for avoiding problems such as hidden terminal in multihop networks. Furthermore, the FD-based MAC-layer protocol should not unduly favor FD opportunities over HD flows, which requires the access mechanism to be capable of providing a fair opportunity for all nodes to access the shared medium.

Low energy consumption issues: Since most wireless terminals are battery-driven and have limited energy harvesting capabilities, the energy dissipation of FD-based MAC-layer protocols remains a challenging issue. It is of great importance to develop cost-efficient FD-based MAC-layer protocols with low energy consumption in order to extend both the devices' battery recharge time and the network's overall survivability.

FDM in the high-SNR/data rate regime: The FDM philosophy was shown to outperform HDM in terms of capacity gain, link robustness, and/or outage probability, provided that the former operates at low to medium SNR values and information rates. Hence, expanding the benefits of FDM to the high-SNR/data rate regime is promising but challenging in practical environments.

Future Research

It is worth pointing out that some of the approaches presented in this article may be further developed, as detailed below.

FDM under a wider bandwidth with a higher transmit power: The feasibility of FD technologies in systems of wider bandwidth with higher transmit power has to be further improved with the aid of improved SI cancellation capability, despite current techniques that can be effectively utilized in systems having relatively narrow bandwidth and low transmit power (e.g., IEEE 802.15.4).

Cost-efficient spatial-domain SI suppression: Complex matrix computations are usually required in many existing spatial-domain SI suppression methods with a complexity burden that significantly hampers the realizability of FD systems. Therefore, more cost-efficient spatial domain SI suppression algorithms have to be designed specifically for MIMO channels.

FDM-based MAC-layer protocols: Many critical issues, such as the problems of hidden terminals, and multiple access collisions of distributed techniques, the requirements of low power consumption, and maintaining backward-compatibility with existing MAC protocols, cannot be readily addressed in the context of FD-based MAC-layer protocols. Thus, an appropriate MAC-layer protocol conceived for fully exploiting the FDM benefits is definitely worthy of further study.

Practical implementation of a hybrid FD/HD scheme: Although a hybrid scheme facilitating dynamic switching between HDM and FDM is capable of outperforming either of its constituent modes, the hybrid-mode-based devices have to be capable of identifying the CSI changes and promptly switching between these two modes. In the absence of a centralized controller, a sophisticated distributed approach relying on self-organization principles [16] could be employed by devices to implement a cost-efficient hybrid protocol.

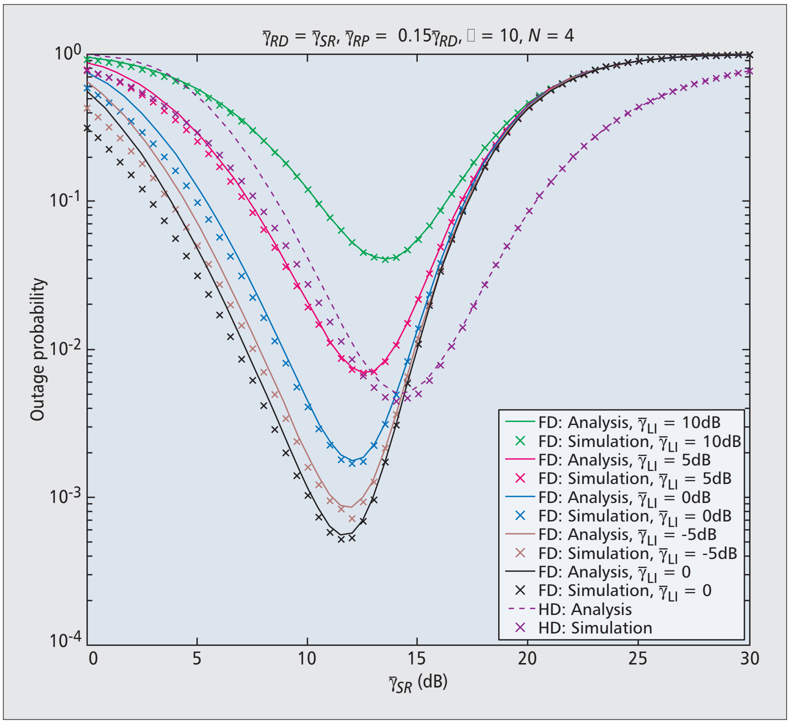


Figure 4. Outage probability vs. average SNR of the $S \rightarrow R$ links in conjunction with $\lambda = 10$ and $N = 4$.

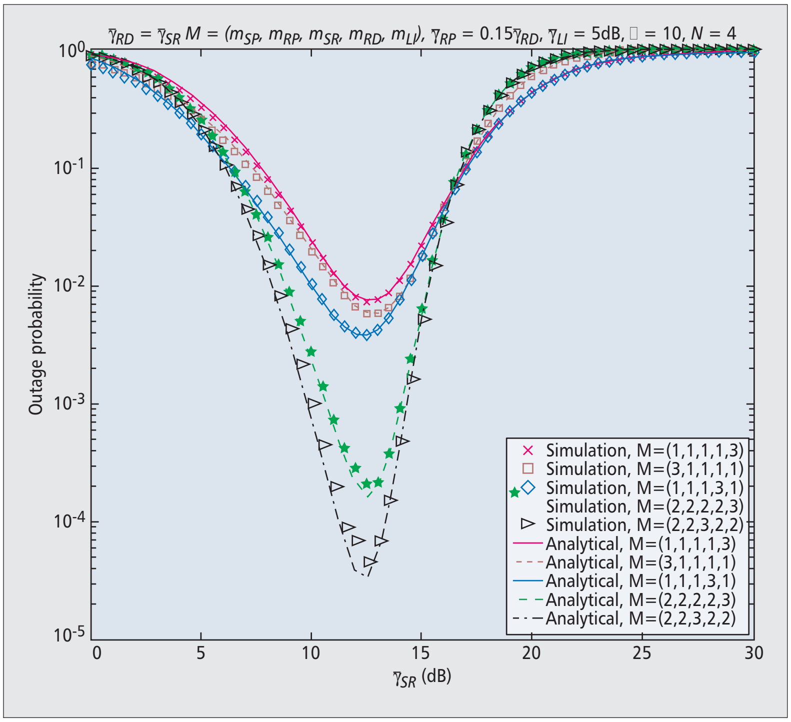


Figure 5. Outage probability vs. the average SNR of $S \rightarrow R$ links, where $\lambda = 10$, $\bar{\gamma}_{LI} = 5$ dB, and $N = 4$. In each simulation, the solid curves are used to stand for the analytical results, and the markers without lines denote the simulation results.

Buffer size vs. PLR/delay trade-off: Since an FDM-based device has to process twice as many packets as an HDM-based device due to its essential capability of concurrently transmitting and receiving in a single time/frequency slot, both the PLR and the delay may become more severe for FDM than for HDM unless the buffer's queue length in the former is significantly increased. Nevertheless, striking the appropriate buffer size vs. PLR/delay trade-off constitutes a promising study item.

Conclusions

Since the throughput requirements cannot be readily satisfied without increasing the achievable SE expressed in bits per second per Hertz, FD technology has been proposed with the promise of nearly doubling the data rate in comparison to its HD counterpart. An FDM-based device potentially facilitates simultaneous transmission and reception within the same frequency band. One of the main challenges in implementing FD communications comes from the performance erosion induced by the SI power, which has to be suppressed/cancelled to a tolerable level. However, the family of existed SI suppression/cancellation solutions is typically based on costly hardware design and/or complex matrix computations, cost-efficient algorithms associated with low complexity are highly required for improving the realizability of practical HDM based devices. Apart from the physical-layer issues, there is also an urgent demand for high-performance low-complexity FD protocols, requiring the impact of the MAC/higher-layer protocols on the practical implementation of FDM-based systems to be investigated more vigorously. Last but not least, FDM-based relay selection will also play a critical role in optimizing the performance gain of multi-relay cooperative communication systems.

References

[1] S. Hong *et al.*, "Applications of Self-Interference Cancellation in 5G and Beyond," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 114–21.

[2] T. Riihonen, S. Werner, and R. Wichman, "Optimized Gain Control for Single-Frequency Relaying with Loop Interference," *IEEE Trans. Wireless Commun.*, vol. 8, no. 6, June 2009, pp. 2801–06.

[3] T. Riihonen, S. Werner, and R. Wichman, "Hybrid Full-Duplex/Half-Duplex Relaying with Transmit Power Adaptation," *IEEE Trans. Wireless Commun.*, vol. 10, no. 9, Sept. 2011, pp. 3074–85.

[4] A. Sahai, G. Patel, and A. Sabharwal, "Pushing the Limits of Full-Duplex: Design and Real-Time Implementation," arXiv:1107.0607, 2011.

[5] A. Sabharwal *et al.*, "In-Band Full-Duplex Wireless: Challenges and Opportunities," arXiv:1311.0456, 2014.

[6] D. Bharadia, E. Mcmilin, and S. Katti, "Full Duplex Radios," *Proc. ACM SIGCOMM*, Hong Kong, 2013, pp. 375–86.

[7] X. Xie and X. Zhang, "Does Full-Duplex Double the Capacity of Wireless Networks?," *Proc. IEEE INFOCOM '14*, May 2014, pp. 253–61.

[8] T. Riihonen *et al.*, "Outage Probabilities in Infrastructure-Based Single-Frequency Relay Links," *IEEE WCNC*, Budapest, Hungary, Apr. 2009.

[9] E. Everett *et al.*, "Empowering Full-Duplex Wireless Communication by Exploiting Directional Diversity," *Asilomar Conf. Signals, Systems and Computers*, 2011, pp. 2002–06.

[10] Y.-C. Cheng *et al.*, "Jigsaw: Solving the Puzzle of Enterprise 802.11 Analysis," *Proc. ACM SIGCOMM*, 2006.

[11] D. W. Bliss, T. M. Hancock, and P. Schniter, "Hardware Phenomenological Effects on Cochannel Full-Duplex MIMO Relay Performance," *2012 Conf. Record 46th ASILOMAR*, Nov. 2012, pp. 34–39.

[12] M. Miyagoshi *et al.*, "Multi-User Transmission Scheduling for A Hybrid of Full-and Half-Duplex Relaying," *2011 8th Int'l. Conf. Info., Commun. and Signal Processing*, 2011, pp. 1–5.

[13] K.-K. Wong, "Maximizing the Sum-Rate and Minimizing the Sum-Power of A Broadcast 2-User 2-Input Multiple-Output Antenna System Using A Generalized Zeroforcing Approach," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, Dec. 2006, pp. 3406–12.

[14] Z. Zhang, W. Zhang, and C. Tellambura, "Cooperative OFDM Channel Estimation in the Presence of Frequency Offsets," *IEEE Trans. Vehic. Tech.*, vol. 58, no. 7, Sept. 2009, pp. 3447–59.

[15] H. Cui, L. Song, and B. Jiao, "Relay Selection for Two-Way Full Duplex Relay Networks with Amplify-and-Forward Protocol," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, July 2014, pp. 3768–3876.

[16] Z. Zhang *et al.*, "On Swarm Intelligence Inspired Self-Organized Networking: Its Bionic Mechanisms, Designing Principles and Optimization Approaches," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 1, 1st qtr. 2014, pp. 513–37.

Biographies

Zhongshan Zhang (zhangzs@ustb.edu.cn) received his B.E. and M.S. degrees in computer science from the Beijing University of Posts and Telecommunications (BUPT) in 1998 and 2001, respectively, and his Ph.D. degree in electrical engineering in 2004 from BUPT. In August 2004 he joined DoCoMo Beijing Laboratories as an associate researcher, and was promoted to researcher in December 2005. In February 2006, he joined the University of Alberta, Edmonton, Canada, as a postdoctoral fellow. In April 2009, he joined the Department of Research and Innovation (R&I), Alcatel-Lucent, Shanghai, as a research scientist. From August 2010 to July 2011, he worked at NEC China Laboratories as a senior researcher. He has served or is serving as a Guest Editor and/or an Editor for several technical journals, such as *IEEE Communications Magazine* and KSII Transactions on Internet AND Information Systems. He is currently a professor with the School of Computer and Communication Engineering at the University of Science and Technology Beijing (USTB). His main research

interests include statistical signal processing, self-organized networking, cognitive radio, and cooperative communications.

XIAOMENG CHAI received his B.Sc. degree in communication engineering from USTB in 2012. He is currently working toward his Ph.D degree with the Institute of Advanced Network Technologies and New Services, USTB. His current research interests include heterogeneous networks, self-organization networks, and resource allocation optimization.

KEPING LONG [SM] (longkeping@ustb.edu.cn) received his M.S. and Ph.D. degrees at UESTC in 1995 and 1998, respectively. From September 1998 to August 2000, he worked as a postdoctoral research fellow at the National Laboratory of Switching Technology and Telecommunication Networks of BUPT. From September 2000 to June 2001, he worked as an associate professor at BUPT. From July 2001 to November 2002, he was a research fellow with the ARC Special Research Centre for Ultra Broadband Information Networks at the University of Melbourne, Australia. He is now a professor and dean at the School of Computer and Communication Engineering, USTB. He is a member of the Editorial Committee of *Science China Series F* and *China Communications*. He has also been a TPC or ISC member for COIN 2003/04/05/06/07/08/09/10, IEEE IWCN 2010, ICON 2004/06, APOC 2004/06/08, Co-Chair of the Organization Committee for IWCMC 2006, TPC Chair of COIN 2005/08, and TPC Co-Chair of COIN 2008/10. He was awarded the National Science Fund for Distinguished Young Scholars of China in 2007 and selected as the Chang Jiang Scholars Program Professor of China in 2008. His research interests are optical Internet technology, new generation network technology, wireless information networks, and value-added service and secure technology of networks. He has published over 200 papers, 20 keynote speeches, and invited talks at international conferences and local conferences.

ATHANASIOS V. VASILAKOS (vasilako@ath.forthnet.gr) is currently a professor at the University of Western Macedonia, Greece. He has authored or co-authored over 200 technical papers in major international journals and conferences. He is author/co-author of five books and 20 book chapters in the area of communications. He has served as General Chair/Technical Program Committee Chair for many international conferences. He has served or is serving as an Editor and/or Guest Editor for many technical journals, such as *IEEE Transactions on Network and Service Management*, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *IEEE Transactions on Information Technology in Biomedicine*, *IEEE Transactions on Computers*, *ACM Transactions on Autonomous and Adaptive Systems*, *IEEE JSAC* Special Issues of May 2009, January 2011, and March 2011, *IEEE Communications Magazine*, *ACM/Springer Wireless Networks*, and *ACM/Springer Mobile Networks and Applications*. He is founding Editor-in-Chief of the *International Journal of Adaptive and Autonomous Communications Systems* and the . He is General Chair of the Council of Computing of the European Alliances for Innovation.

LAJOS HANZO (lh@ecs.soton.ac.uk) [F], FREng, FIET, Fellow of EURASIP, and D.Sc., received his degree in electronics in 1976 and his doctorate in 1983. In 2009 he was awarded an honorary doctorate "Doctor Honoris Causa" by the Technical University of Budapest. During his 37-year career in telecommunications he has held various research and academic posts in Hungary, Germany, and the United Kingdom. Since 1986 he has been with the School of Electronics and Computer Science, University of Southampton, United Kingdom, where he holds the Chair in telecommunications. He has successfully supervised 80+ Ph.D. students, co-authored 20 John Wiley/IEEE Press books on mobile radio communications totalling in excess of 10,000 pages, published 1400+ research entries at IEEE Xplore, acted as both TPC and General Chair of IEEE conferences, presented keynote lectures, and been awarded a number of distinctions. Currently he is directing a 100-strong academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry, the Engineering and Physical Sciences Research Council (EPSRC) UK, the European Research Council's Advanced Fellow Grant, and the Royal Society's Wolfson Research Merit Award. He is an enthusiastic supporter of industrial and academic liaison and offers a range of industrial courses. He is also a Governor of the IEEE VTS. During 2008–2012 he was Editor-in-Chief of IEEE Press and also a Chaired Professor at Tsinghua University, Beijing. His research is funded by the European Research Council's Senior Research Fellow Grant. For further information on research in progress and associated publications, please refer to http://www-mobile.ecs.soton.ac.uk. He has 19,000+ citations.

Full Duplex Cognitive Radio: A New Design Paradigm for Enhancing Spectrum Usage

Yun Liao, Lingyang Song, Zhu Han, and Yonghui Li

Yun Liao and Lingyang Song are with Peking University.

Zhu Han is with the University of Houston.

Yonghui Li is with the University of Sydney.

ABSTRACT

With the rapid growth of demand for ever increasing data rates, spectrum resources have become more and more scarce. As a promising technique to increase the efficiency of the spectrum utilization, cognitive radio has great potential to meet such a requirement by allowing unlicensed users to coexist in licensed bands. In conventional CR systems, spectrum sensing is performed at the beginning of each time slot before data transmission. Unfortunately, this results in two major problems: transmission time reduction due to sensing, and sensing accuracy impairment due to data transmission. To tackle these problems, in this article we present a new design paradigm for future CR by exploring full duplex techniques to achieve simultaneous spectrum sensing and data transmission. With FD radios equipped at secondary users (SUs), the SUs can simultaneously sense and access the vacant spectrum, and thus significantly improve sensing performance while increasing data transmission efficiency. The aim of this article is to transform the promising conceptual framework into a practical wireless network design by addressing a diverse set of challenges such as protocol design and theoretical analysis. Several application scenarios with FD-enabled CR are elaborated, and key open research directions and novel algorithms in these systems are discussed.

INTRODUCTION

Existing and new wireless technologies, such as smartphones, wireless computers, and WiFi home and business networks, are rapidly consuming radio spectrum. Unlike the wired Internet, the wireless world has a limited amount of links to distribute. Consequently, the traditional regulation of spectrum requires fundamental reform in order to allow for more efficient use of the precious resource of the airwaves [1]. Cognitive radio (CR) has been widely recognized as a promising technique to increase the efficiency of spectrum utilization [2]. It allows unlicensed secondary users (SUs) to coexist with primary users (PUs) in licensed bands. There are basically two kinds of CR networks (CRNs): underlay and overlay. In underlay CRNs, SUs transmit at a limited power such that they do not cause harmful interference to the primary network. The restricted power limits the transmit range of SUs. In overlay mode, SUs need to accurately sense the transmission of PUs' spectrum, identify unused spectral holes in which to transmit, and leave it when the incumbent radio system is ready to transmit [3]. In this article, we concentrate on overlay CRNs.

Most existing CRNs deploy half duplex (HD) radios to transmit and receive signals in two orthogonal channels, and SUs employ the well-known Listen-before-Talk (LBT) protocol, in which SUs sense the target channel before transmission [3]. Although proven to be effective, the LBT protocol actually dissipates the precious resources by employing time-division duplexing, and thus unavoidably suffers from two major problems:

- The SUs have to sacrifice transmission time for spectrum sensing, and even if the spectrum hole is long and continuous, the data transmission needs to be split into small discontinuous slots.
- During SUs' transmission, SUs cannot detect changes in PUs' states, which leads to collision when PUs arrive and spectrum waste when PUs leave.

Hence, it would be desirable for SUs to continuously sense spectrum and meanwhile transmit whenever spectrum holes are detected. This, however, seems impossible with the conventional HD systems. A full duplex (FD) system, where a node can send and receive signals at the same time and frequency resources, offers the potential to achieve simultaneous sensing and transmission in CRNs. However, for many years, it was considered impractical [4] because signal leakage from local output to input, referred to as self-interference, may overwhelm a receiver, thus making it impossible to extract the desired signal. However, the recent significant progress in self-interference cancellation has presented great promise for realizing FD communications for future wireless networks [4].

In this article, we explore FD techniques in CRNs, and present a novel Listen-and-Talk

0163-6804/15/$25.00

(LAT) protocol by which the SUs can simultaneously perform spectrum sensing and data transmission [5]. Specifically, by equipping SUs with FD radios, they can sense the target spectrum band in each time slot, and determine if the PUs are busy or idle, and at the same time they can also transmit data or decide to keep silent based on the sensing results. Apparently, the proposed FD-CR system is totally different from the traditional HD-based one in many aspects, including:

- Spectrum sensing: In FD-CR, sensing is continuous, but the received signal for sensing is interfered with by the residual self-interference (RSI), which degrades the signal-to-interference-plus-noise ratio (SINR) in sensing. While in HD-CR, there is no RSI in the received signal for sensing, but the sensing process is discontinuous and only takes a small fraction of each slot. This leads to unreliable sensing performance due to the inefficient number of samples to make decisions.
- Data transmission: In traditional HD-CR, the SUs can only utilize the remaining part of each slot after sensing for data transmission. On contrary, in FD-CR, SUs can continuously transmit as long as PUs are absent. However, in FD-CR, the data transmission affects the sensing process; thus, there is a constraint on transmit power to achieve acceptable sensing performance.

In summary, FD technology enables to explore another dimension of the network resources for increasing the capacity of CRNs. Therefore, this requires a new design of signal processing techniques, resource allocation algorithms, and network protocols. For example, one of the major challenges faced by FD-CR is how to optimize the transmit power for the FD source node to maximize the system throughput. This article comprehensively discusses the novel protocol design issues, key system parameter derivation, and practical algorithms for FD-CR systems. In addition, we also extend the proposed FD-CR to *distributed* and *centralized* network scenarios, demonstrate new research challenges, and provide the latest promising research development. Some potential research directions and open problems are also discussed.

Basics of Full-Duplex Communications and Cognitive Radio

This section briefly introduces a basic FD communication system and the conventional LBT protocol.

Full-Duplex Communication System

Figure 1 presents a simple two-node FD communication system with a transmit and a receive antenna at each node. The FD technique allows two nodes to transmit and receive the signals at the same frequency and time interval, that is, node i can transmit its signal x_i ($i = 1, 2$) with the transmit antenna and receive x_{3-i} from the other node via the receive antenna in the same channel simultaneously. However, this leads to severe self-interference caused by the signal leakage from the transmit RF unit to the receive RF unit, as shown in Fig. 1. In practical FD systems, the self-interference cannot be completely canceled, such that the signals received at each node is a combination of the signal transmitted by the other source, the RSI, and the noise. Specifically, the RSI can be typically modeled as additive white Gaussian noise (AWGN), or Rayleigh or Rician distributed variables [4, 6], of which the variance is proportional to the transmit power.

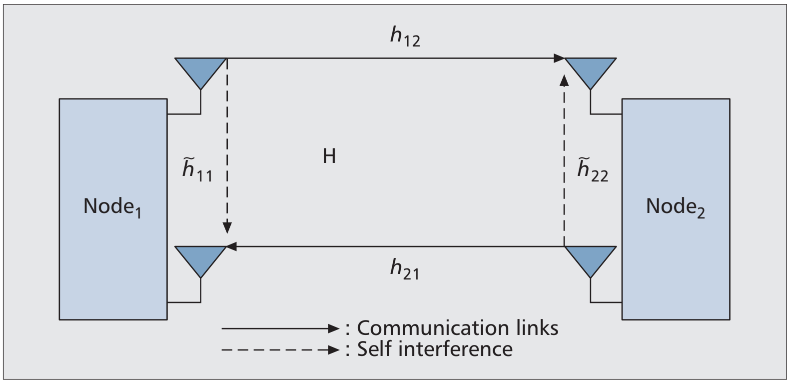


Figure 1. Full-duplex wireless communications.

A Conventional Cognitive Radio Communication System

In a conventional CRN, SUs are allowed to access the spectrum allocated to PUs without causing severe interference to the primary network. The typical assumptions are that PUs do not notify SUs when they begin or stop transmission, and do not wait when SUs are using the channel. Thus, SUs need to reliably identify the spectrum holes to access and back off when PUs appear. Typically, SUs' traffics are slotted, and every SU's time slot is divided into two sub-slots for sensing and data transmission [3]. At the beginning of each slot, all SUs sense the spectrum with a certain sensing strategy such as energy detection, matched filter detection, cyclostationary feature detection, or cooperative detection, and test the following two hypothesises:

- The spectrum is idle.
- The spectrum is occupied by a PU.

Once the spectrum is judged idle, some SUs transmit in the rest time of the slot according to a certain scheduling scheme to minimize the probability of potential collision among SUs [7].

Listen-and-Talk Protocol

In this section, we present our proposed LAT protocol by exploring the FD techniques at SUs in CRNs.

The LAT Protocol

Figure 2 shows the system model of the LAT protocol. Consider a CR system consisting of one PU and one SU pair (the left side of Fig. 2), in which SU_1 is the secondary transmitter and SU_2 is the receiver. Each SU is equipped with

Key parameters	Design methods
Sensing scheme	The LAT protocol uses energy detection as the sensing scheme, which requires less a priori information of PUs' signal patterns than other detection strategies;* however, the impact of RSI needs to be carefully considered since the average received power increases significantly with the transmit power.
Sensing threshold	The expected received power varies with the transmission of SUs due to the RSI, and the expected received power increases with the transmit power. Thus, the sensing threshold at a certain SU under the LAT protocol is no longer fixed, but adjusted with its transmit power. Thus, the SU needs to check its own activity and choose the best threshold accordingly in each slot.
Length of time slot	Generally, a longer time slot length leads to better sensing performances since more samples are considered, but, the probability that the PU changes its state in one slot also increases, which would make the sensing unreliable. Thus, a modest slot length needs to be carefully designed.

Table 1. Key design parameters in the LAT. *Note that other detection methods can also be used in the sensing process [13].

two FD antennas, Ant_1 and Ant_2. SU_1 performs sensing and transmission simultaneously: Ant_1 senses the spectrum continuously, while Ant_2 transmits data when a spectrum hole is detected. Specifically, at the end of each slot, SU_1 makes the decision on the PU's presence, and determines whether to transmit or not in the next slot. Assume that the PU can utilize the spectrum freely; thus, there are the following four cases of spectrum utilization, as shown on the right side of Fig. 2:

- $Case_1$: The spectrum is occupied only by the PU, and SU_1 is silent.
- $Case_2$: The PU is absent, and SU_1 utilizes the spectrum.
- $Case_3$: Neither the PU nor SU_1 is active, and a spectrum hole remains.
- $Case_4$: The PU and SU_1 both transmit, and a collision happens.

Among these four cases, $case_1$ and $case_2$ are the normal cases, and $case_3$ and $case_4$ are referred to as spectrum waste and collision, respectively. The spectrum waste ratio and the collision ratio are the most important metrics to measure sensing performance. One key challenge in using the FD technique is that the transmit signal at Ant_2 of SU_1 is received by Ant_1, causing irreducible RSI, which then has a strong impact on the sensing performance. Specifically, when Ant_2 is idle, the received signal at Ant_1 only contains the potential PU's signal and noise, and is free of RSI; once the data transmission begins, the RSI is introduced to the received signal, which may lead to a sharp decrease of SINR in sensing.

Key Design Parameters

The protocol design needs to satisfy the strict constraint that the interference of SUs to the primary network should not exceed a certain level of the probability of collision:

$$P_c = \lim_{t\to\infty} \frac{\text{Collision duration}}{\text{PU's transmission time}}. \tag{1}$$

From Eq. 1, the sensing strategy, thresholds, and slot length can be derived and optimized as shown in Table 1.

Performance Analysis

Power-Throughput Trade-off — The saturated throughput of SUs can be calculated as

$$C = R \cdot (1 - P_w), \tag{2}$$

where P_w represents the ratio of wasted spectrum hole and R denotes the transmission rate of the channel. Consider the relation between transmit power and the two factors, R and $(1 - P_w)$. On one hand, the rate R is positively proportional to SU's transmit power. On the other hand, the sensing performance degrades with the increase of RSI; that is, with the same RSI level, the false alarm probability increases with the transmit power under the same constraint of the collision probability [5]. This, in turn, leads to increased P_w and a decrease of $(1 - P_w)$.

Thus, there exists a power-throughput tradeoff in this protocol: when the secondary transmit power is low, the RSI is negligible, and the spectrum will be used more efficiently, but the ceiling throughput is limited by R; when the transmit power increases, the channel rate becomes high-

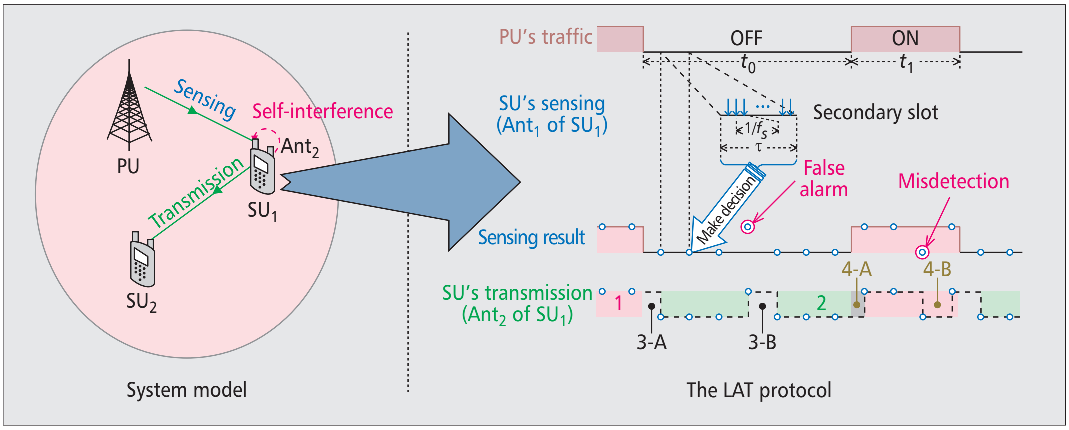


Figure 2. System model of the LAT protocol.

er, but the sensing performance is degraded due to the RSI.

Figure 3 shows the throughput performance of the LAT protocol in terms of secondary transmit power. The small circles are the simulated results, and the thin solid line depicts the ideal case with perfect RSI cancelation. Without the RSI, the sensing performance is not affected by transmit power, and thus the throughput always goes up with the power. It can be used to serve as an upper bound for the LAT performance. The thick dash-dotted, dotted, and dash lines in the middle are the typical cases, in which we can clearly observe the power-throughput trade-off and identify the local optimal power (the asterisks in the figure). With the decrease of the RSI

$$\chi^2 = \frac{\text{power of RSI}}{\text{SU's transmit power}}$$

for –10 dB, – 20 dB, and –30 dB, the optimal transmit power increases, and the corresponding throughput goes to a higher level. This is because the smaller the RSI is, the closer it approaches the ideal case, and the deterioration caused by self-interference becomes dominant under higher power. However, when χ^2 is sufficiently large and close to 1, there is no power-throughput trade-off (see the thick solid line denoting the cases when χ^2 = –0.46 dB = 0.9). The sensing performance is unreliable even if the transmit power is small and no local optimal point can be found in this curve. One noticeable feature of the LAT protocol shown in Fig. 3 is that when self-interference exists, all curves approach the thin dotted line $C = 0.5\,R$ when the power goes up. This line indicates the case that the spectrum waste is 0.5, that is, the sensing is totally ineffective. When the transmit power is too large, severe self-interference largely degrades the performance of spectrum sensing. The larger χ^2 is, the sooner the sensing gets unbearable and the throughput approaches the line for $C = 0.5 \times \log_2(1 + \gamma_t)$.

As shown in both analytical and simulated results, in the low-power region, the SUs' throughput first increases and then decreases such that a local optimal transmit power can be found, while in the high-power region, the SUs' throughput increases monotonically with the power. For each SU, the maximum transmit power is limited by the physical structure. The existence of the local optimal transmit power indicates that the maximum power may not always be the best choice to achieve the highest throughput. Instead, if the maximum power exceeds the local optimal one, but is not high enough, modest power may be much better than the maximum in achieving both higher throughput and longer transmission time.

Comparison between the LAT and Conventional LBT Protocols — Since there are limitations of both LBT and LAT protocols, we make comparisons under the same model shown in Fig. 2. For fairness, in the LBT, SU_1 uses both antennas to sense and transmit, and correlation between the two antennas needs to be considered. In the LBT, the data transmission time is

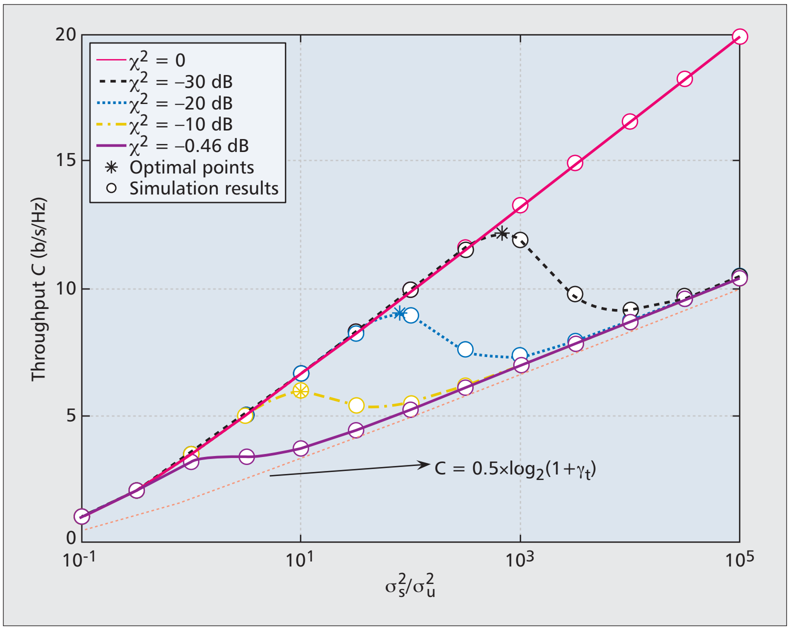


Figure 3. Power-throughput curves, in which the curves represent the analytical performance, and the circles on curves show the simulation results.

reduced because of spectrum sensing, while in the LAT, the RSI is the main cause for decreased performance.[1] Here we list some important parameters that influence the SUs' throughput in Table 2. Note that the proposed LAT may not always outperform the conventional protocol, especially when secondary transmit power is sufficiently large and the impact of RSI becomes unbearable. However, in most cases the LAT performs better due to its higher utilization efficiency of the spectrum holes, and prompt detection and reaction to the PU's state change.

Full Duplex Cognitive Radio Applications

In this section, we focus on some key unique applications of FD-CR systems.

Distributed Scenarios

The distributed scheme has no central infrastructure for coordinating the common channel access procedure. Hence, each FD-CR user that is going to transmit has to take the contention procedures and resolve possible collisions.

Distributed Spectrum Access Scheme: We first introduce a simple distributed FD-CR system. Consider M non-cooperative FD-SUs contending for one channel of PUs [10]. Each SU senses the channel by the LAT protocol independently and accesses the spectrum without communicating with each other. To avoid collision among SUs, a distributed dynamic spectrum access (DSA) scheme is needed. The difference of using the FD technique in the design of a DSA scheme is that PUs are no longer "blind" to SUs when SUs are transmitting data; instead, SUs can detect in real time the state changes of PUs and other SUs continuously, and when collision happens, SUs can back off immediately

[1] Note that if a LAT system is compared with a conventional LBT system consisting of SUs with single antennas, the comparison would be more straightforward since there is no spatial correlation in the LBT. Also, the switching would still exist due to the effects of RSI.

Parameters	LAT	LBT
Spatial correlation	NA	Both sensing and transmission deteriorate with the increase of spatial correlation.
Sensing duration	NA	Non-monotonous impact. An optimal ratio of sensing duration can be found [8].
Secondary slot length	Only sensing performance needs to be considered in the optimization. The optimal slot length is typically similar to sensing duration in LBT.	It requires the joint optimization of the durations of sensing and transmission sub-slots.
Secondary tansmit power	The power-throughput trade-off exists, and a local optimal transmit power can be found.	Throughput increases with the power monotonously.
RSI level	Better suppression level leads to higher local optimal power and throughput.	NA

Table 2. Comparison between LAT and LBT.

before finishing the current packet. This brings about the possibility of a new DSA scheme with less spectrum waste and shorter collision length.

Full Duplex MIMO System — As shown in Fig. 4, it consists of a pair of FD multiple-input multiple-output (MIMO) transceivers, nodes A and B , where each node is equipped with multiple antennas (N), respectively. In each node, some antennas (N_s) are used for sensing, some (N_t) for data transmission, and some (N_r) for receiving data from the other CR node. Both nodes operate in the same frequency band at the same time. Hence, if $N_s = N$, the system becomes the traditional CR with LBT; when all these three parameters are employed, this system supports bidirectional communication while sensing, but the interference is quite complicated among the antennas.

Centralized Scenarios

In the centralized scenarios, an access point (AP) establishes the connection with the mobile users, which are served in the coverage area.

Full Duplex Cognitive Access Point System — Figure 5 shows a simple secondary AP system consisting of an FD cognitive orthogonal frequency-division multiplexing (OFDM)-AP with two antennas and a number of SUs [11]. One antenna in the AP senses the PU's spectrum, and the other antenna provides wireless service for the SUs, in which the downlink transmission brings the RSI to the sensing results. Hence, how to assign SUs to the proper subcarriers and adjust the transmit power becomes essential for system performance.

Full Duplex Cognitive Relay System — By equipping the relay node with an FD radio, the relay can receive and retransmit signals simultaneously, so spectral efficiency can be improved compared to the traditional HD relay. Figure 6 illustrates a simple FD cognitive relay system consisting of one source, one destination, and one FD relay node. Both the source and relay nodes use the same time-frequency resource, and the relay node works in FD mode with three antennas (one for spectrum sensing, one for reception, and one for transmission). The communication process can be briefly described thus:
- The source transmits signals to the FD relay.
- At the same time, the FD cognitive relay performs spectrum sensing.
- The relay forwards the signals received in the previous time slots to the destination.

Note that different from traditional HD relay, the FD relay uses two antennas to receive data from the PU for sensing and the source for data forwarding. Thus, since FD works in the same frequency band, these two receive antennas actually have a combination of the PU, source, and RSI signals.

Research Problems

In this section, we summarize the main research problems for FD-CR communication systems. Similar to traditional wireless systems, multidimensional resource on space, time, frequency, power, and nodes need to be properly managed to optimize overall system performance. In particular, FD communication provides another dimension of resource, and its performance is also greatly affected by the RSI. Some key research problems as well as possible solutions are also summarized.

Signal Processing Techniques

Spectrum Sensing — Non-cooperative narrowband sensing was elaborated on earlier. However, the degradation of sensing performance becomes unbearable when transmit power is high. A promising solution for this problem is *cooperative sensing*, in which the sensing results of several SUs are combined to make a final decision. However, when employing cooperative spectrum sensing in FD-CRNs, there is interference from the transmitting SU to other SUs, which makes the local sensing different from the conventional HD case [9].

Also, since CRNs will eventually be required to exploit spectrum opportunities over a wide frequency range, an FD-enabled *wideband sensing scheme* is needed. With the impact of RSI, the original sparsity, which is the basis of the conventional wideband sensing scheme, would change, and the whole sensing scheme may be different.

Multiple Antenna Techniques: If multiple antennas are equipped at the FD-CRs, beamforming and antenna selection can be employed to further improve the secondary network performance:

• Transmit beamforming: Transmit beamforming is used to control the directionality of transmission in order to provide a large antenna array gain in the desired directions. For an FD cognitive MIMO system, the transmit antenna set at each FD-CR node can perform transmit beamforming to simultaneously transmit information and reduce the interference to its own received sensing signals. The design is to jointly optimize the sum rate of the system. If the FD cognitive AP node, which serves a group of users, is equipped with multiple antennas, it may be able to support multiple downlink transmissions while maintaining reliable sensing performance by using a certain structure of antennas to minimize the RSI.

• Antenna selection: For an FD-CR node, especially one with multiple antennas, each antenna can be configured to sense (receive) or transmit the signals. This creates an important problem to optimally select the antenna configurations that optimize system performance [6]. In an FD cognitive MIMO system, the problem is to choose one group of antennas to sense the spectrum, one group to transmit, and the rest to receive signals from another SU simultaneously. Such a combinatorial problem becomes much more complicated as the number of antennas increases. Similarly, in a general FD cognitive relay system, each relay can adaptively select its sense, transmit, and receive antennas based on the instantaneous channel conditions to achieve reliable sensing as well as maximum SINR in transmissions.

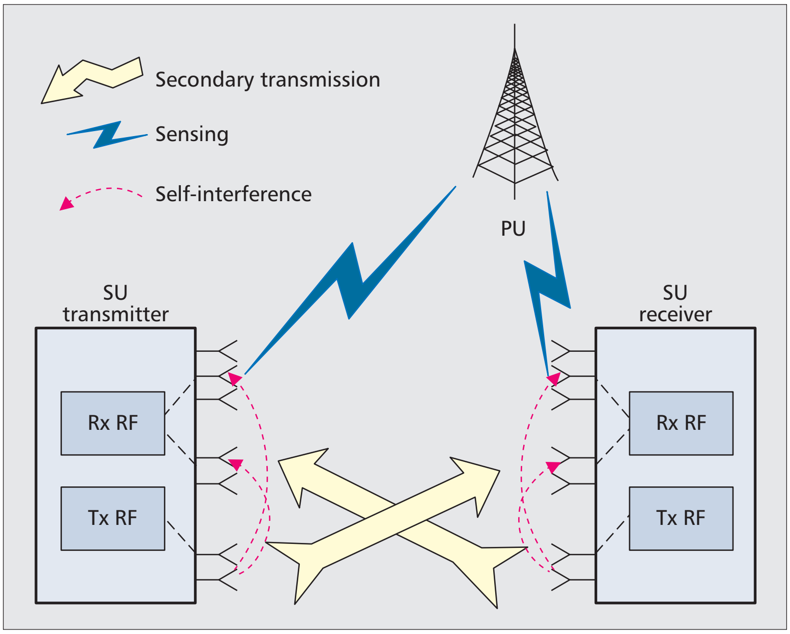


Figure 4. Full duplex cognitive MIMO system.

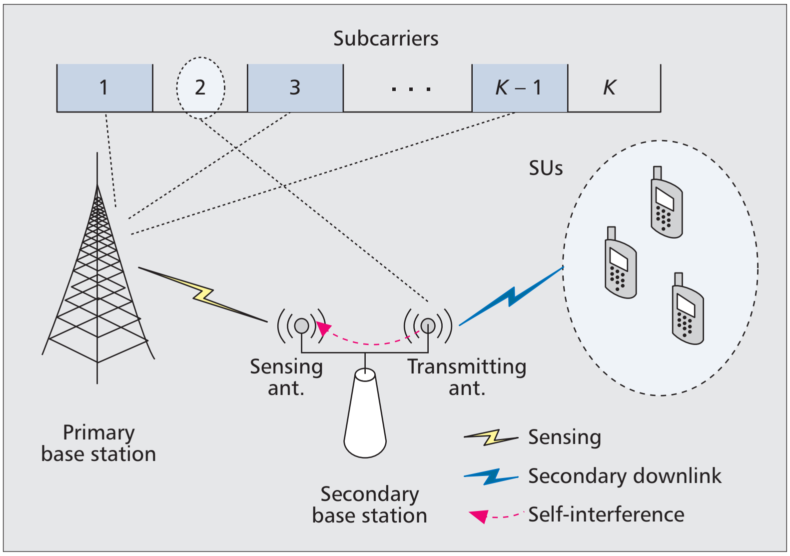


Figure 5. Full duplex cognitive access point system.

Dynamic Spectrum Access and Management

DSA is the key approach in the CRNs through which a cognitive wireless node is able to adaptively and dynamically transmit and receive data in a changing radio environment.

Distributed Dynamic Spectrum Access: In many scenarios such as ad hoc CRNs where the SUs compete for several PU channels, deploying a central controller is not always possible. Therefore, distributed DSA will be required, by which each SU has to independently gather, exchange, and process the information of the wireless environment. The commonly used carrier sense multiple access with collision avoidance (CSMA/CA) in distributed DSA with HD users can effectively reduce collision probability [7], but some problems still exist:

- Collision among the SUs can never be detected if the SUs are synchronized, such that the secondary transmission may fail on a large scale.
- SUs cannot abort transmission when collision occurs, which leads to long collision duration. With FD-CR, not only can the SUs detect the presence of PUs, but can also detect collision with other SUs during transmission such that the collision duration is reduced significantly. However, the RSI may degrade the collision detection accuracy, which cannot be ignored [10].

Centralized Dynamic Spectrum Access: For centralized DSA, a central controller is deployed to gather and process information about the wireless environment. Therefore, the decision of the SUs to access the spectrum can be made such that the desired system-wide objectives are achieved. This centralized control method fits the FD cognitive AP/relay scenarios very well, where the AP/relay performs spectrum sensing

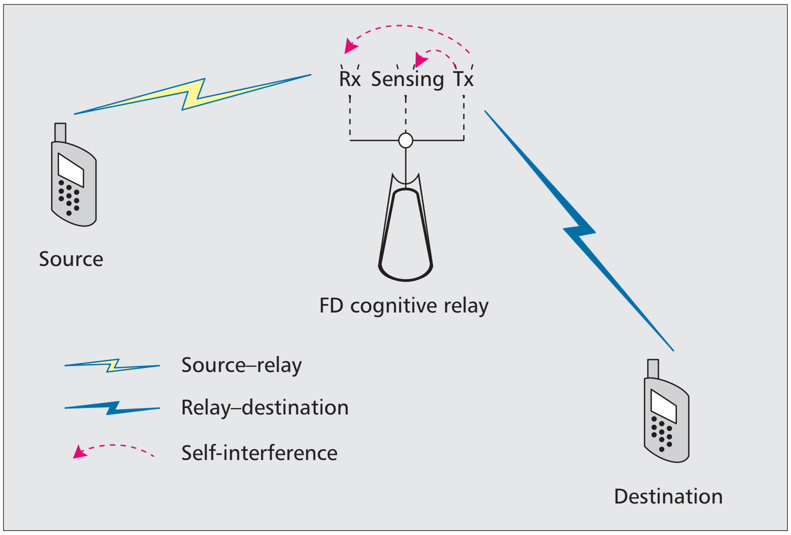


Figure 6. Full-duplex cognitive relay system.

and monitors the users below. Specifically, in an FD cognitive AP system consisting of one FD OFDM AP and some downlink SUs, a fundamental challenge is how to match available subcarriers with downlink SUs to optimize the network performance [11]. In this scenario, the consideration of tolerable RSI at the AP for reliable sensing is the main difference from conventional radio allocation problems. Similarly, in an FD cognitive relay system consisting of multiple secondary source and destination pairs and a FD relay node, the corresponding subcarriers should be properly allocated taking account of the RSI.

Power Control — Power control has been a commonly used approach in multi-user communication systems to optimize system performance such as link data rate, network capacity, and coverage. Unlike traditional wireless networks, FD-CR communication suffers from RSI, which deteriorates sensing performance when transmit power increases. Therefore, the corresponding power control algorithm needs to be properly redesigned in order to optimize system performance. Different power constraints (e.g., *total* or *individual* transmit power) will lead to different designs and final solutions. Moreover, as detailed below, different FD-CR systems require different power control algorithms:

- FD cognitive MIMO system: The antennas at the FD node are divided into sensing, receive, and transmit antenna sets with individual power constraints. Considering the RSI at the sensing set, the optimal power pouring mechanism can be significantly different from the conventional water-filling.
- FD cognitive relay system: The relay is under individual power constraint, and both the relayed signals and sensing results are corrupted by the RSI. Increasing the transmit power at the relay will increase the signal-to-noise ratio (SNR) at the destination, but on the other hand decrease the accuracy of sensing and blur the received signals from the source. The optimization needs to consider these issues.
- FD cognitive AP system: The FD cognitive AP should sense the spectrum while communicating with the SUs at the same time. The transmit power can be allocated at the AP side with total power constraint by splitting the power among all the subcarriers for different SUs. Thus, the optimal power control needs to jointly consider sensing and transmission in different subcarriers.

Multiple Networks Coexistence

Spectrum sharing has been recognized as a key remedy for the spectrum scarcity problem, especially after the successful deployment of WLAN and wireless personal area network (WPAN) devices on an unlicensed band (e.g., industrial, scientific, and medical, ISM, band). However, severe performance degradation has been observed when heterogeneous devices share the same frequency band due to mutual interference rooted in the lack of coordination. The cooperative busy tone (CBT) algorithm allows a separate node to schedule a busy tone concurrently with the desired transmission, thereby improving the visibility among different sorts of devices [12]. But preventing the busy signal from interfering with a data packet still remains a problem. By deploying FD techniques, the coexistence between heterogeneous networks may become more flexible. The research problem is to further reduce the RSI impact and realize efficient spectrum access management.

Conclusions

This article presents a new paradigm for future CR by exploring FD technology to allow SUs to simultaneously sense and access the vacant spectrum. Novel protocol design and key parameter derivation are explained in depth. Both analytical and simulated results indicate that the proposed LAT protocol can efficiently improve the spectrum utilization. Feasible applications with FD enabled CR have been elaborated on in centralized and distributed scenarios. The associated signal processing and spectrum access problems in these systems are also outlined. Future work may further study the scenario with multiple PUs and SUs in CRNs, and introduce basic economic theories as a tool to study and analyze FD-CRNs, such as dynamic spectrum access and spectrum trading problems.

References

[1] L. Song and J. Shen, *Evolved Network Planning and Optimization for UMTS and LTE*, Auerbach Publications, CRC Press, 2010.
[2] J. Mitola and G. Q. Maguire, "Cognitive Radio: Making Software Radios More Personal," *IEEE Personal Commun.*, vol. 6, no. 4, Aug. 1999, pp. 13–18.
[3] T. Yucek and H. Arslan, "A Survey of Spectrum Sensing Algorithms for Cognitive Radio Applications," *IEEE Commun. Surveys & Tutorials*, vol. 11, no. 1, Mar. 2009, pp. 116–30.
[4] J. Choi *et al.*, "Achieving Single Channel, Full Duplex Wireless Communication," *16th Annual Int'l. Conf. Mobile Computing and Networking*, New York, NY, Sept. 2010.
[5] Y. Liao *et al.*, "Listen-and-Talk: Full-Duplex Cognitive Radio Networks," *Proc. IEEE GLOBECOM*, Austin, TX, Dec. 2014.
[6] M. Zhou *et al.*, "Transmit-Receive Antenna Pair Selection in Full Duplex Systems," *IEEE Wireless Commun. Lett.*, vol. 3, no. 1, Feb. 2014, pp. 34–37.

[7] S. Huang, X. Liu, and Z. Ding, "Opportunistic Spectrum Access in Cognitive Radio Networks," *Proc. IEEE INFOCOM*, Phoenix, AZ, Apr. 2008.
[8] Y. C. Liang *et al.*, "Sensing-Throughput Tradeoff for Cognitive Radio Networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 4, Apr. 2008, pp. 1326–37.
[9] Y. Liao *et al.*, "Cooperative Spectrum Sensing for Full-Duplex Cognitive Radio Networks," *Proc. IEEE Global Commun. Confs.*, Macau, China, Nov. 2014.
[10] Y. Liao *et al.*, "Decentralized Dynamic Spectrum Access in Full-Duplex Cognitive Radio Networks," *Proc. IEEE ICC*, London, U.K., Jun. 2015.
[11] T. Wang *et al.*, "Joint Spectrum Access and Power Allocation in Full-Duplex Cognitive Cellular Networks," *Proc. IEEE ICC*, London, U.K., June 2015.
[12] X. Zhang and K. G. Shin, "Enabling Coexistence of Heterogeneous Wireless Systems: Case for ZigBee and WiFi," *Proc. ACM Int'l. Symp. Mobile Ad Hoc Networking and Computing*, New York, NY, May 2011.
[13] T. Riihonen and R. Wichman, "Energy Detection in Full-Duplex Cognitive Radios under Residual Self-Interference," *Proc. Cognitive Radio Oriented Wireless Networks and Commun.*, Oulu, Finland, June 2014, pp. 57–60.

BIOGRAPHIES

YUN LIAO [S'14] has been pursuing her B.S. degree at the School of Electrical Engineering and Computer Science of Peking University since 2011. Her research interests mainly include cognitive radio networks, full-duplex communications, cooperative communications, wireless network virtualization, and optimization theory.

LINGYANG SONG [S'03, M'06, SM'12] received his Ph.D. from the University of York, United Kingdom, in 2007, where he received the K. M. Stott Prize for excellent research. He worked as a research fellow at the University of Oslo, Norway, and Harvard University, until rejoining Philips Research UK in March 2008. In May 2009, he joined the School of Electronics Engineering and Computer Science, Peking University, China, as a full professor. His main research interests include MIMO, cognitive and cooperative communications, physical layer security, and wireless ad hoc/sensor networks. He has published extensively and has written three textbooks. He was the recipient of the 2012 IEEE Asia Pacific (AP) Young Researcher Award, and has received seven best paper awards. He is an IEEE Distinguished Lecturer.

ZHU HAN [S'01, M'04, SM'09, F'14] received his B.S. degree in electronic engineering from Tsinghua University in 1997, and his M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1999 and 2003, respectively. From 2000 to 2002, he was an R&D engineer at JDSU, Germantown, Maryland. From 2003 to 2006, he was a research associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is an associate professor in the Electrical and Computer Engineering Department at the University of Houston, Texas. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, wireless multimedia, security, and smart grid communication. He has been an Associate Editor of *IEEE Transactions on Wireless Communications* since 2010. He was awarded the IEEE Fred W. Ellersick Prize 2011. He was an NSF CAREER award recipient in 2010. He is an IEEE Distinguished Lecturer.

YONGHUI LI [M'04, SM'09] received his Ph.D. degree from Beijing University of Aeronautics and Astronautics, China, in 2002. From 1999 to 2003, he was affiliated with Linkair Communication Inc, where he held the position of project manager with responsibility for the design of physical layer solutions for the LAS-CDMA system. Since 2003, he has been with the Centre of Excellence in Telecommunications, University of Sydney, Australia. He is now an associate professor in the School of Electrical and Information Engineering, University of Sydney. He was the Australian Queen Elizabeth II Fellow and is currently the Australian Future Fellow. His current research interests are in the area of wireless communications, with a particular focus on MIMO, cooperative communications, coding techniques, and wireless sensor networks. He holds a number of patents granted and pending in these fields. He is an Executive Editor of *European Transactions on Telecommunications*. He has also been involved in the Technical Committees of several international conferences, including ICC, GLOBECOM, and others.

Exploiting Full Duplex for Device-to-Device Communications in Heterogeneous Networks

Li Wang, Fei Tian, Tommy Svensson, Daquan Feng, Mei Song, and Shaoqian Li

Li Wang, Fei Tian, and Mei Song are with the Beijing Key Laboratory of Work Safety Intelligent Monitoring at Beijing University of Posts and Telecommunications.

Tommy Svensson is with Chalmers University of Technology.

Daquan Feng and Shaoqian Li are withUESTC.

The work was supported in part by the National Natural Science Foundation of China under Grants No. 61201150, the Beijing Higher Education Young Elite Teacher Project under Grant No. YETP0442, the National High Technology Research and Development Program of China under Grant No. 2014AA01A701, the Swedish Research Council VR under the project 621-2009-4555 Dynamic Multipoint Wireless Transmission, the European Sharing Project, and the Doctor Foundation of Ministry of Education under Grant No. 20110185130003.

ABSTRACT

In this article, we discuss full duplex for heterogenous networks that accommodate the coexistence of device-to-device communications. The short link distance and lower transmit power of device-to-device communications make them excellent candidates to exploit full duplex inband transmission. By incorporating power allocation for self-interference cancellation based on antenna isolation, analog cancellation, and digital cancellation, full-duplex device-to-device, FD-D2D, nodes can potentially improve spectrum efficiency in HetNets. We provide a comprehensive overview on FD-D2D communications in HetNets. Additionally, we identify several challenges, provide potential solutions to interference mitigation based on power control, beamforming, and resource scheduling, and further discuss applications of FD in 5G networks.

INTRODUCTION

The craving for high data transmission rates is continuously fueled by the relentless growth of wireless data traffic volume due to the rapidly expanding population of broadband users and new emerging applications, such as mobile social networks and electronic commerce. However, available spectrum for wireless communications has already been pushed to the brink of exhaustion. Hence, new and disruptive techniques are crucial to meet this growing mobile traffic demand. As a result, heterogenous networks (HetNets) [1], full duplex (FD) wireless transmission [2], and device-to-device (D2D) communications [3] have been proposed to improve spectrum efficiency and deal with spectrum limitation.

FD communications allow a user equipment (UE) to transmit and receive signals at the same time over the same frequency, and thus can potentially double spectral efficiency. However, it has long been regarded as impractical in the past due to strong self-interference (SI) [2]. The critical challenge to exploit FD wireless transceivers lies in the huge power discrepancy between its own transmit and receive signal powers, with a ratio generally exceeding 100 dB. Therefore, reducing transmit power is a direct way to alleviate SI.

D2D communications and HetNets with small cells are the two key complementary technologies in Long Term Evolution (LTE)-Advanced systems. Both D2D links and small cells can offload traffic from macrocell base stations (BSs), extend network coverage, and improve spectrum and energy efficiency. They require low-transmission power in general due to short communication distance. As a result, it is natural and effective to integrate FD into HetNets with coexisting small cell and D2D communications [3, 4]. A number of exciting and challenging issues are emerging with regard to integrating FD communications into wireless networks, including antenna deployment, circuit design, physical layer protocol design, network foundations, and protocol design [2]. In this article, we discuss potential scenarios and applications of exploiting FD nodes in HetNets with three-tier architectures consisting of macrocell, femtocell, and D2D links, as shown in Fig. 1. Moreover, we focus on algorithms and protocols that are dedicated to reducing SI through resource allocation, power control, transmission mode selection, and scheduling.

The rest of this article is organized as follows. We elaborate SI cancellation (SIC) techniques, and demonstrate the role of FD transceivers in HetNets. We discuss full duplex D2D (FD-D2D) nodes and the relay-aided problem in HetNets. We present several potential solutions for interference control before introducing some open issues.

FD TRANSCEIVERS IN HETEROGENOUS NETWORKS

In this section, we first elaborate on three effective SIC techniques. We also articulate the advantages of utilizing FD nodes in various applications and major configurations of HetNets involving D2D links and small cells.

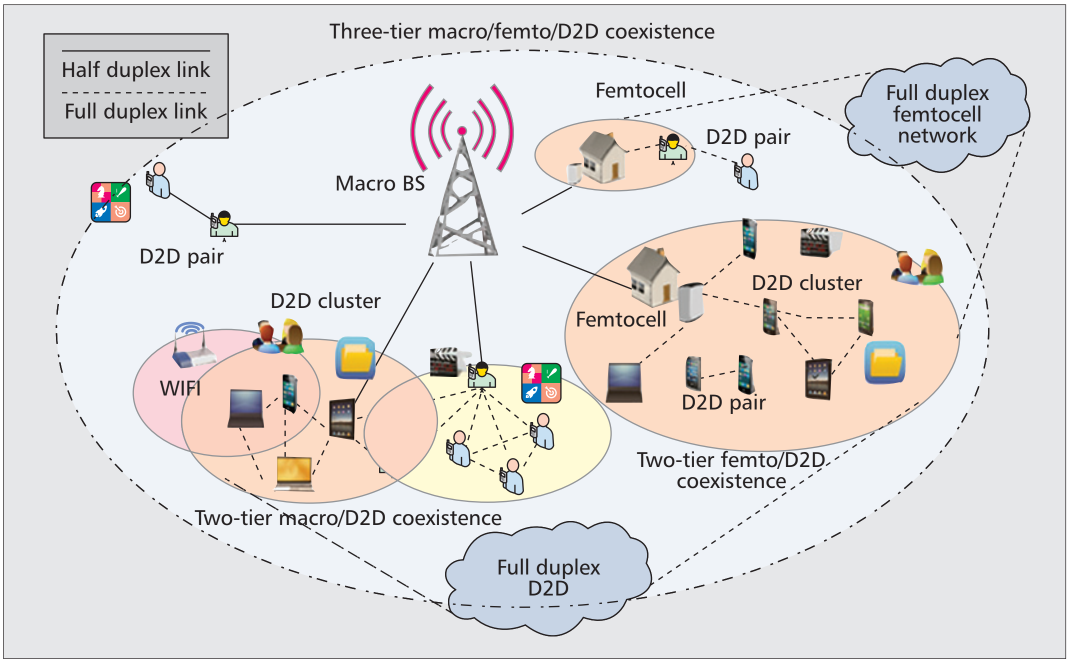


Figure 1. A heterogenous network with full duplex consideration.

D2D communications allow cellular UEs in close proximity to communicate directly without routing through base stations. D2D also presents substantial benefit in terms of energy saving because the shorter link distance requires smaller transmit power.

Self-Interference Cancellation in FD Radios

Three key techniques have been broadly utilized to mitigate SI [2] in FD nodes: physical separation of transmit and receive antennas, analog cancellation, and digital cancellation. Physical separation of antennas is a simple passive method, and can be further categorized into two types: antenna separation and shared antenna. Antenna separation uses multiple transmit antennas to generate signals with different phase shifts at receive antenna(s) such that the received signals from multiple transmit antennas can cancel one another because of destructive superposition. On the other hand, a shared antenna system isolates SI through a circulator. Analog and digital cancellation are achieved via circuit signal processing. The analog cancellation technique aims to remove substantial SI before the analog-to-digital converter (ADC). It may be dependent on or independent of channel state information (CSI). Channel-aware techniques can cancel interference from both the direct path and reflective paths, while channel-unaware techniques can only mitigate that of the direct path. The frequency selectivity of broadband wireless channels makes analog SIC much more challenging than flat fading channels. On the other hand, digital interference cancellation after the ADC applies digital signal processing techniques to the received signal. However, the dynamic range of ADC limits the capability of SI reduction in the digital domain.

To achieve the best interference cancellation performance, two or more SIC techniques are usually utilized together in FD communications. In particular, all three key techniques are combined together in [5], as shown in Fig. 2. In the figure, a two-antenna FD radio configuration includes antenna isolation, and allows the same signal to be sent from two transmit antennas. As a result, the received signals from the two transmit antennas are with opposite sign and therefore cancel each other. Interference cancellation performance from antenna suppression is affected by the accuracy of antenna positioning, signal bandwidth, transmit antenna, and RF calibration. Therefore, analog and digital cancellation are further adopted to achieve the desired level of SIC.

Most early works involving FD nodes assume perfect SIC based on accurate information of the SI channel [6]. The major challenges in SIC arise from estimating SI channels, uncertainties in RF circuitry such as oscillator phase noise, power amplifier nonlinearity, and limited dynamic range, and precision of ADC. However, recent works from both academia and industry have made FD communications more practical, due to advances in interference cancellation techniques, as listed in Table 1 for the three key SIC techniques. Inspired by the pioneering work in [5], other recent advanced SIC works have studied the impact of oscillator phase noise, designed antennas with appropriate form factor, and developed transmission scheduling and networking jointly to overcome SI [2, 7].

Why FD in D2D Communications?

D2D communications allow cellular UEs in close proximity to communicate directly without routing through BSs [8, 9]. D2D also presents substantial benefit in terms of energy saving because the shorter link distance requires smaller transmit power. Neighboring UEs may form D2D pairs (links). Thus, in addition to the cellular mode, many cellular UEs may select a D2D mode.

D2D links reuse the same spectrum resource concurrently with some cellular users simultaneously, which is called D2D underlay communications. To manage mutual interference in D2D underlay, resource pairing between cellular users

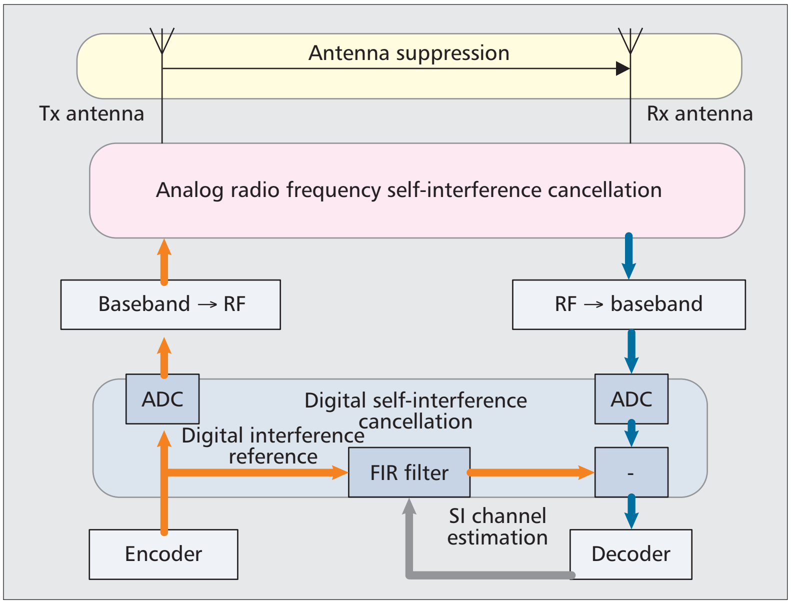


Figure 2. Block diagram of FD interference cancellation.

and D2D links (pairs) is highly critical to achieving high system throughput and low outage.

Integration of FD nodes in D2D may further improve spectrum efficiency [10]. In FD D2D links, SI is much weaker due to lower transmit power required by shorter communicate distance.

To optimize system performance, SI in FD communications and co-channel interference of D2D underlay should be addressed jointly through power control, beamforming, scheduling, and link adaptation. Given FD nodes, scheduling, mode-selection, and link adaptation in D2D communications become even more important and challenging since more accurate CSI and neighbor discovery are required. Furthermore, we should pay more attention to whether it is viable to exploit FD transmission according to interference constraints, or just keep half duplex (HD) instead. HD may use either frequency-division duplexing (FDD) or time-division duplexing (TDD).

FD Deployment in Femtocell and D2D Networks

To reduce hardware retrofitting cost, we may only consider FD radio applications in two-tier femtocell and D2D HetNets, while limiting the macrocell to operate under traditional HD. Self-interference seriously affects FD radios, whereas different types of co-channel interference also exist. For simplicity, we assume nodes within femtocell coverage will likely access femto BSs instead of macro BSs. Nodes within femtocell coverage but in close proximity may establish direct D2D links. Only D2D nodes outside femtocell coverage may collaborate by taking advantage of forming relays. Note that, also for cost consideration, only intermediate nodes are always assumed to use FD. Considering two-tier HetNets that involve femtocell and D2D communications, we divide the networking with FD radios into three different categories, as shown in Fig. 3:

- **Bidirectional direct link**: Source (S) and destination (D) nodes communicate directly in FD mode with asymmetric traffic and symmetric channels. In this case, there may be strong SI at both source and destination nodes. Note that both node A and node B can be source node and destination node.
- **Relay assisted link**: One or more FD relay nodes are used to transmit data from source to destination with symmetric or asymmetric channels. Self-interference takes place at relay (R) nodes.
- **Base station assisted link**: Simultaneous downlink and uplink take place at femto BS for two HD users. The BS provides FD communications for both downlink and uplink users with asymmetric traffic and symmetric channels with the assistance of successful SIC.

Cooperative FD-D2D Communications

In this section, we introduce cooperative D2D communications in FD-enabled HetNets. In cooperative D2D networks, FD UEs can form clusters that facilitate reconfigurable connection topology. The optimization of link topology can improve the efficacy of FD SIC for better transmission and reception. Configuring clusters of FD nodes to form cooperative relay networks provides a promising solution to achieve better wireless network coverage and throughput. Relay selection, FD or HD mode determination, and power control in such cooperative networks can effectively facilitate higher throughput and better energy conservation. Because multihop D2D cooperation, D2D cluster, and one-hop relay all pose different challenges to FD, we discuss them separately below.

Multihop D2D Cooperation

As seen in Fig. 4, in multihop D2D communications, the source node (S) and destination node (D) can communicate via direct link as well as different relay hops, which demonstrates a typical scenario where both S and D are in HD, while relay nodes (Rs) are in FD. In the figure, the red solid lines, black dashed lines, and black solid lines denote FD SI, direct link between transceivers, and inter-relay hop link, respectively. Before deciding which way to communicate (direct or relay hop) between the source and the destination, an overall assessment is performed to compare specific physical metrics, such as total throughput under limited power, energy consumption, delays, and interference to other nodes. Note that short hops make SIC easier, whereas longer hops require high power. Thus, we may choose HD with fewer but longer hops or FD with more but shorter hops. In brief, HD is preferred for the direct link between the source and the destination, possibly at lower spectrum efficiency compared to FD. In addition, the specific time duration of each hop in multihop also affects performance. In general, we have to consider the overall system performance.

FD-Enabled D2D Clusters

Clustered D2D nodes can form efficient and reconfigurable networks to minimize the impact of SI in FD transceivers. In particular, we focus on D2D nodes equipped with multiple antennas that can form separate transmit and receive beams. In a general two-node bidirectional link, it would make little sense for a UE to apply different transmit and receive beams as the UE would prefer to transmit to, and receive from, its counterpart with maximum gain. Hence, such a traditional D2D link has less flexibility in optimizing transmit and receive beams to mitigate SI without substantially affecting their transmit or receive signal-to-noise ratio (SNR). In a user cluster, however, a simple network can be formed by recruiting other members as relay nodes such that the transmit and receive directions (and hence the beams) need not be similar. For this reason, FD D2D clusters, through reconfigurable networking, present another rich and practical research area for FD communications.

FD Relay Links

Optimizing FD relay to achieve optimal system performance should take SIC into account. In addition to the three key SIC techniques aforementioned, we may also deploy antenna array beamforming to further strengthen the ability to isolate SI for FD relays. In particular, by letting FD radios transmit to and receive from two nodes, we can flexibly design transmit and receive beams to achieve better SI mitigation. More recently, information-theoretic analyses of FD multiple-input multiple-output (FD-MIMO) relaying in [11] have also considered the effect of channel estimate error and amplifier nonlinearities on residual SI. For practicality, such factors should also be integrated into our problem formulation.

Interference Mitigation for FD D2D Networks

As mentioned earlier, interference mitigation plays a critical role in both FD and D2D communications. A comprehensive solution should jointly mitigate SI caused by FD and co-channel interference from D2D underlay. Existing works have demonstrated the potential of interference control to facilitate FD D2D communications. This section discusses several major solutions involving power control, mode selection, beamforming and precoding, as well as resource scheduling.

SIC solutions	Claimed SIC suppression
Antenna isolation	20–30 dB
Analog cancellation	20–45 dB
Digital cancellation	25–30 dB
Sum suppression of SIC techniques	113 dB

Table 1. Traditional self-interference cancellation techniques [5].

Power Control and Mode Selection

Power control plays a key role in FD communications. In addition to interference cancellation, it also contributes to energy savings. D2D links may share cellular spectrum resource with a cellular user only when the cellular user can achieve its own required quality of service (QoS). However, in FD-D2D networks, there is also SI in FD in addition to co-channel interference among cellular users and D2D pairs. Hence, D2D UE power control is important to limit co-channel interference from D2D links to cellular receivers, and guarantee a minimum data rate or a required outage probability threshold to cellular users [8]. A coverage area is defined in [10] such that only users therein will be potentially allocated D2D underly resources to avoid severe interference to the cellular users.

MIMO relay has attracted considerable attention recently due to its ability to increase spectral efficiency and its natural ability to exploit FD by spatial SI suppression [12]. In FD MIMO relaying systems, appropriately allocating limited power to each node (in a D2D pair or cluster), and assigning different transmission modes and different time durations, can improve the overall throughput and energy efficiency.

In an FD relaying system, different QoS requirements of UEs motivate mode selection

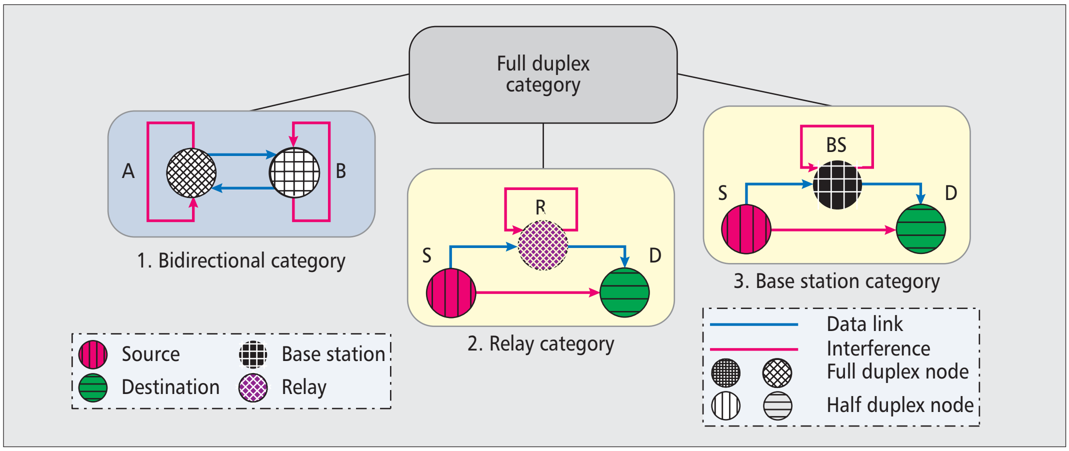


Figure 3. Types of FD communication links in a two-tier heterogenous network.

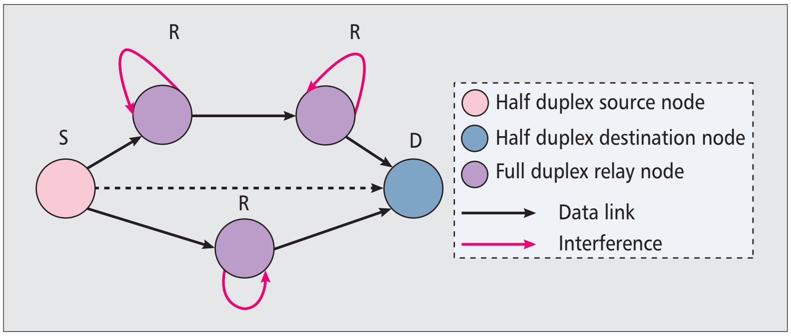


Figure 4. Direct link vs. multihop.

between HD and FD. The need to reduce energy consumption and transmission delay may also require an appropriate selection of HD or FD relay mode. Nevertheless, we should incorporate the effect of residual SI after SIC when comparing the performance of HD and FD for mode selection. In general, whether the FD or HD mode performs better depends on many factors, including the number of antennas, SNR, and residual SI.

Multi-Antenna Beamforming and Precoding

FD-MIMO has exhibited great potential for better SI mitigation. Through beamforming, FD-MIMO can significantly reduce SI via jointly optimizing transmit precoding and receiver processing.

MIMO relay can generate spatial nulls using receive antenna array in the directions of the transmit relays to mitigate SI. In [13], FD MIMO relay in both amplify-and-forward (AF) and decode-and-forward (DF) modes has been discussed, where beamforming is jointly considered with power control to maximize signal-to-leakage-plus-noise ratio (SLNR). In practice, limited transmitter/receiver dynamic range at various nodes and channel estimation errors should also be considered [12]. In [11], a gradient projection algorithm has been developed to optimize the power allocation vectors in a bidirectional system and a relay system with imperfect CSI.

We note that despite the advantages of FD-MIMO radios, multiple antennas are not commonly equipped on most UEs. Since many current D2D UEs only have a single antenna, to harness the advantages offered by FD-MIMO, it is more convenient and effective to form multiple cooperative (relay) networks to construct virtual MIMO arrays. Hence, more efforts are desired to overcome the challenges posed by synchronization errors and channel estimation errors in virtual FD MIMO systems.

It is worth noting that there is no inherent contradiction between spatial multiplexing and FD. First of all, spatial multiplexing requires the receiving node to be equipped with a sufficient number of receive antennas. As long as adequate SIC can be achieved, multiple transmit antennas can take part in spatial multiplexing in FD mode. On the other hand, when SI is so strong that it requires transmit antenna preprocessing or beamforming, the ability for spatial multiplexing will be constrained. Hence, the interesting trade-off between HD spatial multiplexing and FD SIC poses another important research problem.

Resource Scheduling

By exploiting FD radio in HetNets, we can improve system spectrum efficiency by using the same radio resource for transmission and reception simultaneously at different nodes within the same network. Careful scheduling in a local network is crucial to suppress SI so as to fully benefit from the capabilities of FD radios. Next, we elaborate on how FD techniques may support different multiple user access protocols in deployment scenarios that accommodate the coexistence of D2D links and small cells.

Traditionally, different multi-user access techniques can multiplex data streams to/from multiple users on a certain radio resource, typically involving HD on the same frequency bands. However, it remains unclear how to schedule multi-user access in FD networks so as to minimize the effect of co-channel interference and SI. In this section, depending on how FD is utilized, we discuss some potential resource scheduling solutions with FD/HD or hybrid mode in joint femtocell/D2D scenarios in terms of frequency-division multiple access (FDMA) and time-division multiple access (TDMA).

We consider a femto BS that can operate in both FD and HD modes in Fig. 5. In the first scenario, the femto BS has an HD transceiver. Up to four simultaneous data streams can be supported over two orthogonal radio channels with FDMA (case 1) and TDMA (case 2), respectively. In the first scenario, D2D UEs (UE2 and UE3) use lower transmit powers to control interference to femto communications. Both UE2 and UE3 can operate in FD (case 1) while reusing the uplink resource of cellular UE1. In case 2 of the first scenario, only UE2 is able to use FD; therefore, the D2D link between UE2 and UE3 needs to be supported by other means. Here, we can use the cellular downlink resource of UE2 instead to serve its own transmission to UE3 to achieve FD, whereas UE3 shares the cellular uplink resource of UE1.

The second scenario of Fig. 5 illustrates a hybrid case of an FD BS, one FD cellular node, and two HD D2D UEs. Similar to the first scenario, there are two cases of resource scheduling by means of FDMA (case 3) and TDMA (case 4). Taking case 3 as an example, the BS allocates a radio resource for UE1's FD transmission and reception, and also allocates another orthogonal radio resource for three links to share. Here each of the three radio nodes in the three circular links can transmit and receive simultaneously in FD mode. Case 4 carries the same features as case 3 from the perspective of the time domain but based on the same frequency band.

Opportunities and Challenges

Although FD has demonstrated potential in practical D2D and HetNets, thus far there are still a myriad of open issues for large-scale adoption in future wireless networks. Here we discuss a few possible directions.

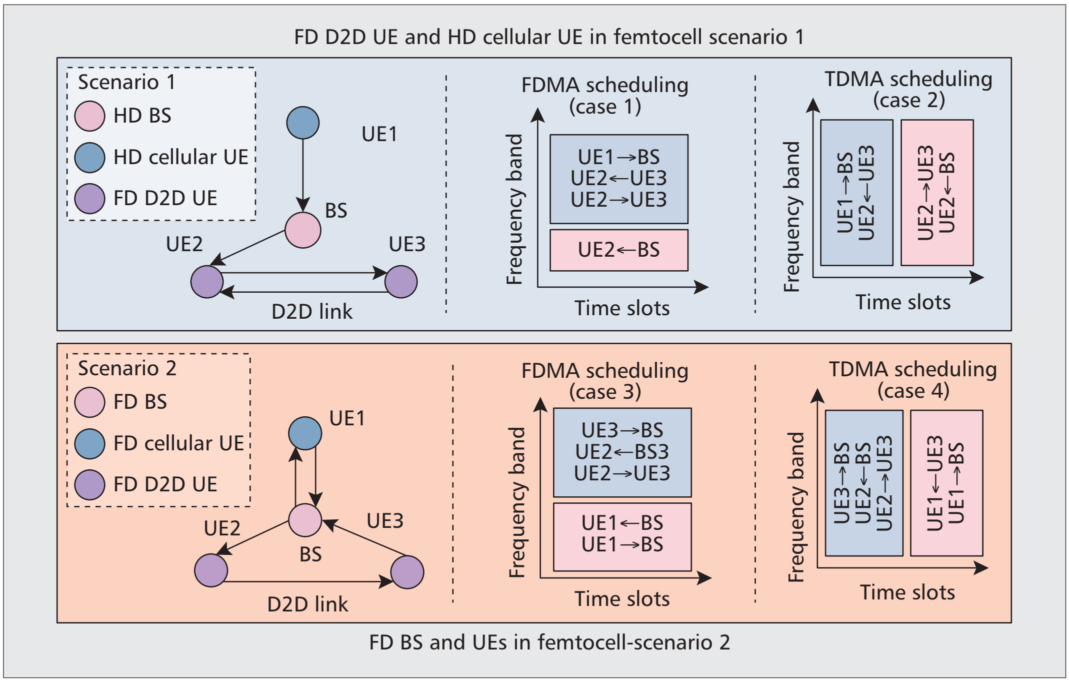


Figure 5. Resource scheduling for a D2D pair in a two-tier FD femto/D2D network.

FD-MIMO has exhibited great potential for better SI mitigation. Through beamforming, FD-MIMO can significantly reduce SI via jointly optimizing transmit precoding and receiver processing.

Exploiting FD in Emerging 5G Millimeter Wave Technology

Densely deployed small cells in the future may place a high burden on infrastructure, such as radio network control and fiber optic backhaul. Wireless solutions have been considered to replace current wireline backhaul. Furthermore, millimeter-wave (mmWave) has also attracted considerable attention as possible radio access technology in small cells because of its distinct characteristics, such as massive available bandwidth, large propagation attenuation (path loss), weaker building penetration as well as diffraction effects, and smaller-sized components due to shorter wavelengths [14].

With highly directional beamforming and weak building penetration, mmWave may facilitate co-channel interference cancellation among participating nodes in multihop relay networks, especially when the intermediate relay node in FD mode is in the middle of forwarding desired data. On the other hand, one of the major SIC techniques is antenna isolation, which means the physical distance between transmit antenna and receive antenna has to be accurately controlled and calibrated. For mmWave, the very short wavelength makes accurate antenna positioning and calibration much more difficult. Hence, SIC in mmWave can potentially pose greater challenges for FD radios. Other issues include the costly fabrication and difficult calibration of RF hardware for analog SIC.

Accuracy and Complexity Trade-off in Channel Estimation

Another open issue lies in the trade-off between accurate channel estimation for effective SIC and training overhead in fast fading and frequency-selective fading channels. For complex MIMO relays in FD D2D systems, it is possible to design a protocol that prepends a training phase to estimate CSI before payload. With the estimated CSI, power allocation can be optimized for payload transmission epochs. However, more practical and challenging factors need to be considered including channel estimation error, channel fading, residual SI, distortive RF amplifiers, oscillator frequency and phase noises, as well as imperfect ADC and digital-to-analog converter (DAC). Usually, to reduce training overhead, a single training epoch is responsible for a large number of payload epochs. One problem is the design of training epoch length vs. payload duration for both signal detection and SIC. In addition, a complex transmit precoding matrix is hard to realize, particularly when UE radio transceivers are not computationally highly powerful. Hence, the trade-off between complexity and performance is always a critical but challenging issue. Furthermore, unlike most works that target slow flat fading wireless channels, more practical channels with fast fading or frequency-selective characteristics should also be investigated.

Improving Scalability and Robustness

Thus far, there has been only limited work on network throughput improvement for FD communications. In fact, one issue lies in the scalability and robustness of spectrum efficiency doubling from end-to-end capacity to network level. In a single point-to-point link, FD can double wireless link capacity for a single pair of nodes. However, in practical networks, multiple nodes may share the spectrum based on protocols, such as carrier sense multiple access (CSMA). Unlike HD links, FD links have a much larger interference footprint at all times, thereby preventing nearby nodes from accessing the same spectrum to achieve spatial frequency

It is important to investigate means for routing traffic in order to maximize the total profit for multiple users with power constraints while meeting required QoS guarantee. We must also study and quantify the impact of non-ideal FD transceivers on the networks before their wide-spread integration into 5G systems.

reuse. In other words, FD does not necessarily double the network capacity even when assuming perfect SIC. FD may not provide scalable and robust spectrum efficiency doubling as one would hope when deployed within networks of various access protocols, which has been illustrated in asynchronous contention-based wireless networks [15]. In FD-D2D systems, however, this problem may be alleviated through better BS centralized control and resource scheduling, thereby presenting a research opportunity.

There are other issues with respect to higher network layers under FD. For example, it is important to investigate means for routing traffic in order to maximize the total profit for multiple users with power constraints while meeting required QoS guarantee. Furthermore, we must also study and quantify the impact of non-ideal FD transceivers on the networks before their wide-spread integration into 5G systems.

Conclusions

This article presents an overview of FD communications in heterogenous networks, and discusses the integration of FD radios in small cells and coexisting D2D underlay communications. Confronted by the critical issue of SI cancellation in FD radios, we demonstrate how D2D communications can better take advantage of FD radios, particularly in cooperative relay networks. We further analyze, in detail, several technical challenges and solutions on FD SI mitigation through power control, mode selection, beamforming, and precoding, as well as link and resource scheduling.

References

[1] M. Peng *et al.*, "Self-Configuration and Self-Optimization in LTE-Advanced Heterogeneous Networks," *IEEE Commun. Mag.*, vol. 51, no. 5, May 2013, pp. 36–45.

[2] A. Sabharwal *et al.*, "In-Band Full-Duplex Wireless: Challenges and Opportunities," *IEEE JSAC*, vol. 32, no. 9, Sept. 2014, pp. 1637–52.

[3] D. Feng *et al.*, "Device-to-Device Communications in Cellular Networks," *IEEE Commun. Mag.*, vol. 52, no. 4, 2014, pp. 49–55.

[4] Full-Duplex Radios for Local Access (DUPLO) Technical Report, DUPLO Deliverable D1.1, "System Scenarios and Technical Requirements for Full-Duplex Concept," Apr. 30, 2013.

[5] M. Jain *et al.*, "Practical, Real-Time, Full Duplex Wireless," *Proc. MobiCom.*, Las Vegas, NV, Sept. 2011.

[6] S. Simoens *et al.*, "On the Gaussian MIMO Relay Channel with Full Channel State Information," *IEEE Trans. Signal Process.*, vol. 57, no. 9, Sept. 2009, pp. 3588–99.

[7] D. Bharadia, E. Mcmilin, and S. Katti, "Full Duplex Radios," *Proc. ACM Sigcomm*, Hong Kong, 2013.

[8] L. Wang and H. Wu, "Fast Pairing of Device-to-Device Link Underlay for Spectrum Sharing with Cellular Users," *IEEE Commun. Lett.*, vol. 18, no. 10, Oct. 2014, pp. 1803–06.

[8] L. Wang, H. Tang, and M. Cierny, "Device-to-Device Link Admission Policy based on Social Interaction Information," *IEEE Trans. Vehic. Tech.*, no. 99, 2014.

[9] S. Ali, N. Rajatheva, and M. Latva-aho, "Full Duplex Device-to-Device Communication in Cellular Networks," *Proc. EuCNC*, Bologna, Italy, June 2014.

[10] A. C. Cirik, Y. Rong, and Y. Hua, "Achievable Rates of Full-Duplex MIMO Radios in Fast Fading Channels with Imperfect Channel Estimation," *IEEE Trans. Signal Proc.*, vol. 62, no.15, Aug. 2014, pp. 3874–86.

[11] B. P. Day, *et al.*, "Full-Duplex MIMO Relaying: Achievable Rates under Limited Dynamic Range," *IEEE JSAC*, vol. 30, no. 8, Sept. 2012, pp. 1541–53.

[12] U. Ugurlu *et al.*, "Power Control and Beamformer Design for the Optimization of Full-Duplex MIMO Relays in a Dual-Hop MISO Link," *Proc. CROWNCOM*, Oulu, Finland, June 2014, pp. 545–49.

[13] T. S. Rappaport *et al.*, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!," *IEEE Access*, vol. 1, May 2013, pp. 335–49.

[14] X. Xie and X. Zhang, "Does Full-Duplex Double the Capacity of Wireless Networks?," *Proc. IEEE INFOCOM*, Toronto, Canada, May 2014.

Biographies

Li Wang [S'08, M'14] (liwang@bupt.edu.cn) is an associate professor in the School of Electronic Engineering, Beijing University of Post and Telecommunications (BUPT), where she directs the Lab of High Performance Computing and Networks in BUPT. She received her Ph.D. degree in 2009 from BUPT. She received the 2013 Beijing Young Elite Educator for Higher Education Award. From December 2013 to January 2015, she was a visiting researcher in the School of Electrical and Computer Engineering at Georgia Tech, Atlanta. Her research interests include wireless networking, secure communications, device-to-device communication systems, and peer-to-peer networks. She has served on the Technical Program Committees of IEEE CCNC 2009, IEEE CCNC 2010, IEEE WCSP 2013, IEEE GLOBECOM 2014, IEEE WCNC 2015 , IEEE ICC 2015, IEEE ICNC 2015, and IEEE ICCC 2015.

Fei Tian received his Bachelor's degree from Hainan University, China, in 2013. Since 2013, he has been pursuing graduate studies at BUPT.

Tommy Svensson [S'98, M'03, SM'10] (tommy.svensson@chalmers.se) is an associate professor in the Communication Systems Department at Chalmers University of Technology, where he is leading the research on air interface and wireless backhaul networking technologies for future wireless systems. He received a Ph.D. in Iinformation theory from Chalmers in 2003, and worked at Ericsson AB with core networks, radio access networks, and microwave transmission products. He was involved in the European collaborative research projects WINNER I, II, WINNER+, and ARTIST4G, which have contributed to 3GPP LTE standards, and is currently active in the EU FP7 METIS project targeting solutions for 5G. His main research interests are in design and analysis of physical layer algorithms, multiple access, resource allocation, cooperative systems, moving relays/cells/networks, as well as satellite systems. He has co-authored two books, and more than 110 journal and conference papers. He is Chairman of the IEEE Sweden Vehicular Technology/Communications/Information Theory Chapter, and coordinator of the Communication Engineering Master's program at Chalmers.

Daquan Feng (fdquan@gmail.com) received his B.S. degree in communication engineering in 2008 from Henan University, Kaifeng, China. He is currently pursuing his Ph.D. degree at National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China (UESTC), Chengdu. His research interests include device-to-device communications, energy-efficient wireless network design, and heterogeneous networks.

Mei Song is a professor at BUPT. Her research interests include heterogenous networks, green communications, cognitive radio, and spectrum sharing technologies.

Shaoqian Li [SM] received his M.E. degree in information and communication systems from UESTC in 1984, and his B.S.E. in wireless engineering from Northwest Institute of Telecommunication Engineering, Xi'an, China, in 1981. Currently, he is a professor at UESTC, director of the National Key Laboratory of Science on Communications, and dean of the School of Communications and Information Engineering of UESTC. His research interests include anti-jamming techniques for wireless communications, frequency-hopping techniques, cognitive radio, and spectrum sharing technologies. He has co-authored two books, and published more than 40 referred journal papers and 200 conference papers. He is an inventor of more than 20 issued patents and more than 50 filed Chinese patents.

Call for Papers
IEEE Communications Magazine
Communications Standards Supplement

Background

Communications standards enable the global marketplace to offer interoperable products and services at affordable cost. Standards development organizations (SDOs) bring together stakeholders to develop consensus standards for use by a global industry. The importance of standards to the work and careers of communications practitioners has motivated the creation of a new publication on standards that meets the needs of a broad range of individuals, including industrial researchers, industry practitioners, business entrepreneurs, marketing managers, compliance/interoperability specialists, social scientists, regulators, intellectual property managers, and end users. This new publication will be incubated as a Communications Standards Supplement in *IEEE Communications Magazine*, which, if successful, will transition into a full-fledged new magazine. It is a platform for presenting and discussing standards-related topics in the areas of communications, networking, and related disciplines. Contributions are also encouraged from relevant disciplines of computer science, information systems, management, business studies, social sciences, economics, engineering, political science, public policy, sociology, and human factors/usability.

Scope of Contributions

Submissions are solicited on topics related to the areas of communications and networking standards and standardization research in at least the following topical areas:

Analysis of new topic areas for standardization, either enhancements to existing standards or in a new area. The standards activity may be just starting or nearing completion. For example, current topics of interest include:
- 5G radio access
- Wireless LAN
- SDN
- Ethernet
- Media codecs
- Cloud computing

Tutorials on, analysis of, and comparisons of IEEE and non-IEEE standards. For example, possible topics of interest include:
- Optical transport
- Radio access
- Power line carrier

The relationship between innovation and standardization, including, but not limited to:
- Patent policies, intellectual property rights, and antitrust law
- Examples and case studies of different kinds of innovation processes, analytical models of innovation, and new innovation methods

Technology governance aspects of standards focusing on both the socio-economic impact as well as the policies that guide them. These would include, but are not limited to:
- The national, regional, and global impacts of standards on industry, society, and economies
- The processes and organizations for creation and diffusion of standards, including the roles of organizations such as IEEE and IEEE-SA
- National and international policies and regulation for standards
- Standards and developing countries

The history of standardization, including, but not limited to:
- The cultures of different SDOs
- Standards education and its impact
- Corporate standards strategies
- The impact of open source on standards
- The impact of technology development and convergence on standards

Research-to-standards, including standards-oriented research, standards-related research, and research on standards

Compatibility and interoperability, including testing methodologies and certification to standards

Tools and services related to any or all aspects of the standardization life cycle

Proposals are also solicited for Feature Topic issues of the Communications Standards Supplement.

Articles should be submitted to the IEEE Communications Magazine submissions site at

http://mc.manuscriptcentral.com/commag-ieee

Select "Standards Supplement" from the drop-down menu of submission options.

TECHNOLOGIES AND APPLICATIONS OF FUTURE SATELLITE NETWORKING

Claudio Sacchi

Kul Bhasin

Naoto Kadowaki

Fred Vong

The second part of the Feature Topic "Satellite Communications and Networking: Emerging Techniques and New Applications," published in this issue of *IEEE Communications Magazine*, presents the overflow from the first part published in the March issue. In the first part, the accent was on a renewed vision of satellite communications and networking that we claim as "Space 2.0" [1] in order to mark a clear discontinuity with the "Space 1.0" era, begun in 1945 with A.C. Clarke's article "Extra Terrestrial Relays." The contributions published in the first part mainly focused on techniques that will characterize future satellite networking, such as exploitation of higher frequency bands, cognitive spectrum utilization, delay- and disruption-tolerant networks (DTNs), software-defined networking (SDN), and network virtualization. In the remaining part of the Feature Topic, we aim at shifting the emphasis to the application aspects without losing sight of technological investigation.

Indeed, one of the basic questions that arises among the social and business communities is "How can we exploit the cost efficiency, resilience, and ubiquity naturally provided by satellite networking to design new services for the emerging mass market of the next decades?" We believe that the five articles published in this issue can contribute an answer to this question.

The first article of the issue, "Challenges for Efficient and Seamless Space-Terrestrial Heterogeneous Networks," by J. P. Choi and C. Joo, analyzes the main challenges that should be solved in order to realize seamless space-terrestrial heterogeneous networks. The final aim is to provide multi-purpose readily available platforms for both commercial and no-profit services.

The second article, "Design Challenges in Contact Plans for Disruption-Tolerant Satellite Networks," by J. Fraire and J. Finochietto investigates one of the most critical aspects of DTN architectures: the provision of reliable and efficient contact plan design.

The third article, "Emergency Satellite Communications: Research and Standardization Activities," by T. Pecorella, L. S. Ronga, F. Chiti, S. Jayousi, and L. Franck, considers in detail the coordinated endeavors carried on at various levels (academic, industrial, international standardization bodies, national space agencies, governmental institutions, etc.) targeted at realizing a global infrastructure for emergency communications. In such a framework, satellite communications plays a key role.

The fourth article, "Alerting over Satellite Navigation Systems: Lessons Learned and Future Challenges," by T. De Cola and C. Pàrraga Niebla, proposes a very interesting application of satellite navigation systems (i.e., the deployment of public alerting messages in case of danger for citizens). Global navigation services can profitably support the efficient dissemination of alerting messages. In the article, the challenges and constraints characterizing this safety-critical service coexisting with the regular geolocalization service are carefully analyzed.

The last article, "Flexible Heterogeneous Satellite-Based Architecture for Enhanced Quality of Life Applications," by E. Del Re, S. Morosi, L. S. Ronga, S. Jayousi, and A. Martinelli, analyzes the synergistic use of communication, positioning, and monitoring techniques, provided by meshed terrestrial-satellite heterogeneous network architectures, to satisfy specific requirements coming from everyday life: e-Health and well-being, public safety, and mobility.

At the conclusion of this Feature Topic, the Guest Editors wish to thank the anonymous reviewers for their commitment in sending their timely and always valuable comments, the editorial staff of *IEEE Communications Magazine* and IEEE ComSoc for their continuous assistance, and, last but not least, all of the contributors. We received a large number of high-quality submissions, covering all aspects of satellite communications and networking. In our opinion, this is clear proof that, 70 years after the publication of Dr. Clarke's visionary contribution, satellite communication remains a field of increasing interest for advanced research and novel technology development.

REFERENCES

[1] C. Sacchi *et al.*, "Towards the Space 2.0 Era," Guest Editorial, *IEEE Commun. Mag.,* Part 1, Feature Topic on Satellite Communications and Networking: Emerging Techniques and New Applications, Mar. 2015.

BIOGRAPHIES

CLAUDIO SACCHI [SM] (sacchi@disi.unitn.it) obtained his Laurea degree in electronic engineering and Ph.D. in space science and engineering from the University of Genoa, Italy, in 1992 and 2003, respectively. Since August 2002 he has held a permanent position as an assistant professor on the Faculty of Engineering at the University of Trento. He is currently affiliated with the Department of Information Engineering and Computer Science at the University of Trento. He is the author and co-author of more than 80 contributions published in international journals, edited books, international conference proceedings, and so on.

KUL BHASIN [SM] is a senior space communications architect at NASA Glenn Research Center in Cleveland, Ohio. He has led a number of NASA-wide architecture teams who design space communication systems and ground networks. Currently, he is leading and designing NASA's next-generation space-based communications satellite architecture, as well as designing a network service portal system. Prior to that he held the position of system engineering and formulation manager for space communications projects at NASA GRC. He is an associate fellow of AIAA and a Fellow of the Society of International Optical Engineers (SPIE).

NAOTO KADOWAKI [M] received a B.S. in communications engineering, a Master's degree in information engineering, and a Ph. D. from the University of Tohoku, Sendai, Japan, in 1982, 1984, and 2010, respectively. From July 2004 to December 2006 he was the managing director of the Strategic Planning Department at NICT and director general of the Wireless Network Research Institute at NICT from July 2008 to June 2013. He is currently senior executive director and executive director of the Strategic Planning Department at NICT. He is a member of AIAA and IEICE of Japan.

FRED VONG is director of engineering at Asia Satellite Telecommunications Co. Ltd. in Hong Kong. He has over 19 years' experience in the satellite telecommunication industry. He is managing teams of engineers to support satellite procurement, engineering support to the satellite control center, customer satellite network design and qualification, service product development, frequency spectrum coordination, and ITU activities related to satellite communications. He holds Bachelor's and Master's degrees in engineering from the University of Hong Kong.

Challenges for Efficient and Seamless Space-Terrestrial Heterogeneous Networks

Jihwan P. Choi and Changhee Joo

Jihwan P. Choi is with Daegu Gyeongbuk Institute of Science & Technology.

Changhee Joo (corresponding author) is with Ulsan National Institute of Science and Technology.

This work was supported in part by the Daegu Gyeongbuk Institute of Science and Technology through the MIREBraiN program and in part by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT & Future Planning (No. NRF-2014R1A1A1A05 002557).

Abstract

By interconnecting satellites in the sky with ground stations on earth, space-terrestrial networks can extend the coverage area and increase the throughput for both commercial and non-profit applications. To realize the seamless and economic sky-to-earth networks, we still have to advance the technologies that bridge different types of channels, and find killer-applications that are essential for our life. In this article we review the current state-of-the-art technologies of multibeam satellites and relaying as fundamental cornerstones, and then address the issues of low-latency random access, mobility support, and multipath protocols for realizing a seamless heterogeneous network. We then illustrate two impactful applications: for emergency communication as a near-future and non-profit application, and for simultaneous transmission of information and power as a long-term application.

Introduction

The capabilities of satellites can be enhanced when they are interlinked with ground components, which can fill some coverage holes and improve system throughput. The non-line-of-sight (NLOS) environments inside buildings or underground where satellite signals are not easily received can be covered by relaying temporarily built WiFi/Bluetooth or partially working 3G/LTE signals to satellite networks. When the network architecture interconnects the different components over the sky and on the ground with the system heterogeneity taken into account, the performance of space-terrestrial "heterogeneous" networks can be significantly enhanced, especially for disaster recovery and rescue coordination. In addition to the traditional application for information transmission, the simultaneous transmission of information and power from satellites can allow uninterrupted connectivity to isolated areas. Space-based solar power transmission from satellites has been a long-term project and is finally planned to be realized in a few decades [1]. Space-terrestrial heterogeneous networks can feed solar power from space to large rectennas (rectifying antennas) located in unpopulated areas, and then distribute to each user through terrestrial power line or wireless links.

In this article we investigate the enabling technologies for multi-purpose space-terrestrial networks that achieve both high spectrum efficiency and low latency. We explore the technical challenges in interconnecting satellites with terrestrial networks such as LTE and WLAN. In particular, the main thrusts to enhance the economics of the space-terrestrial network are as follow:

- Multibeam satellites that generate agile narrow spotbeams by controlling interbeam interference.
- Relaying architecture that interconnects multibeam satellites and ground stations for the multi-dimensional performance objectives of throughput, delay, power consumption, and reliability.
- Low-latency random access that overcomes long propagation delay with distributed resource allocation schemes in satellite uplink.
- Mobility support at high frequency bands that enhances the economics of satellite communications despite delayed channel state information (CSI).
- Multipath protocols that are optimized in cross layers for quality of service (QoS) guarantees across space and terrestrial network components.

We then discuss a few examples where the space-terrestrial networks can be applied as killer applications in the near term and long term, respectively:

- Emergency communications that are essential for disaster recovery and rescue missions by connecting non-satellite users through (partially working or temporarily deployed) terrestrial networks.
- Simultaneous transmission of information and power that can distribute solar energy from multibeam satellites to the Earth, as illustrated in Fig. 1 for a scheme of emergency networks.

In the rest of this article we focus on the current status and future outlook of the technologies and applications listed above, which range from those being actively studied or currently being implemented (e.g. multibeam satellites, relaying architecture, and emergency communications), to the others being considered for possible use in the future but still to be proven beneficial. To fill the gap between practical importance and future interests of our selection, many ongoing research activities will be presented as well, including interference mitigation techniques, multi-input-multi-output (MIMO) systems, network coding, delay diversity techniques, and advanced diversity/handover concepts for mobility support.

Enabling Technologies

Multibeam Satellites

Multibeam satellites are currently a state-of-the-art technology for commercial satellite networks such as Inmarsat and Globalstar. Narrow spotbeams can project high power density and thus can support high data rates to small user terminals. In addition, an identical frequency spectrum can be reused in multiple distant cells, increasing the total system capacity. An advanced multibeam satellite equipped with a phased array antenna can effectively serve a large number of users over its coverage area by beamforming narrow spotbeams and managing interbeam interference. It is shown in Fig. 2a that phased array antenna techniques such as agile antenna gain patterning and beam scheduling can achieve 74 percent of the maximum capacity while serving interfering users located within 1/4 of the mainlobe beamwidth from each other [2].

The core physical-layer components that realize multibeam satellites are active array antenna and solid-state power amplifiers (SSPA). The phased array antenna (Fig. 2b) can adjust the size and/or shape of a beam by feeding many array elements from SSPA, and linearly superimpose signals by controlling an antenna-patterning matrix. Moreover, the phased array antenna can work together with transmission scheduling switches faster than 1 msec, which is desirable for delay-constrained transmission. Flexible antenna gain patterning allows for simultaneous service of users in a populated area by suppressing possibly significant interbeam interference. The impact of interference can be addressed by a new class of high-layer protocols. Note that the static beamforming with traveling wave tube amplifiers (TWTAs), which require a high power margin and have been widely used in satellite communications traditionally, achieves only a fraction of the throughput achieved by SSPA and phased array antenna, as shown in Fig. 2(a). It is still a challenging research topic to design high-efficiency gallium nitrate (GaN) SSPA on a large scale. Interference mitigation using beamforming techniques [3] and exploiting dual satellites or polarization for MIMO communications [4] are being adopted to improve the efficiency of multibeam schemes.

Relaying

Relaying is one of the key elements of 4G/5G cellular systems as it can extend the coverage outside the cell-edge and take traffic load off the macro base stations. The seamless relaying through heterogeneous radio access technologies has yet to be realized in practice due to the difficulty of controlling different protocols through different channels. The benefits and difficulties of realizing relay networks are equally applied to the space-terrestrial network. The overall system throughput of multibeam satellites can be enhanced when ground stations such as gateways, gap fillers, or feeder antennas are also utilized for path diversity. If packets have to be retransmitted, the ground stations can fulfill the request, eliminating long propagation delays from satellites.

The frequency bands of satellite-to-earth and terrestrial links may overlap entirely or partially, in which case the allocation of the shared spectrum should be coordinated not to degrade the entire system performance. As cellular communication technology advances rapidly and the demand for spectrum increases unexpectedly, the standards community has been paying more attention to spectrum sharing (for example, around 2.1 GHz) between satellites and terrestrial cellular networks. Since onboard resource management has so far only been analyzed without taking terrestrial resources into account, resource allocation of the space and ground components should be jointly coordinated. By incorporating ground stations to receive satellite signals, the multibeam satellite has a choice of routing signals to gateways/feeder antennas and/or sending directly to end user terminals. A cross-layer approach takes into joint account the PHY layer rate/power allocation, the MAC layer user scheduling, the network layer routing path selection, and the transport layer congestion control. The scheduling scheme seeks a trade-off between the use of satellite and terrestrial resources, and compensates for the weakest link in the space-terrestrial heterogeneous network. For the heterogeneous network to guarantee seamless service in practice, more robust integration of heterogeneous channels should be realized, e.g. hybrid automatic repeat request (HARQ) packet retransmission based on punctured codes in the terrestrial link.

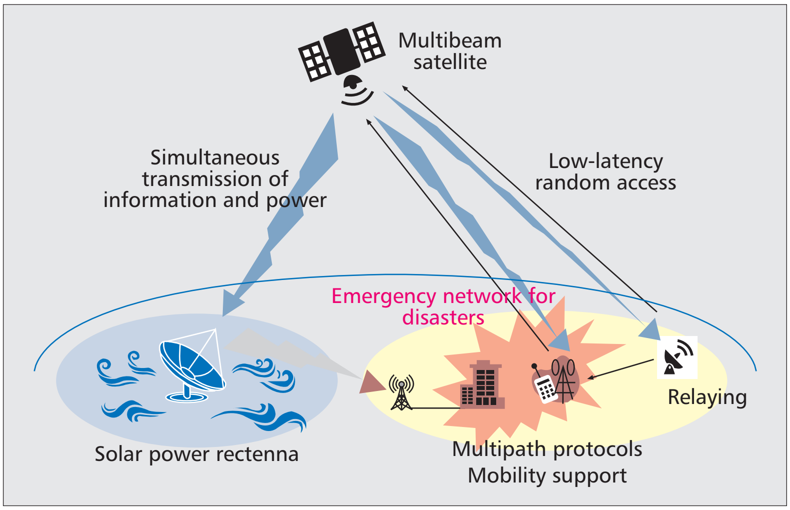


Figure 1. Scheme of emergency networks with simultaneous transmission of information and power.

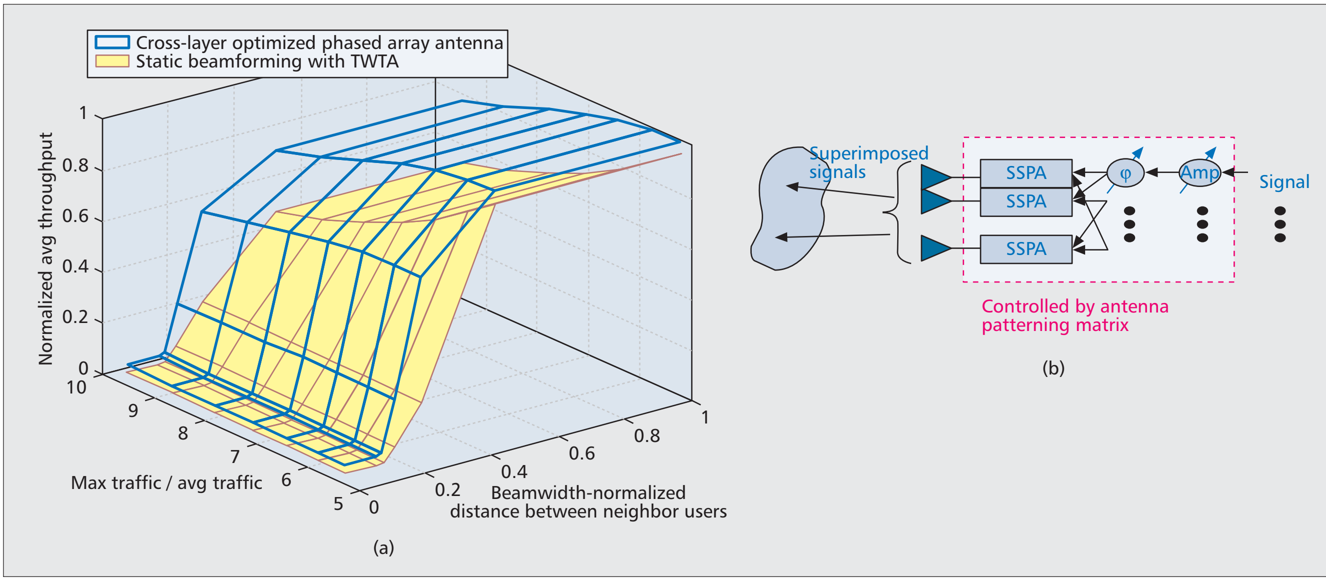


Figure 2. (a) Performance and (b) diagram of phased array antenna (adopted from [2]).

Low-Latency Random Access

In the satellite coverage area, many users compete for uplink access to the satellite transponder. In random access such as the Aloha protocol, users rely on randomness of their signal transmissions in a distributed manner, achieving low latency because data can be transmitted without waiting for the resource allocation from the central coordinator. On the other hand, in heavily crowded communication systems, collisions from random access can cause severe throughput degradation and additional delay. A potential solution would be the use of carrier-sensing functionality as in terrestrial networks, called carrier sense multiple access (CSMA). Under CSMA, each user monitors the status of the channel before its signal transmission, and starts transmission only when the channel is unused. A hybrid approach that combines the reservation scheme is also possible. For example, on recognition of an idle channel, a user can send a short packet ahead of real data and reserve longer slots for the data, obtaining higher throughput without excessive waste of resources due to collision.

A critical weakness of the CSMA approach in the satellite network is again the excessively long delay for signal propagation. When a user senses an idle channel, the information is already stale and its signal transmission may collide with others when it eventually arrives at the satellite antenna, leading to a failed attempt for reservation. One way to tackle the issue is to develop random access schemes based on each user's understanding of the impact of its transmission on the others instead of making reservations. For example, each user measures the signal to interference ratio (SIR) over downlink channels, infers the uplink SIR under the assumption of channel reciprocity, and dynamically adjusts its random access time, backoff parameters, and transmission power. Since many users on Earth will access a much smaller number of satellites in a distributed manner, more rigorous coordination of random transmissions and power control for the system stability is necessary, imposing additional control overhead and latency for information exchange. Otherwise, as in the heavy load scenarios of Aloha networks, multiple transmissions will severely interfere with each other and a significant amount of resources will be lost. It is worth noting that a joint design of admission control and random access could be beneficial to guarantee QoS and fairness among users.

Mobility Support at High Frequency Bands

For cross-layer optimization of the satellite network, it is critical to input the accurate PHY information, such as channel state information and geographical user locations, into high layers and to adapt high-layer protocols according to time-varying satellite channel conditions and user mobility. In an example of successful mobility management, the location information update is decoupled with handovers that can be frequent for non-geostationary satellites, thus reducing the amount of information uploaded to the satellite [5].

As for channel state information, channel modeling for high frequency band communication satellites has so far been mainly focused on weather impairments, such as rain attenuation and atmospheric scintillation. The coherence time of weather-induced channel variation is in the order of seconds and longer, and does not cause severe performance degradation to prediction of the channel states and the corresponding system adaptation. On the other hand, channel variation due to multipath and mobility is much faster than the round-trip delay. It is known that the capacity of the time-varying channels can increase only if the channel feedback delay is within the channel correlation time [6]. The performance of satellite-terrestrial heterogeneous networks at higher frequencies can be enhanced by adding mobility support and system adaptabil-

		LOS (line-of-sight) environments	NLOS (non-line-of-sight) environments
Fixed users	Channel characteristics	Slow time-varying (channel coherence time > 1 second)	
		Channel modeling of weather-induced impairments using log-normal and Rician distribution	Not a major interest (If necessary, bad channels in the Markov channel modeling)
	Communication strategies	Channel prediction and system adaptation	
Mobile users	Channel characteristics	Fast time-varying (channel coherence time < 1 millisecond)	
		Channel variation mainly due to Doppler effects	Most vulnerable channels due to multipath, shadowing, blocking, and Doppler effects
	Communication strategies	Path diversity and co-operation, such as HARQ and multi-layer coding, using terrestrial networks	

Table 1. Fixed and mobile user support comparison for LOS and NLOS environments.

ity to fast time-varying channels, for which the current practical system design does not fully make up.

It is still an open problem how to overcome the difficulty of predicting fast time-varying channels, and to develop a new class of protocols for adapting systems accordingly. Not only the technical challenges, but low demands for deploying satellite networks in the urban and/or high-speed environments have been obstacles. Another bottleneck is the speed and accuracy of the current beam switching technology. Unless the advanced phased array antenna is used, provisioning agile spotbeams to mobile users in the urban channel condition is not a simple task even if the perfect channel state information is known to satellites. Thus, instead of channel prediction and adaptation as for weather-induced channels, it has been suggested to provide an option of more redundancies, so that users in fast time-varying channels can take advantage of them in the case of poor channel conditions, but do not have to do so otherwise.

Retransmissions via ground stations or advanced coding schemes such as multi-layer coding [7] can improve the supportability of mobile users and mitigate the blocking/multipath effects in the form of retransmitted packets and the longer coded packet length. The network coding technique can be combined with ARQ protocols to reduce delay further and increase throughput [8]. The new protocols in the MAC layer will incur the proper modification in other layer protocols that have been optimized only for slowly time-varying weather-induced channels. For example, by considering potential retransmissions, more packets are forwarded to ground stations than when redundancies are not taken into account. At the same time, since the delay-sensitive packets should be headed to users directly, multi-path routing and the corresponding multi-path transport layer protocol should be deployed for network performance optimization. Diversity and co-operation using terrestrial stations are the promising solution for adding mobility support to high-frequency satellite networks despite incurring redundancies and coordination/synchronization between satellites and Earth stations. If the co-operation with terrestrial stations is enough to support user mobility on the fast time-varying channels, the system performance degradation due to the imperfect CSI of the fast time-varying channel can be alleviated since the measurements and modeling of the satellite channels at high frequencies are extremely time-consuming. Table 1 summarizes the current technologies and future trends for fixed and mobile user support in LOS and NLOS environments, respectively.

Multipath Protocols

Multipath protocols are intended to support diversity/co-operation and low-delay transmission in the space-terrestrial heterogeneous network that operates different types of protocols over multiple signal paths with different channel characteristics. The current MP-TCP (Multipath Transport Control Protocol) has been developed to support parallel TCP subflows and shows a drastic change in performance based on multipath characteristics. Compatibility with legacy single-path TCP is one of the key issues in the community. To prevent starvation of legacy TCP, a MP-TCP connection is supposed to not overwhelm a legacy TCP in each path, while its overall performance is better than a single-path TCP in the best-quality path [9]. To exploit multi-path diversity while providing backward compatibility, it is imperative to couple transmission window sizes between subflows. The problem becomes more challenging when multiple paths have significantly different characteristics.

It is still an open problem how to overcome the difficulty of predicting fast time-varying channels, and to develop a new class of protocols for adapting systems accordingly. Not only the technical challenges, but low demands for deploying satellite networks in the urban and/or high-speed environments have been obstacles.

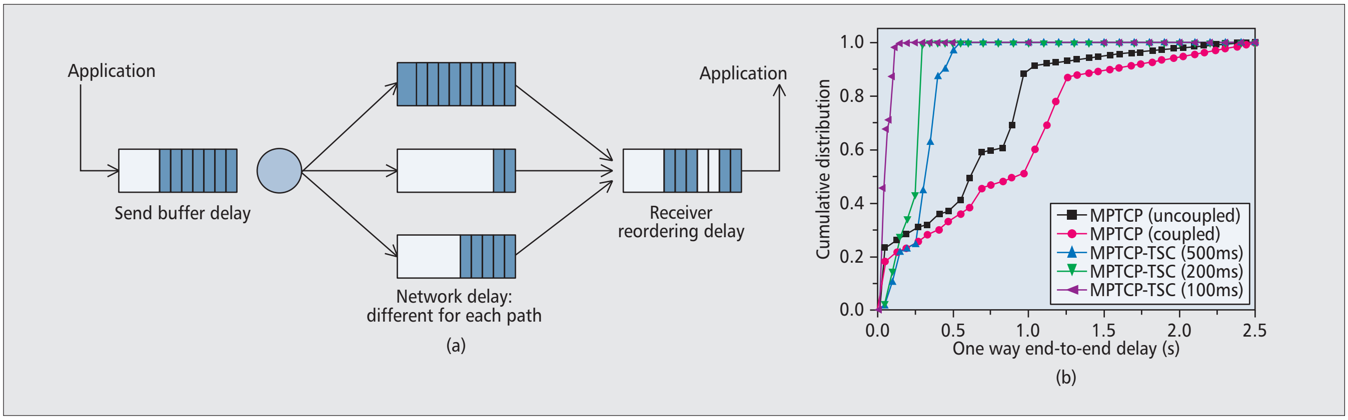


Figure 3. (a) Latency model of MP-TCP and (b) comparison of MP-TCP schemes (adopted from [10]).

In particular, if one of the subflows is through Long-Fat Networks (LFN) such as satellite networks, detecting changes in system dynamics takes more time, which often leads to a waste of scarce wireless resources and degradation of throughput performance.

Guaranteeing delay performance becomes harder due to complicated interplays between the sender, intermediate routers, and the receiver. Since a single flow is split into multiple subflows, its packets not only experience different network delays, but also have to wait for previous packets at the reordering buffer of the receiver, as shown in Fig. 3(a). Recently, by controlling packet transmission times based on queuing delays, channel conditions, and propagation delay, it has been shown that a modified version of MP-TCP can reduce the end-to-end delay at a reasonable user cost, as shown in Fig. 3b [10]. The development of a multipath protocol should be focused on both frontiers of extending the existing protocols to the multipath environment and developing a new class of protocols optimized for heterogeneous types of channels. Since the space-terrestrial network has extremely distinct characteristics for each signal path, we first have to analyze the asymmetry of multipath in terms of channel fading, propagation delay, and available bandwidth. The optimal control of uplink/downlink traffic should be able to compensate for the path heterogeneity and consider transmitter/receiver buffer occupancies as well. However, since the multipath control may impose additional complexity on already sophisticated resource management or result in unnecessary packet transmissions, a careful tradeoff study between performance and complexity is required.

Applications

Emergency Communications

One applicable scenario to use a space-terrestrial network for emergency communications is when a disaster recovery team is deployed for rescue coordination and relief efforts in the area where terrestrial infrastructure is heavily damaged. The recovery team sets up mountable antennas that can play the role of gateways or relay stations between the satellite and end user terminals. The satellite plays the role of a core network while the temporarily deployed or partially working terrestrial links relay satellite signals to surrounding areas including NLOS environments. Each end user terminal for a recovery team member can also receive the satellite signal directly from the satellite if the terminal supports the dual mode at the same time. User mobility support and low-latency random access will be indispensable for receiving direct satellite signals and sending direct requests to the satellite if the member's mission is time-sensitive or the mission area is too wide to be covered by gateway stations.

Co-operation between space and terrestrial networks has attracted much attention as a promising technology for emergency communications. For example, the WISECOM (Wireless Infrastructure over Satellite for Emergency Communications) project in Europe uses Inmarsat and DVB-RCS (Digital Video Broadcast-Return Channel via Satellite) to backhaul terrestrial traffic from 3G/4G, WiFi, TETRA (Terrestrial Trunked Radio), etc. [11]. A simple amplify and forward scheme with diverse delay paths can be shown to improve bit error rate (BER) performance by introducing additional frequency selectivity in the NLOS channel [12].

The path diversity of the multibeam satellite can be exploited to improve reliability in emergency communications. However, the multipath-support protocols such as network coding and MP-TCP face challenging problems of choosing a better signal path, optimizing coding block size, considering transmission diversity for higher reliability, and segmenting a packet for better throughput. If multimedia packets can be received from other members or headquarters with guaranteed QoS levels, a rescue member can carry out missions more efficiently and understand potential dangers more easily. Since the terrestrial ad-hoc network can be used for direct communications among rescue members, the satellite signals should be scheduled not to interfere with (and also not to be interfered by) the terrestrial signals if they operate in the same or adjacent bands. The problem of user scheduling and resource allocation for emergency communications can be formulated to give answers to other questions, such as the location and the number of ground gateways, the single or dual mode of receiver terminals, and so on. Since it is critical to minimize power consumption and save battery power in a disaster where power distribution infrastructure does not function at full capability, the power consumption of satellites and ground stations should be the major consideration for the CAPEX (capital expenditure) and OPEX (operation expenses) of the emergency network.

Simultaneous Transmission of Information and Power

Wireless power transfer technology has recently drawn the attention of communications and energy experts since it can alleviate the power shortage problem of communication nodes and devices. However, the state-of-the-art practical application is limited to a very short range transfer (less than 1m) only using resonant inductive coupling, and ongoing research for solar power transmission has primarily focused on the mechanical and physical technologies, such as very wide solar panels, satellite body structure, and a large scale of rectennas [13]. On the other hand, simultaneous wireless information and power transfer (SWIPT) [14] is also under active research in academia since it can ensure the timely delivery of information under harsh environments. For example, in disaster scenarios, power stations can also be damaged and the rescue effort can be restrictive due to unreliable communications. Since a short halt of communications can be critical for the mission, simultaneous transmission of power and information in the space-terrestrial network will be extremely beneficial, as illustrated in Fig. 1. Although wireless power transfer technology is not close at hand in practice, the development of the high-layer architecture and protocols for simultaneous transmission of information and power should be accompanied to advance the realization of crucial applications such as emergency networks.

When power transmission from satellites is mathematically modeled and practically implemented, we have to compare common and different features between information and power transfer. Signal attenuation due to free-space loss is governed by the Friis transmission equation in the same way as for information. The power transmitted from the Ka band GEO satellite is attenuated by more than 200 dB by free-space loss only. Assuming rectenna receiver gain up to 130 dB with a wide diameter of 3 km, the beamforming gain of nearly 80 dB at the transmitter with a diameter on the order of 10~100 m would be preferred. Since the rectennas for solar power are currently planned to be located in unpopulated areas such as deserts and oceans due to safety reasons, multipath effects are not considered to be a major concern. Rain and moisture in the atmosphere are still big causes for power attenuation, making it hard to utilize high frequency bands. Channel prediction and system adaptation techniques should be more aggressive for power transmission. Nevertheless, a simple method of shutting down power transmission to an area with a bad channel condition can still improve the transmission efficiency. The trade-off between power and information transfer in terms of bandwidth allocation, which is shown in the short-range resonant coupling [15], is also applied to the long-range transmission. The wider the bandwidth, the bigger the information capacity, but power transfer is less efficient. High-efficiency wideband antennas for power transmission and reception can be made of GaN solid-state power amplifiers, which is one of the major research areas for solar power transmission. Table 2 compares the differences between information and power transmission.

	Information transmission	Power transmission
Free space loss	Proportional to the squared distance	
Transmitted power level	~ 1kW	Up to 1GW
Satellite antenna size	< 10 m	~ 100 m
Ground station antenna size	~ 10 m	~ 3 km
Mobility support	Yes	No
Power attenuation	Rain, atmospheric moistures, multipath, shadowing, blocking, Doppler shifts	Rain and atmospheric moistures
Power amplifiers	TWTA, SSPA	Klystron, magnetron, SSPA (made of GaN)

Table 2. Comparison of information and power transmission.

A multibeam-satellite resource allocation problem for simultaneous transfer of information and power to the terrestrial relaying network should be set up for the maximum power transfer under the constraint of guaranteeing the minimum information transmission. Power transmission spotbeams should have maximum power constraints, which are not imposed for information transmission beams unless TWTAs are used. Due to health, safety, and environmental concerns, the amount of received power should be controlled under an allowable threshold. By extending the result on relaying information transmission via ground stations, the optimal solution for power transmission should also be based on channel conditions, user demand for power and information, QoS requirements, and interference among spotbeams themselves and with terrestrial radios. For the feedback of such information from/to ground stations, low-latency random access and multipath protocols will play the key role for information delivery in a timely and reliable manner. Further, for unmanned devices and robots that conduct critical missions in harsh environments, mobility support will be crucial in simultaneous transmission of information and power.

Formulation and modeling can be an initial step toward realizing solar power transmission to Earth in the near future. Following more detailed performance analysis of satellite resource alloca-

	Emergency communications	Simultaneous Tx of information and power
Multibeam satellites	Diversity and increased throughput	Simultaneous delivery with diversity
Relaying	Coverage extension and reliability	Feasible location for rectennas
Low-latency random access	Time-sensitive responses and feedback	Interactive communications for smart distribution
Mobility support at high frequencies	Direct reception for time-sensitive packets	Direct power reception of mobile end users
Multipath protocols	Reliable delivery of timely information	Efficient resource allocation using alternative paths

Table 3. Use of technologies for applications in the space-terrestrial heterogeneous network.

tion, we will need to design efficient control and feedback signals from ground to satellite, ultimately interconnecting solar power satellites and terrestrial smart grids for direct power reception of mobile end users.

Conclusion

Starting from the basic building blocks of multibeam satellites and relaying technology for space-terrestrial heterogeneous networks, this article addressed technical challenges of developing advanced technologies of low-latency random access, mobility support over fast time-varying channels, and multipath protocols. High-impact applications of emergency communications and simultaneous information/power transmissions can benefit from each technology, as summarized in Table 3. We believe that the space-terrestrial network will broadly impact different industry sectors of the commercial Internet, remote sensing/monitoring, disaster response, green energy distribution, and so on, and be of invaluable significance for geopolitics, diplomacy, and military defense.

References

[1] S. Sasaki, "It's Always Sunny in Space," *IEEE Spectrum*, May 2014, pp. 38–43.

[2] J. P. Choi and V. Chan, "Resource Management for Advanced Transmission Antenna Satellites," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, March 2009, pp. 1308–21.

[3] G. Zheng, S. Chatzinotas, and B. Ottersten, "Generic Optimization of Linear Precoding in Multibeam Satellite Systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, June 2012, pp. 2308–20.

[4] P. Arapoglou *et al.*, "MIMO over Satellite: A Review," *IEEE Commun. Surveys & Tutorials*, vol. 13, no. 1, First Quarter 2011, pp. 27–51.

[5] H. Tsunoda *et al.*, "Supporting IP/LEO Satellite Networks by Handover-Independent IP Mobility Management," *IEEE JSAC*, vol. 22, no. 2, Feb. 2004, pp. 300–07.

[6] H. Viswanathan, "Capacity of Markov Channels with Receiver CSI and Delayed Feedback," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, Mar 1999, pp. 761–71.

[7] J. Arnau, A. Rico-Alvarino, and C. Mosquera, "Adaptive Transmission Techniques for Mobile Satellite Links," *Proc. 30th AIAA Int'l. Communications Satellite Systems Conf.* (ICSSC), 2012.

[8] S. Sorour and S. Valaee, "A Network Coded ARQ Protocol for Broadcast Streaming over Hybrid Satellite Systems," *Proc. IEEE 20th Int'l. Symposium on Personal, Indoor and Mobile Radio Communications* (PIMRC), Sept. 2009.

[9] A. Ford *et al.*, "Architectural Guidelines for Multipath TCP Development," IETF RFC 6182, March 2011.

[10] S.-Y. Park *et al.*, "Impact of Traffic Splitting on the Delay Performance of MPTCP," *Proc. 2014 IEEE Int'l. Conf. Communications* (ICC), June 2014.

[11] M. Berioli *et al.*, "Aerospace Communications for Emergency Applications," *Proc. IEEE*, vol. 99, no. 11, Nov. 2011, pp. 1922–38.

[12] S. Morosi, S. Jayousi, and E. Del Re, "Cooperative Delay Diversity in Hybrid Satellite/Terrestrial DVB-SH System," *Proc. IEEE Int'l. Conf. Communications* (ICC), May 2010.

[13] S. Sasaki, K. Tanaka, and K. Maki, "Microwave Power Transmission Technologies for Solar Power Satellites," *Proc. IEEE*, vol. 101, no. 6, June 2013, pp. 1438–47.

[14] K. Huang and V. K. N. Lau, "Enabling Wireless Power Transfer in Cellular Networks: Architecture, Modeling and Deployment," *IEEE Trans. Wireless Commun.*, Feb. 2014, pp. 902–12.

[15] P. Grover and A. Sahai, "Shannon Meets Tesla: Wireless Information and Power Transfer," *Proc. 2010 IEEE Int'l. Symposium on Information Theory* (ISIT), pp. 23632367, June 2010.

Biographies

Jihwan P. Choi [S'00-M'06] received his B.S. degree from Seoul National University, Seoul, Korea in 1998, and the Ph.D. in EECS from Massachusetts Institute of Technology, Cambridge, MA, USA in 2006. From 1999 to 2006 he participated in the Next Generation Internet (NGI) satellite network project at MIT. From 2006 to 2012 he was with the Wireless R&D Group at Marvell Semiconductor Inc., Santa Clara, CA, USA, for mobile system design and standardization of WiMAX and 3GPP LTE as principal system engineer. He is now an assistant professor at Daegu Gyeongbuk Institute of Science & Technology (DGIST), Korea. His research interests are in cross-layer system design for space/wireless networks and wireless power transfer technologies.

Changhee Joo [S'98-M'05-SM'13] received the Ph.D. degree from Seoul National University, Korea. From 2005 he was with Purdue University and the Ohio State University, USA. In 2010 he joined Korea University of Technology and Education, Korea. He is now with Ulsan National Institute of Science and Technology (UNIST), Korea. His research interests include resource allocation in wireless networks, network optimization, wireless sensor networks, and Internet protocols. He is an editor of the *Journal of Communications and Networks*, and has served several primary conferences as a committee member, including IEEE INFOCOM, ACM MobiHoc, and IEEE SECON. Dr. Joo received the IEEE INFOCOM 2008 best paper award and the KICS Haedong Young Scholar Award (2014).

Design Challenges in Contact Plans for Disruption-Tolerant Satellite Networks

Juan A. Fraire and Jorge M. Finochietto

ABSTRACT

During the past 20 years, space communications technologies have shown limited progress in comparison to Internet-based networks on Earth. However, a brand new working group of the IETF with focus on DTN promises to extend today's Internet boundaries to embrace disruptive communications such as those seen in space networks. Nevertheless, several challenges need to be overcome before operative DTNs can be deployed in orbit. We analyze the state of the art of effective design, planning, and implementation of the forthcoming network communications opportunities (contacts). To this end, different modeling techniques, system constraints, selection criteria, and methods are reviewed and compared. Finally, we discuss the increasing complexity of considering routing and traffic information to enrich the planning procedure, yielding the need to implement a contact plan computation element to support space DTN operation.

INTRODUCTION

The Internet has enabled seamless, transparent, and heterogeneous communication, migrating centralized functions toward scalable and efficient distributed systems such as modern banking or education platforms. Back in the 1960s, the concept of the Internet grew out of military studies on how to build robust networks. As a result, addressing and routing were decentralized; but most important, the primary purpose of the network was to remain connected beyond any catastrophe. Unknowingly, the Internet inherited an *end-to-end connectivity* paradigm that shaped modern networked communications including the popular TCP/IP protocol stack. In this context, and under different orbital environments, satellite networks intended not to be the exception.

Traditional geostationary Earth orbit (GEO) satellite relay systems implement bent-pipe repeaters to transmit from one location on Earth to the satellite and back to another location on Earth. Therefore, GEO relays are appealing for broadcasting information to a large geographical area; however, when considered for bidirectional and interactive data communication, challenges such as long round-trip times (RTTs) and frequent channel disruptions must be addressed. This effect is even more dramatic in deep space (DS) systems as longer distances provoke higher delays and severe disruptions due to planet rotation, making permanent and conversational Internet-like communications infeasible [1]. For example, TCP simply cannot tolerate the 20-min propagation time a signal might take from Earth to Mars. On the other hand, disruptions are the prevalent effect in low Earth orbit (LEO) satellite systems. In order to provide voice services, an Iridium satellite constellation system had to be designed to sustain stable end-to-end multihop paths in a highly dynamic and extensive topology. To this end, connectivity was achieved at the expense of a highly complex, expensive, and controversial system. Most recently, the ambitious Defense Advanced Research Projects Agency (DARPA) F6 distributed spacecraft architecture project, which aimed at deploying an Internet-like mesh LEO network, was canceled due to significant increase in budget and overall complexity.

Whether by system complexity and cost in LEO, or physical infeasibility in GEO and DS systems, the end-to-end connectivity paradigm has proven to be hard to adapt to the space environment. In this context, TCP/IP-based Internet will always impose a frontier on networks challenged by delay or disruption, commonly known as delay or disruption tolerant networks (DTNs) [2]. DTN for space applications has been under discussion at the core of the Consultative Committee for Space Data Systems (CCSDS) alongside Internet standards during the past 20 years. Still, the recent formation of a DTN Working Group (DTNWG) in the Internet Engineering Task Force (IETF) might relax the original end-to-end connectivity axiom, allowing DTN applications to finally be embraced by the Internet architecture with all the benefits it entails.

In this scenario, Internet data would be routed through network nodes not necessarily having end-to-end connectivity with the final destination. Indeed, connectivity among nodes (i.e., contacts) could be sporadic but predictable due to orbital mechanics. Nodes would then store and carry these data until forwarding opportuni-

The authors are with the Universidad Nacional de Córdoba — CONICET.

In contrast to most terrestrial applications, space-oriented DTN behavior can be predicted in advance, enabling unique network planning and design opportunities further discussed below.

ties (contacts) are available. Since not all feasible contacts may be required to route data, previous works have investigated topology design issues related to providing connectivity among nodes at the lowest cost (e.g., minimum number of contacts) considering the time-evolving nature of the contact topology [3]. However, resource constraints (available transponders, power consumption, etc.) also need to be considered in the selection of these contacts. Besides connectivity, other criteria related to capacity and fairness are also required to select those contacts that, still satisfying a given set of constraints, can provide the best operational performance. We refer to this problem as contact plan design (CPD), where the contact plan (CP) is the resulting set of contacts that complies with restrictions and maximizes some performance metrics. Among the challenges that need to be addressed before DTN services can be fully implemented, the design of CPs is a critical problem since satellite networks have limited resources (transponders, power, fuel, etc.), which have to be considered. Recently, the design of CPs with resource constraints was proposed and investigated in [4].

In this article, we provide a detailed description of the CPD problem, and a comparison of the performance and complexity of existing solutions. In the next section we provide an overview of the DTN architecture in the space environment and discuss the CPD problem following that. Next we bring an example network under the design process to illustrate the benefits and drawbacks of the proposed CPD mechanisms. Finally, we conclude our work and set future work directions.

DTN Overview

Delay or disruption tolerant networks have received much attention during the past few years as they have been proposed for several environments where communications can be challenged by latency, bandwidth, errors, or stability issues [2]. Originally studied to develop an architecture for the interplanetary Internet (IPN) [1], DTNs have also been recognized as an alternative solution for building future satellite applications [5]; in particular, to cope with typical intermittent channels of LEO constellation systems [6]. Furthermore, DTNs have also been considered for underwater sensor networks, battlefield networks, unmanned aerial vehicle (UAV) communications systems, and connectivity in developing areas, among others.

DTNs overcome the problem of channel delays and disruptions by using a store-carry-forward message scheme. This is analogous to postal systems where messages are stored in a given place (node) until able to move (forward) to another one before reaching its final destination. Internet nodes are designed with little memory (buffer) since they can likely deliver data to the next hop after a routing decision has been made; however, DTN nodes require persistent storage as the link to the next hop might not be available for a long time. As a consequence, Internet communications can be seen as a particular case of DTN with insignificant delay or disruptions.

Among the efforts to implement practical DTNs, the definition of a new communication protocol that does not assume end-to-end connectivity between source and destination nodes has been addressed by the specification of the Bundle protocol in RFC 5050, also resulting in the availability of several software implementations of the protocol, NASA's Interplanetary Overlay Network (ION) [7] being the most popular for space-borne applications. ION was one of the first DTN-capable protocol stacks successfully tested in space in the DINET mission [8]. In contrast to most terrestrial applications, space-oriented DTNs' behavior can be predicted in advance, enabling unique network planning and design opportunities further discussed below.

DTN for Satellite Networks

Technology advances, electronic miniaturization, and the industry of smaller launchers are enabling new business cases for cubesats projects like QB50, EDSN, PlanetLab, and even small-satellite companies like SkyBox (recently acquired by Google) to deploy large LEO constellations systems in the next few years. Therefore, LEO constellations promise to become the first large-scale networks to shift the current monolithic paradigm to a more efficient, scalable, and distributed Internet-like approach. Instead of continuous end-to-end connectivity, DTN would provide these constellations with an effective way to transport data in a store-carry-forward fashion not only by Earth-satellite links (ESLs) but also via sporadic inter-satellite links (ISLs). However, given the degree of conservatism in the traditional space industry, research effort is mandatory as DTNs must still go under severe scrutiny before being considered for large-scale deployments.

In particular, LEO systems are challenged by channel disruptions rather than by propagation times, which are similar to the delays experienced on TCP/IP Internet applications. However, in contrast to the Internet, space-borne DTNs' behavior is under management of a mission operation and control (MOC) center that can deterministically predict (by means of orbital mechanics and communications models) the expected contacts among nodes. Indeed, a contact can be defined as the opportunity to establish a temporal communication link among two DTN nodes when physical requirements are met (antenna pointing, received power, etc.). Henceforth, ISLs are solely thought of as point-to-point, disregarding shared medium access schemes, which fail to perform properly in extensive networks as they assume physical adjacency of many nodes. The latter is either unlikely in a free-flying constellation or demands strict flight-formation requirements to the satellite attitude and orbital control system (AOCS). Furthermore, DTN architecture can handle routing on higher layers, enabling simpler communications architectures, especially if mission requirements can be met in a disruptive scenario.

Figure 1a illustrates the concept of contact with a 4-polar-orbit-satellite (98 inclination angle and 650 km height each) DTN example network we use throughout this article. This scenario is of particular interest for Earth observation missions as satellites account for maximum distance in populated areas while approaching each other in the poles. In these areas, contacts become feasible between adjacent spacecrafts, producing a train-like formation where two directive point-to-point antennas (placed in front and back) can optimize the link budget, producing longer contacts. Henceforth, the network iterates among these contacts, but for the sake of simplicity we base further analysis solely on the half-orbit topology interval. Furthermore, contacts with ground stations are disregarded on the example, but should be transparently considered as another node with which to communicate.

The set of all feasible contacts within a topology interval in a given DTN network can be defined as the *contact topology*. However, it is possible that conflicts or constraints (interference, power restrictions, etc.) need to be addressed before committing the set of planned contacts to the network. As a result, the set of forthcoming contacts to be finally implemented in the network can be defined as the *contact plan* (CP), which is a subset of the original contact topology [9]. Henceforth, the process of selecting the definitive contact set is referred to as *contact plan design* (CPD). As it is typically assumed that all potential contacts between DTN nodes can belong to the CP, the design of CPs has thus far received little attention. However, this problem quickly becomes nontrivial in large-scale systems, and detrimental for resource-constrained scenarios such as satellite missions. Therefore, applying efficient CPD procedures can significantly improve the performance of large DTN satellite constellations.

The Contact Topology Model

In order to tackle the contact plan design problem, a topology modeling technique needs to be specified. Consider the four-satellite network example shown in Fig. 1a. The time evolving nature of these contacts can be captured by means of graphs, with vertex and edges symbolizing nodes and links, respectively. In other words, this representation can be thought of as a finite state machine (FSM) where each state is characterized by a graph with arcs, in turn, that represent a communication opportunity during a period of time (i.e., a contact). Each state can be identified by $k = 1, 2, \ldots, K$ conforming K graphs comprising the same set of nodes but different arcs among them. Particularly, in the suggested scenario, three states can describe the contact topology, which represents the communications evolution during half an orbit topology interval. The FSM model of the example network is illustrated in Fig. 1b.

In particular, a contact topology consists of $p_{k,i,j}$ links between nodes i and j at state k, where $p_{k,i,j}$ may adopt an integer identifying the communication interface (antenna). If no contact is feasible, $p_{k,i,j} = 0$, while $p_{k,i,j} = a$ if the contact among i and j is possible through interface a. Furthermore, at state $k = 1$, $p_{1,2,3} = p_{1,3,2} = 0$ since no physical link exists between N_2 and N_3. In general, the contact topology can be defined

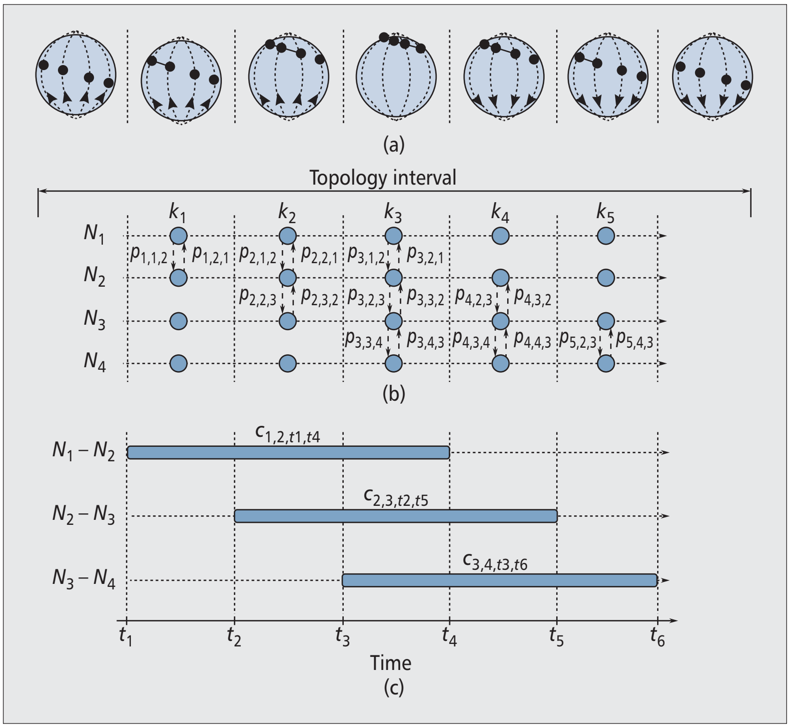


Figure 1. a) Example of a DTN satellite network modeled with b) FSM; c) CL.

by a three-dimensional physical adjacency matrix, from which the contact plan can be designed by removing $p_{k,i,j}$ edges.

Alternatively, a topology can be represented by a contact list (CL) where each contact is in the form of source, destination, start time, and stop time (as in $C_{1,2,t1,t4}$). Therefore, the example network basically consists of three contacts: N_1 to N_2 from t_1 to t_4, N_2 to N_3 from t_2 to t_3, and N_3 to N_4 from t_1 to t_4. The CL modeling for the example topology, illustrated in Fig. 1 c, is more compact than the FSM since it can be expressed as a contact table instead of an adjacency matrix. As a result, CL is the format adopted by ION [7] DTN stack implementation for CP distribution and storage. However, for CP design and engineering, the FSM model granularity might turn out to be convenient to work with, especially when applying mixed integer linear programming (MILP) optimization techniques [9]. Furthermore, as we do in the next section, the FSM model can benefit from discrete state fractionation in order to provide a more detailed and precise topology description. No matter which modeling technique is chosen, translation between FSM and CL is straightforward.

Contact Plan Design

In the initial phase, communications subsystem attributes, including transmission power, modulation, bit error rate, and so on, and orbital dynamics such as position, range, and attitude (orientation of the spacecraft and antenna in the inertial system) can be used to determine the feasibility of future contacts that will form the aforementioned contact topology. This technique is no different from how single-spacecraft missions currently determine space-to-earth contact

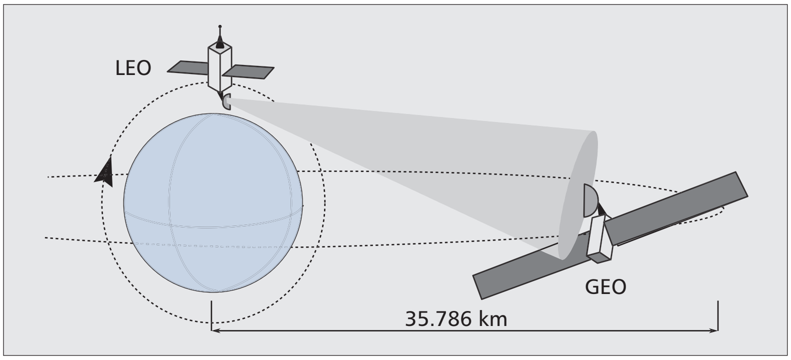


Figure 2. Interference to GEO ssatellites.

opportunities. Nevertheless, the contact topology at this stage does not necessarily encompass all system restrictions or constraints. For instance, interference generated to and from other space assets can turn a given contact unfeasible. In addition, a node may have potential contacts with more than one node at a given time but be limited to only make use of one of these opportunities due to conflicting resources. These conflicts comprise node power budgets or architectural limitations typically found on spacecraft operating in the harsh space environment. As a consequence, further work is required to design a contact plan that considers these scenarios. Finally, once the contact plan is designed, it has to be distributed throughout the network to let nodes execute the contacts as planned. The frequency and mechanisms of contact plan distribution, as well as the topology interval length, are topology-dependent and remain an open research topic.

Design Constraints

In general, we can classify the contact topology constraints in two groups: one representing those that render a particular contact (in a given timeframe) infeasible, and another that limits the number of contacts a DTN node can simultaneously support. We name the former *time-zone constraints* (TZCs) and the latter *concurrent-resources constraints* (CRCs), where both can relate not only to communications but to general system operations issues.

Time-zone constraints: In general, TZCs are those that can forbid communications in a specific geographical area or time for interference or other agency-specific reasons. As LEO constellations systems basically orbit over a wide area of Earth regions, complying with international regulations can be challenging. Moreover, as shown in Fig. 2, since ISLs in LEO constellations are held tangentially in respect to the Earth surface, GEO satellites can be interfered when LEO nodes orbit in the pole area. In particular, some GEO satellites are to be specially considered as they support manned mission communications such as the International Space Station. As a consequence, a proper irradiation policy must be considered not to generate (or receive) interference beyond the International Telecommunication Union (ITU) normative [10].

Furthermore, many other agency-specific reasons might exist for irradiating in a particular geographical area. This can be addressed by time and zone constraints that prevent the availability of a contact in the corresponding CP in a given interval. It should be noticed that these strategies can be considered as the network tolerates disruption, while systems like Iridium have had to request a complete band allocation to sustain interference-free end-to-end communications.

In general, TZCs can be applied directly to the contact topology structure by disabling the conflicting contacts. For instance, if considering the unlikely interference from the short-range antennas of the example topology, contacts from the k_3 state in the FSM model could just be removed. However, interference is not only measured in signal energy, but also in the percentage of time it reaches the interfered-with node [10]. Also, satellites might exhibit energy constraints limiting the fraction of time a transponder can be used. Therefore, there is the possibility to select which contact to disable and when, derived in a combinatorial problem similar to those found on CRC constraints.

Concurrent-Resources Constraints — CRCs are not as straightforward as TZCs as they usually involve the spacecraft architecture and resources. They end up defining the quantity of simultaneous communications (i.e., contacts) a DTN node is able to establish. Consider the architectures of Fig. 3 that apply to the example topology. In Fig. 3a, a simple power splitter divides the transponder signal energy to the two antennas of the spacecraft, while in Fig. 3b a power switch concentrates all the power in one of the available antennas. In either of the aforementioned architectures, only one contact can be established at a given time (i.e., belong to the CP) even if more are feasible through each antenna. A much more complex architecture is shown in Fig. 3c, where two simultaneous contacts are implemented by two cooperative communications subsystems. The latter is the only possible architecture if a non-DTN solution is adopted for the example network, but at the expense of further requirements over the platform power subsystem and weight budget, among others.

In general, CRCs require a selection process. To illustrate the latter, suppose that the example satellite network makes use of the architecture shown in Fig. 3b. In the contact topology of Fig. 1b, a decision must be made for N_2 and N_3 at k = 2, 3, and 4 in the FSM model. Indeed, two possible CPs are illustrated in Figs. 4a and 4b. If the first CP is chosen, the network will provide maximum overall contact time, while if the second one is selected, a more fair and connected network is obtained. Both solutions are defined as feasible CPs the network can implement with the specified resources, but they honor different selection criteria: overall throughput or link assignment fairness.

It should be noticed that while being useful for an illustrative example with two feasible solutions, as more nodes, antennas, transponders, or longer topology interval, CRCs are derived in a nontrivial combinatorial problem with exponentially increasing complexity that a network planner must solve before defining the final CP. Furthermore, CPD might be mandated by more complex selection criteria that take into account not only single-hop consideration as in the present example, but also multihop routing path (slashed arrow in Fig. 4), or even the user traffic that is expected to flow on the system.

Selection Criteria

In general, the problem of CPD lies in selecting from among those contacts that satisfy the communications opportunities represented in the contact topology, and at the same time fulfill TZCs and CRCs. Figure 5 illustrates this group classification. If no constraints exist, the feasible CP solution space expands to that of the contact topology, implying that all combinations are valid for the final network. Also, the original contact topology might reside in the feasible CP space, meaning that it can be directly used for configuring the system without changes. However, the most common scenario requires a selection among the possible CPs, for which a criterion must be defined. The latter is mandatory if network planning automation is required as for large-scale DTN orbiting systems.

In the example topology of previous sections, two initial topology-driven criteria were appointed: maximum contact time and contact assignments fairness. In spite of the fact that these criteria solely depend on topological information, the finest selection can be considered if routing or traffic information is provided. To this end, we provide an overview of different criteria.

Topology-Driven — The topology-driven or single-hop criterion is the simplest and requires only topological information since routing is expected to be dynamically solved in-network. The analysis is solely based on the CP observation, disregarding other system parameters. The most common criteria of this kind are the maximum contact time (MCT) and fair CP (FCP), illustrated in Fig. 4. Models and algorithms are provided in [9] to design the CP with both criteria, proving that MCT delivers CPs with high contact density, but does not necessarily guarantee acceptable network connectivity (as shown in Fig. 4a). On the other hand, FCP prioritizes contacts that are relatively scarce in the topology, providing a fair distribution of them in the final plan. An optimal MILP CPD fairness formulation also exists in [9], authors demonstrate that FCP performs better in routed DTN networks.

Traffic-Driven — A pure traffic-driven criterion is the most controlled scheme and assumes the traffic prediction is fully accurate and can be centrally routed. This implies that the designed contact plan is accompanied by precise and extensive route path information for each traffic data type. Despite this can be challenging, centralized path distribution mechanisms exist such as CGR Extension Block in [11]. With a traffic-driven criterion, the CPD problem can be optimized by means of precise MILP formulations based on the models of [12]. We name this method traffic-driven linear programming (TDLP). On the other hand, a suboptimal yet computationally efficient heuristic mechanism has also been proposed in [4]. This kind of selection criterion provides maximum design optimization and control to the network planner at the expense of flexibility. In other words, if a contact prediction turns out to be inaccurate, or a transponder fails, the traffic-driven selection leaves no place for autonomous network adaptation unless alternative routes are provided.

Besides these selection criteria, others could arise depending on the particular purpose of each network. Since several selection methodologies could be required to find out the most appropriate CP, the complexity of such a procedure might be challenging for network operators. Therefore, we envision a CP computation element (CPCE) that can assist or even automate the design of CPs for future spaceborne DTNs. A CPCE shall be capable of determining suitable CPs to support connectivity among nodes and data transfers through the network.

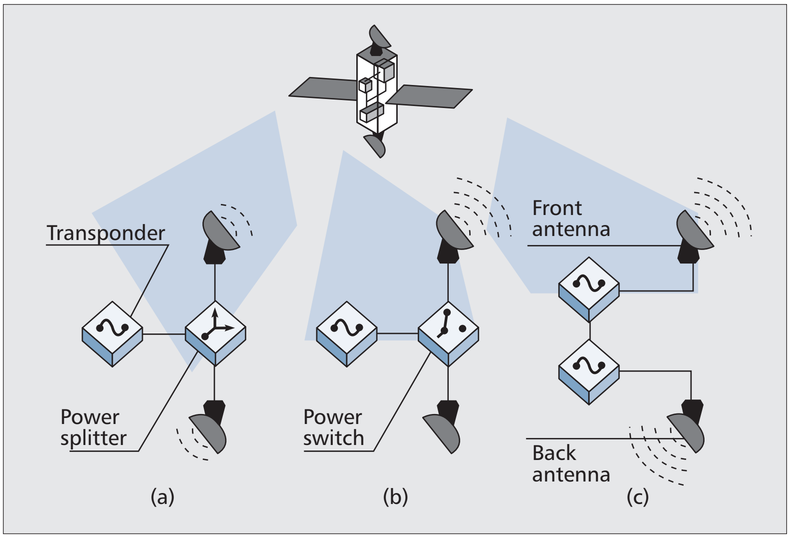


Figure 3. Satellite architecture with a) a power splitter; b) power switch; c) two transponders.

Contact Plan Analysis

In order to evaluate the performance of the existing criteria and solutions described in the previous section, we consider the example topology in the second section. However, we extend the topology interval from the case of half an orbit (illustrated in Fig. 1) up to four orbital periods (spanning a total of 3 h 30 min) in order to better reflect the impact of the CPD mechanisms. Also, in order to allow a higher granularity and accuracy in the design, the topology states K_n longer than 500 s are further partitioned into sub-states. On the other hand, all communications systems are constrained to up to one inter-satellite link per node configured with a 1 Mb/s full-duplex throughput within a 700 km range. The traffic of the scenario is expected to flow equitably from all nodes (N_2, N_3, and N_4) toward N_1, which is expected to deliver the data by means of a space-to-earth high-speed downlink transponder.

In this scenario, we propose to compare topology-driven (FCP) and traffic-driven (TDLP) criteria to illustrate the significance of

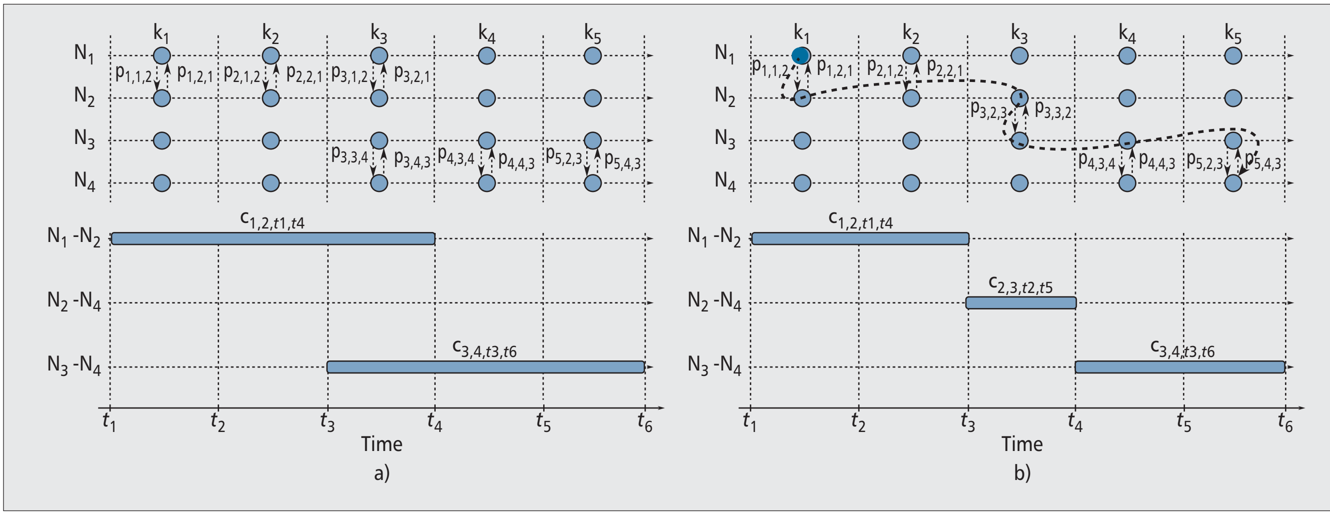


Figure 4. Two possible contact plans: a) maximum throughput; and b) link fairness.

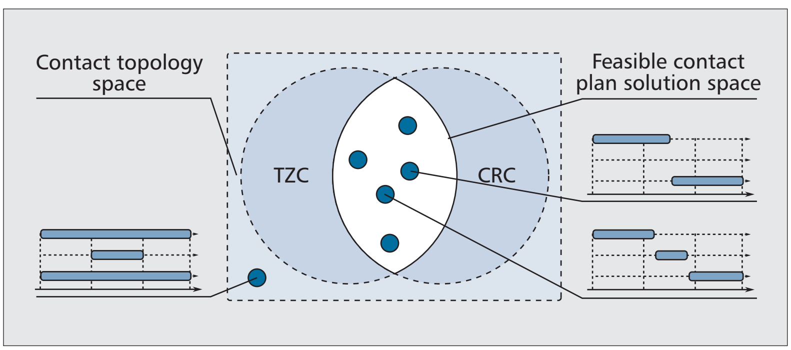


Figure 5. Set of feasible contact plans.

traffic information on the CPD procedures. Besides, we evaluate the physical (PHY) system without any CPD with the aim of illustrating an upper (unconstrained) performance bound. For each of this three methods, we vary the traffic load from R_o = 1 (540 Mb/node) to 0.1 (54 Mb/node), where R_o = 1 is taken as the traffic that saturates one orbit of the unconstrained CP (PHY). In other words, when enabling all feasible links, the R_o = 1 network load can be evacuated to N_1 within a single orbital period. The general hypothesis is that when CPD is necessary (i.e., constraints are considered), the data delivery time will be degraded, especially the measured in FCP as this procedure ignores specific traffic information for contact selection. Besides delivery time, we also evaluate the total contact time as the total accumulated communication time the contact plan uses until all traffic finally reaches N_1.

Results for this scenario are plotted in Fig. 6. It is interesting to note that for R_o = 1, the PHY CP can accommodate all traffic within the first orbit period. As the single-interface constraint is taken into consideration by FCP and TDLP, the delivery time (Fig. 6a) increases with different proportions. As expected, TDLP delivers a better CP in terms of delivery time as it takes advantage of the traffic knowledge the CP is expected to serve. Furthermore, since TDLP contacts are chosen as part of a path toward a traffic destination, the CP is more energy-efficient (outperforming PHY), hence minimizing the total contact time. On the other hand, based only on topology parameters, FCP shows an accumulated effect as R_o increases, penalizing the delivery time up to the fourth orbit of the system. Also, the FCP CP indicates several unused contacts that are not considered for routing the traffic to N_1, which, combined with a high delivery time provokes an excessive total contact time metric (Fig. 6b).

Conclusions

Delay and disruption tolerant networking is emerging as an extension to the current Internet architecture capable of implementing effective networked communications in satellite systems. Among the many benefits, accounting with an autonomous DTN distributed framework can significantly enhance and accelerate the deployment of reliable modern LEO constellations. Furthermore, embracing network disruptions provides an unsought flexibility in traditional Internet-based applications; however, it requires of novel, reliable, and validated mechanisms before being considered for large-scale deployments.

Contact plan design methodologies can be used as a means to optimize limited resources available on satellite networks, and eventually exploit the predictability of these networks. To this end, different modeling techniques, system constraints, selection criteria, and methods were reviewed. We demonstrate that effectively facing the contact plan design can be rewarding but increasingly complex as more information is considered in the planning stage.

Finally, due to the complexity of the design process, we envision the development of CPCEs that can support future operations of future space DTN networks by periodically delivering contact plans that can optimize the performance of these networks. A first commercial-grade CPCE is currently being implemented under the supervision of the authors.

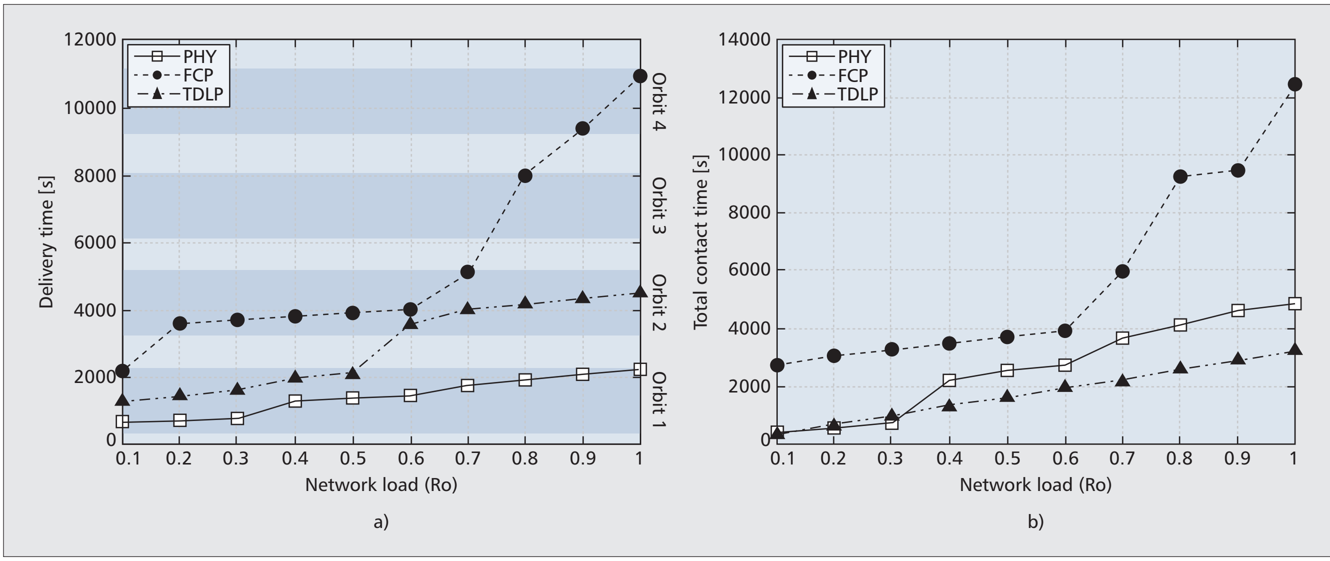


Figure 6. Performance of CPs designed with FCP and MILP model.

References

[1] S. Burleigh *et al.*, "Delay-Tolerant Networking: An Approach to Interplanetary Internet," *IEEE Commun. Mag.*, vol. 41, Jun. 2003, pp. 128–36.

[2] K. Fall, "A Delay-Tolerant Network Architecture for Challenged Internets," *SIGCOMM Proc.*, Karlsruhe, Germany, Aug. 2003, pp. 27–34.

[3] Huang *et al.*, "Topology Control for Time-Evolving and Predictable Delay-Tolerant Networks," *IEEE Trans. Computers*, vol. 62, no.11, Nov. 2013, pp. 2308–21.

[4] J. Fraire, P. Madoery, and J. Finochietto. "Routing-Aware Fair Contact Plan Design for Predictable Delay Tolerant Networks," *Elsevier Ad-Hoc Networks*, vol. 25, Feb. 2015, pp. 303–13.

[5] C. Caini *et al.*, "Delay and Disruption Tolerant Networking: An Alternative Solution for Future Satellite Networking Applications," *Proc. IEEE*, vol. 99, Nov. 2011, pp. 1980–97.

[6] C. Caini and R. Firrincieli, "DTN for LEO Satellite Communications," *Personal Satellite Services*, Springer, vol. 71, Feb. 2011, pp. 186–98.

[7] S. Burleigh. "Interplanetary Overlay Network: An Implementation of the DTN Bundle Protocol," *Proc. 4th IEEE Consumer Commun. and Networking Conf.*, Las Vegas, NV, Jan. 2007, pp. 222–26.

[8] J. Wyatt *et al*, "Disruption Tolerant Networking Flight Validation Experiment on NASA's EPOXI Mission," *Proc. IEEE Advances in Satellite and Space Commun.*, Comar, July 2009, pp. 187–96.

[9] J. Fraire, P. Madoery, and J. Finochietto. "On the Design of Fair Contact Plans for Delay Tolerant Networks," *Proc. Wireless for Space and Extreme Environments*, Baltimore, MD, Nov. 2013, pp. 1–7.

[10] ITU-R Rec. S.1325-3, 2003

[11] E. Birrane, S. Burleigh, and N. Kasch, "Analysis of the Contact Graph Routing Algorithm: Bounding Interplanetary Paths," *Acta Astronautica*, vol. 75, Jul. 2012, pp. 108–19.

[12] J. Alonso and K. Fall, "A Linear Programming Formulation of Flows over Time with Piecewise Constant Capacity and Transit Times," tech. rep., Intel, IRB-TR-03-007, June 2003.

Biographies

JUAN A. FRAIRE received his Telecommunications Engineering degree from the Instituto Universitario Aeronáutico and is completing his Ph.D. in engineering sciences at the National University of Córdoba, Argentina. His research focuses on predictable DTN networking topology design for space-oriented applications. He is a communications protocol architect for STI Company, where he is developing DTN solutions for Earth-observation LEO constellation projects defined in the Argentinian Space Agency (CONAE) National Space Plan.

JORGE M. FINOCHIETTO is a full professor at Universidad Nacional de Córdoba and a senior researcher at the National Research Council (CONICET) of Argentina. He has been involved in several national and international research projects in the fields of communication networks. He has coauthored over 50 papers published in international journals and presented at leading international conferences.

Emergency Satellite Communications: Research and Standardization Activities

Tommaso Pecorella, Luca Simone Ronga, Francesco Chiti, Sara Jayousi, and Laurent Franck

Tommaso Pecorella, Francesco Chiti, and Sara Jayousi are with Università di Firenze.

Luca Simone Ronga is with CNIT Firenze.

Laurent Franck is with Telecom Bretagne.

ABSTRACT

Space communications is an ideal candidate to handle critical and emergency situations arising on a regional to global scale, provided there is effective integration among them. The article presents a review of solutions offered by space communication systems for early warning and emergency communication services. It includes an up-to-date review of public research and standardization activity in the field, with a specific focus on mass alert. The main technical issues and challenges are also discussed along with the cutting-edge research from the scientific community.

INTRODUCTION

Climate changes and complex political scenarios have generated unseen contexts for public authorities called to react to emergency situations. Fast chained events require exceptional capacity for monitoring and action, often over wide areas. In the response phase, satellite communication technologies provide operative communications regardless of the availability of regular terrestrial infrastructures. These capabilities are built on three major properties of satellite communications: broadband capabilities with flexible management, inherent broadcasting, and resilience with respect to Earth damage. By taking advantage of these properties in a timely manner, it is possible to effectively apply them to manage the early warning and emergency response phases. With this aim, we present a review of research and standardization activities, specifically focusing on mass alerting. Additionally, we discuss some relevant scientific and technical challenges toward improving the effectiveness of satellite communications.

EARLY WARNING AND EMERGENCY RESPONSE

Satellite communications solutions are deployed at various level of the end-to-end early warning chain. Cosmicheskaya Sistyema Poiska Avariynikh Sudov — Search and Rescue Satellite-Aided Tracking (COSPAS-SARSAT) payloads sent onboard non-geostationary (low Earth orbit SAR, LEOSAR) or geostationary (GEOSAR) satellites are able to detect, locate, and forward emergency signals sent from compliant beacons. LEO satellite constellations, such as Iridium or Globalstar, provide short message services that are used for sending warning messages. For example, the national plan for flood detection and warning in Haiti relies on a combination of general packet radio service (GPRS) and Iridium short burst data message service to send flood detection alerts to a crisis center. Satellite communications can also be used to send warning messages to the actual recipients. In the Haitian flooding warning systems, sirens are triggered through satellites. Finally, satellite backhaul links can be used as backup trunks for critical communications, among them early warning networks.

Early warning systems may benefit from the integration of different heterogeneous space systems. Galileo (Global Navigation Satellite System — GNSS) can provide small data broadcasts to low-cost receivers; real-time sensors over wide areas can be efficiently polled by GEO-based phone systems (i.e., Immarsat, Turaya). A global picture of the evolving context is depicted in Fig. 1. All the main space components (data collection, navigation, interactive data, and broadcast) are here integrated to provide a common early warning system.

When a disaster event strikes, a priority for first responders is to conduct a rapid assessment of the situation and proceed with the initial response missions. Voice and facsimile are favorite bearers of information. Because time is critical, satellite phones and portable satellite terminals, such as the Inmarsat Broadband Global Area Network (BGAN) are the preferred technologies. They are easily available, portable, and provide worldwide voice and limited data capabilities. However, as the operation continues to be rolled out, the intermediate stages of emergency management (e.g., field headquarters) are also deployed with two consequences: the need for long-haul communications shifts from the field to headquarters, and the communications requirements are increased in both volume and services. Satellite very small aperture terminals (VSATs) are then the favorite choice, as they provide broadband capabilities with the capability to backhaul terrestrial traffic coming from professional mobile radio (PMR), cellular, voice over IP (VoIP), and wireless networks.

 0163-6804/15/$25.00 © 2015 IEEE

These satellite communication hubs are often located so that the satellite access resources can be shared among the various organizations involved in emergency response.

General Communications Architecture

Although information broadcasting is a natural feature of satellites, the recent growth of machine-to-machine (M2M) data exchange and massive improvements in space-borne radio links have created the basis for new emergency-related services provided by space systems. These features might be advantageously applied to more complex scenarios for supporting critical missions, improving the flexibility, reactivity, robustness, and effectiveness of intervention by means of an efficient and flexible coordination of all the involved nodes. It may be achieved with ad hoc and application-driven networking, complementing the classical telecommunications infrastructure and possibly involving all the potential actors, which are federated in order to accomplish data gathering to achieve context awareness, processing and dissemination tasks, as well as distributed decision making and reconfiguration of the main parameters of the critical mission. In particular, it may involve heterogeneous satellites together with both manned aerial vehicles (MAVs) and unmanned (UAVs). To generalize as much as possible, the architecture consists of a dynamic topology relying on the following nodes, as explained in Fig. 2:

- Heterogeneous satellites.
- Satellite Earth stations (SESs).
- High altitude platforms (HAPs), which are in charge of improving coverage by interconnecting isolated domains. This may be accomplished by directly providing connectivity to ground forces or forming a backbone so that the resulting topology is a mesh, or also involving intermediate nodes acting as relays.
- UAVs, normally in charge of monitoring and sensing a specific region. They can be isolated or act in a group with predetermined instructions.
- MAVs, which make the decisions and intervene on the basis of the fused and processed information.
- A network control center (NCC).

Research Projects and Standardization Activities

A review of public research and standardization activity in Europe is reported in the following sections.

Public Research Activity

The increasing interest in emergency services provided by space is proven by the growing number of research projects addressing this theme. A non-exhaustive list of ongoing or recently concluded projects is provided below.

Alert4All: The Alert4All (Alert for All, EU FP7 2011-2013) project developed an advanced concept for alert and emergency communications from a European perspective. Alert4All investigated five major investigation areas:

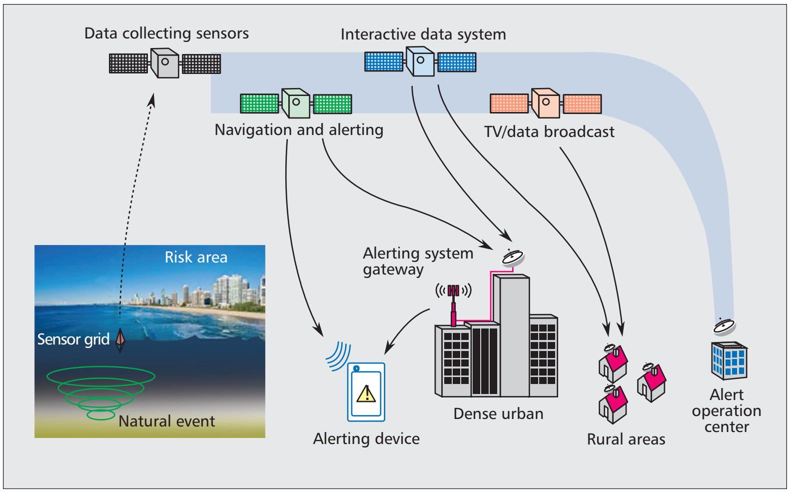


Figure 1. Environmental monitoring and early warning system: architecture overview.

Authorities and Responders Operations, Human Behavior, Role of New Media, Information Management, and Communications Technologies.

MAIA: The MAIA (Mobile Alert Information system using satellites, EU FP7 2008-2010) project is a feasibility study of a satellite-based system aimed at supporting alert and information dissemination using mobile satellite services in S-band over Europe. The MAIA Alerting System provides one-way messaging from civil protection (CP) agencies to citizens, where terrestrial and satellite-based communications are integrated to enable message reception diversity.

CHORIST: CHORIST (Integrating Communications for Enhanced Environmental Risk Management and Citizens Safety, EU FP6 2006–2009) addressed environmental risk management in relation to natural hazards and industrial accidents. The CHORIST project developed a system composed of three subsystems: a risk assessment report subsystem, a warning message dissemination subsystem, and a rapidly deployable telecommunication subsystem.

MASSCRISCOM: MASSCRISCOM (Mass Crisis Communication with the Public, EU FP7 2009–2011), with the general objective of achieving an increased common capability in society to communicate between competent authorities and the public in crises. Participating partners in the study were mainly emergency agencies from northern European countries. The main component devised in the MASSCRISCOM project was the Crisis Communication Centre (CCC).

ALIVE (ESA): The ESA ALIVE (Alert Interface via EGNOS, 2007–2008) concept is a proposal put forward by ESA to provide emergency communications through a satellite-based augmentation system (SBAS, e.g., EGNOS). ALIVE is conceived to act as an interface between the various disaster management centers and the users (general public) in distress. The purpose is to provide users in distress with useful information about the emerging threat, ways to avoid it, and specific rescue measures to be taken.

MLUTB: The MLUTB (Multi-Constellation

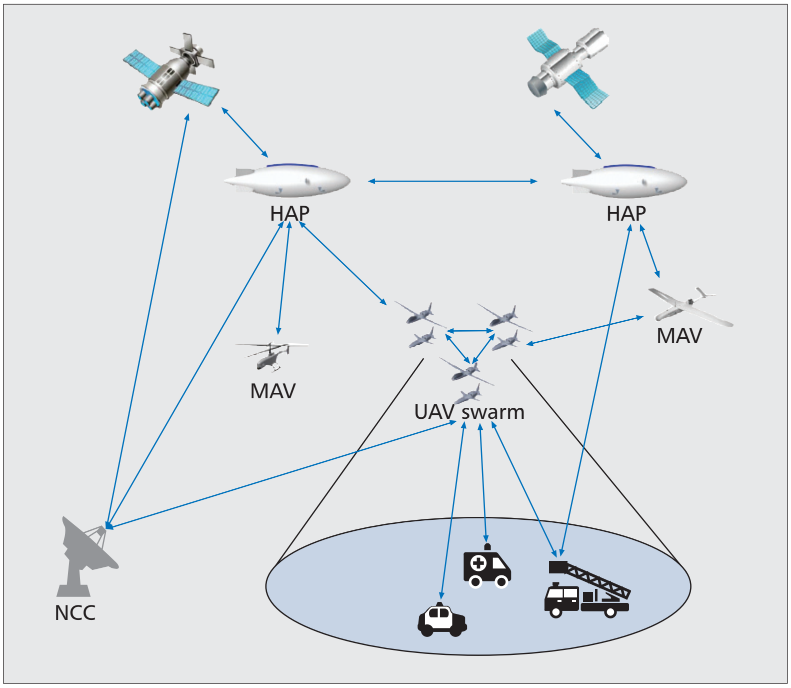


Figure 2. Envisioned integrated communications architecture supporting emergency situations.

Regional System Land Users Testbed, 2010–2011) project is part of ESA's GNSS Evolution program. MLUTB is one of several testbeds intended to support European agencies (ESA and EC), users, and industry in evolving GNSS. The focus of this project is on multi-constellation land users (MLUs), in particular on two different services, the Proof of Position Service (POPS) and Emergency Service (ES).

ABSOLUTE: This project (Aerial Base Station with Opportunistic Links for Unexpected and Temporary Events, EU FP7 2012–ongoing) proposes a rapidly deployable network infrastructure capable of supporting reliable high data rate applications to serve disaster emergency situations. It is obtained through the opportunistic combination of aerial, terrestrial, and satellite communication links with the aim of maximizing network availability, and allowing rapid and incremental network deployment.

EULER: (European Software Defined Radio for Wireless in Joint Security Operations, FP7 2009–2012) aims to define and demonstrate how the benefits of software defined radio (SDR) can enhance interoperability and fast deployment in case of crisis. Considerable effort has been devoted to the definition of SCA-compliant abstraction layers for the integration of heterogeneous components — general-purpose port (GPP), digital signal processor (DSP), and field programmable gate array (FPGA) — for SDR.

WINTSEC: (Wireless Interoperability for Security, Preparatory Activity for Security Research PASR, 2007–2009) is mainly focused on interoperability in public and governmental security (P&GS) systems implementing a "system of systems" approach. WINTSEC explores a mixture of complementary solutions, including SDR, to overcome the barriers for wireless interoperability across different security agencies.

As a general comment, most public research is oriented to the exploitation of available satellite and terrestrial technologies appropriately integrated in a system able to provide a new service. This indeed appears to be the key evolutionary trend in the delivery of emergency applications: no new expensive technologies are strictly required to provide efficient and robust emergency services; instead, a high level of system integration is sufficient. More details on each project can usually be found through the founding institution website (e.g., [1, 2]).

Standardization Activities

A growing number of standardization actions have been initiated by the main institutions involved. Here some of the most relevant bodies and documents in the European context are summarized.

ETSI SES-SatEC: The Satellite for Emergency Communications (SatEC) working group is probably the most focused on satellite applications. Part of the European Telecommunications Standards Institute (ETSI) Satellite Earth Stations (SES), it produced a set of documents related to various aspects of satellite assisted emergency communication services. TR-103-166, "SatEC; Emergency Communication Cell over Satellite," outlines the concept of emergency communication cells over satellite (ECCS). An ECCS is understood as a temporary emergency communication cell supporting terrestrial wireless and wired standards (based on IEEE 802.11, dPMR, IEEE 802.16, 3GPP LTE, or ETSI TETRA), which are linked/backhauled to a permanent infrastructure by means of bidirectional satellite links. The document, besides describing ECCS architectures based on existing products, introduces the challenges in providing interoperable services, and gives an overview of commercial available products and ECCS solutions. TS-103-284 [3] presents a formal classification of ECCS devices taking into account communication, mobility, and energy consumption capabilities, together with robustness and physical features.

TR-102-641, "SatEC; Overview of Present Satellite Emergency Communications Resources," presents an overview of concepts, systems, and initiatives related to the use of space resources in the context of disaster management. The SatEC working group also promoted two EU-funded specific task forces (STFs) devoted to the production of standardization documents in this field. In particular, STF472 is devoted to the definition of reference emergency scenarios, while STF473 will define MAMES, a multi-protocol alert message encapsulation (see below).

ETSI EMTEL: Although not strictly related to satellites, the Emergency Communications (EMTEL) group addresses a broad spectrum of aspects related to the use of telecom services in emergency situations. TR-102-180/182 define the requirements for communication during emergencies among the various entities (citizens, authorities, individuals). TR-102-144 considers the use of short message service (SMS) and cell broadcast service (CBS) for emergency applications, while TR-102-476 presents some solutions for the adoption of VoIP during crises.

IETF ECRIT: The Internet Engineering Task Force (IETF) Emergency Context Resolution with Internet Technologies (ECRIT) initiative, collects several aspects related to the use of Internet services during emergencies. Among them we can mention RFC 5012, "Requirements for Emergency Context Resolution with Internet Technologies", where the baseline requirements are investigated; RFC 5031, "A Uniform Resource Name (URN) for Emergency," leading to a standard resource naming for emergencies; RFC 5069, "Security Threats and Requirements for Emergency Call Marking and Mapping," investigating on VoIP security aspects; RFC 5222, "LoST: A Location-to-Service Translation Protocol," where a context-aware service access mechanism is formulated; RFC 5582, "Location-to-URL Mapping Architecture and Framework," for location aware naming; and RFC 6443, "Framework for Emergency Calling Using Internet Multimedia," where the adoption of multimedia streaming for emergencies is considered.

Emergency Information Dissemination: Systems and Protocols

One of the key requirements for efficient delivery of emergency information (communication from authorities to citizens) is to quickly reach the highest number of people within the affected areas coping with critical operation conditions. Taking into account a common notification service for all types of emergency, [4] reports a collection of operational, organizational, and technical requirements, providing a comparison of the main communication technologies adopted for the dissemination of emergency notifications. Focusing on satellite technologies, the summary below includes an overview of the capabilities of SATCOM/SATNAV systems, highlighting their compliance/non-compliance with some of the main requirements identified in [4] and taken as a reference for the purpose of this article.

Examples of Existing Alerting Systems

Existing alerting techniques rely on basic traditional methods (e.g., loudspeakers, sirens, public displays), broadcast transmission (e.g., paging, radio, and TV broadcast), and personal interactive communications devices (e.g., terrestrial and satellite mobile networks, web/social media), ranging from small local footprint to continental or global reach. However, the use of satellite networks, despite their resilience against terrestrial damage, are currently used mostly for niche applications. Two examples of dedicated alerting systems that use satellite broadcast as dissemination technology for emergency information delivery are the Satellite-Based Warning System (SatWaS) and Modular Warning System (MoWaS).

SatWaS is a German alerting system developed by the Bundesamt für Bevlkerungsschutz und Katastrophenhilfe (BBK) (Federal Civil Protection Agency) to disseminate via satellite urgent alert information in case of major national security incidents or threats. This system is gradually replacing sirens in Germany. However, it does not aim to directly alert the population; in fact, the SatWaS warnings are sent to regional situation centers, public/private media broadcasters, Internet providers, paging services, and press agencies, which in turn are in charge of forwarding the alert messages to the general population.

MoWaS is a modular upgrade of SatWaS, which is currently being developed with the objective of allowing local civil protection authorities to activate all alarm and warning systems in their area of responsibility in a decentralized manner and without discontinuity of media use. Unlike SatWas, MoWaS warnings can directly reach the population via any connected means, and in order to disseminate alerts CBS and Terrestrial Trunked Radio (TETRA) are introduced as communication technologies.

SATCOM/SATNAV Systems for Emergency Information Delivery

The adoption of satellite technology for distributing alert messages allows reliable warning of the population and the authorities about a hazard. From the point of view of the deployment of an efficient alert network, the following categories of satellite systems can be considered: broadcast fixed/mobile satellite system (without return link) and bidirectional fixed/mobile satellite systems. The adoption of the first class of the satellite systems (e.g., DVB-S2, DVB-SH based) enables the alert messages to broadcast over a large area, reaching both home users with traditional satellite TV receivers or additional temporary communication terminals installed for emergency purposes and mobile users equipped with satellite radio and multimedia receivers. On the other hand, the bidirectional feature of the second class of systems and, in particular, the intrinsic robustness of satellite two-way systems allow alert message transmission to users equipped with VSAT terminals and to mobile users with satellite mobile devices subscribed to one of the data services provided by Inmarsat, Thuraya, Globalstar, and Iridium. The bidirectional systems allow the transmission of acknowledgment of the delivery success of the alert message (an optional feature useful mainly for authority-to-authority communications). Among the existing systems, it is worth mentioning Iridium Short Burst Data Service, which is particularly suitable for a global alert message distribution network.

Besides the traditional SATCOM systems, the use of the available data channels of SATNAV systems, such as SBAS and GNSS (e.g., Galileo), has recently been investigated for the provision of alert messages. Combining the mass diffusion of mobile location-aware receiving devices and the location information provided by a SATNAV system, alert messages can be transmitted over a large area or a specific target area, and at the user level the receiving

Combining the mass diffusion of mobile location-aware receiving devices and the location information provided by a SATNAV system, alert messages can be transmitted over a large area or a specific target area and at user level the receiving device could be able to determine based on its current location, whether or not that particular message is relevant to it.

At the physical layer, the exploration of higher frequency bands (W band 80–90GHz) will allow the creation of precise shaped and compact multi-spot antennas increasing dramatically the throughput per unit area. New antenna technologies like meta-materials will also support this trend.

	SATCOM				SATNAV
Requirement	Broadcast fixed	Broadcast mobile	Bidirectional fixed	Bidirectional mobile	GNSS
Mass distribution of alerts	✓	✓	*	*	✓
Reach mobile users	–	✓	–	✓	✓
Adequate capacity (data rates)	✓	✓	†	†	‡
Alert delivery acknowledgment	–	–	✓	✓	–
Legend	✓ = compliant – = non-compliant * = potentially suitable, mainly designed for point-to-point communications † = medium/low data rate ‡ = very low data rate				

Table 1. Satellite systems comparison: SATCOM/SATNAV systems vs. requirements.

device could be able to determine, based on its current location, whether or not that particular message is relevant to it. Although SATNAV systems can be used for alerting purposes, some of their main limitations need to be taken into account in the design of an alert message protocol (available bandwidth, time to alert, etc.). Focusing on the Galileo system, both commercial service (CS) and public regulated services (PRS) could be considered as two main candidates for the delivery of alerts. The former mainly targets the market of commercial applications, and thus can be used for the transmission of alert messages to the population. While the latter is particularly suitable for sensitive applications, since it is restricted to government-authorized users, it can be used for the delivering of alert messages to public protection and disaster relief (PPDR).

Message Format for Exchanging Emergency Alerts

Among existing formats for exchanging alert messages, Common Alerting Protocol (CAP) is currently the international de facto standard for alert messages. It is developed by the Organization for the Advancement of Structured Information Standards) and adopted by the International Telecommunication Union (ITU) as X.1303. CAP is part of the Emergency Data Exchange Language (EDXL) family, which is a suite of XML-based messaging standards aimed at improving emergency information sharing among emergency management entities. It is a simple but general format for exchanging emergency alerts and warnings over different networks. CAP allows all kinds of alerting and public warning systems to be activated, ensuring information consistency over multiple delivery systems and increasing emergency management effectiveness. While a CAP message can also be used to acknowledge reception of an earlier CAP message, the CAP specification does not support or mention any compression schemes. Therefore a typical human-readable XML-based CAP Alert Message has a size of about 1.5 kB, which can be reduced to 1 kB using zip compression. To cope with this issue, which may exclude the transmission of CAP messages over narrowband channels, an efficient compression scheme for XML-based CAP messages was defined in the framework of the Alert4All project as better described in the following subsection.

Alert Message Transmission Protocols

This section gives an overview of the existing initiatives for alert message protocol definition, focusing on a project-driven solution (A4A protocol) and the ongoing ETSI standardization activity within SES/SatEC (Specialist Task Force STF473) for the encapsulation of alert messages (MAMES — Multiple Alert Messages Encapsulation over Satellite). However, for completeness, the following past standardization activities are mentioned: Lightweight Emergency Alerting Protocol (LEAP) and Encoding Secure Common Alert Protocol Entities (ESCAPE), produced by the IETF Authority-to-Citizen Alert (ATOCA) Working Group (whose work concluded in November 2012), and the Post Office Code Standardization Advisory Group (POCSAG), currently used in many countries to distribute urgent information, including emergency and alarm communications.

A4A protocol — Aiming at the transmission of alert messages within size-limited frames of GNSS (e.g., GPS, EGNOS, Galileo) systems (possibly avoiding fragmentation), the A4A protocol consists of processing and translating CAP messages [5]. In detail, only the mandatory fields of a CAP message are encoded into specific protocol header fields. A modular structure of the message allows the definition of new functional blocks according to system requirements, resulting in a flexible protocol toward extensions. Moreover, for an efficient encoding and compression scheme, the use of an alert library structure was considered, and additional functionalities are implemented for fragmentation, priority management, source authentication, security mechanisms (encryption, message integrity), and robustness against link errors.

MAMES Protocol — The main objective of the MAMES protocol comes from the need to transport alert messages over satellite links for fast distribution of an alert to the population [6]. Specifically, MAMES aims to define a flexible encapsulation scheme for carrying diverse alert messages over different satellite communication technologies, which include the Galileo system (CS and PRS services). To cope with the heterogeneity of the alerting devices, which may need to be activated simultaneously, MAMES investigates the possibility of supporting the transmission of multiple concatenated alert messages, formatted according to different alert protocols. Moreover security mechanisms and the optional use of acknowledgments are considered for authority-to-authority communications. To provide a protocol flexible to extensions and enable transmission over narrowband channels such as the GNSS systems, the MAMES message structure is inspired by the packet data unit (PDU) structure of IPv6 (next header concept), where the header consists of a primary header followed by extension blocks. This allows the definition of MAMES mandatory and optional functions related to the message encapsulation and ancillary functionalities, respectively.

Summary: Satellite System and Protocol Comparisons

A comparison among the different classes of satellite systems, which may be adopted for alert messages delivery, is reported in Table 1, which shows the capabilities of each system in terms of compliance or non-compliance with some of the key requirements of an emergency notification service. It is worth highlighting that among the listed requirements, compliance with *mass distribution of alerts* and *adequate capacity* allows the simultaneous distribution of alerts in a short predictable period of time to the targeted large areas (and intended audience). Moreover, the highlighted limited bandwidth capacity of SAT-NAV systems requires the development of efficient protocols opportunely designed to cope with such a constraint. On the other hand, an overview of the previously described protocols for alert dissemination is provided in Table 2, where CAP, A4A, and MAMES are compared in order to highlight their benefits and limitations.

Key feature	CAP	A4A	MAMES
Alert message format	Yes	No	No
Encapsulation protocol	No	Only CAP	Multiple (concatenated) alert messages
Encoding scheme	No	Only for CAP	No
Tx over SAT narrowband channels	Out of scope	Efficient CAP encoding	Within the scope

Table 2. Protocol comparison: CAP vs. A4A vs. MAMES.

Scientific Challenges

Recently, research on emergency communication scenarios is mainly focused on system interoperability, higher frequency band utilization, transmission efficiency, and fast system reconfiguration.

Concerning system interoperability, it may be achieved using different approaches, such as satellite-independent service access points (SI-SAP [7]). Moreover, it is generally assumed that IP represents an unifying technology.

At the physical layer, the exploration of higher frequency bands (W band, 80–90 GHz) will allow the creation of precisely shaped and compact multi-spot antennas, dramatically increasing the throughput per unit area. New antenna technologies like meta-materials will also support this trend.

Another relevant challenge in satellite communications is the improvement of the transmission efficiency. Satellite networks are high delay-bandwidth product networks. The network delay cannot be reduced below a physical limit depending on the satellite orbit, and the jitter caused by packet retransmissions can negatively affect communications performance. In order to enhance protocol performance, several techniques have been proposed, ranging from predictive bandwidth allocation [8] to the use of network coding (NC) principles [9].

NC can be useful in several satellite communication scenarios; for example, when a satellite multicasts a flow to a selected set of receivers, a receiver may decode information from more than *one* satellite, or satellite *trunking* services.

NC was introduced a decade ago in [10] to improve network capacity. It consists of combining several packets using random coefficients in a finite Galois field. It can be done at the physical layer (physical NC, PNC) or at higher layers (digital NC, DNC). As a matter of fact, NC represents a bandwidth-efficient (spectrum and energy efficiency) technique for wireless multicast broadcast services (MBSs) in comparison to traditional automatic repeat request (ARQ) or hybrid ARQ (HARQ) protocols [11].

NC may improve transmission efficiency and reduce delay, even though its optimization is a typical NP-hard problem. Consequently, it requires the development of specific heuristics in order to reduce the number of packet (re)transmissions.

Moreover, NC can be used for strict delay-constrained contents. In this case it is useful to resort to block-based transmission schemes such as random linear network coding (RLNC) [12]. It is worth mentioning that the goal may be to maximize the packet delivery ratio or, alternatively, minimize the block *completion* time for a set of users [13].

Another interesting application field is represented by satellite *relay* networks, where relays may have to provide connectivity to multiple sources. In this case, a promising approach

Environmental changes are leading to ever more frequent natural events with regional (sometimes global) impact on peoples' lives. Space communications have great potential to help in emergency management, even though this potential is still not fully exploited due to limited or even nonexistent integration among different satellite systems.

encourages users to interfere and exploit the interfered signals to increase the network capacity. This can be achieved by using PNC [14], which could potentially double the capacity of two-way relay networks in both synchronous or asynchronous scenarios.

PNC can be effectively used in conjunction with spatial diversity (SD) techniques in order to overcome losses in wireless fading channels, particularly for relay scenarios where multiple sources cooperate and communicate with a single destination through a relay. This represents the typical case of uplink satellite systems, where this approach could reduce the probability that the receiver is unable to collect all the information packets sent by the source [15].

At the IP level, multiple techniques have been proposed to increase the satellite network efficiency. Protocol compression (e.g., ROHC, RFC 5795) may be used, but the system scalability for sporadic connections must be evaluated. The transport layer, and in particular TCP, has been extensively studied. The proposed approaches are to rely on specific TCP flavors (e.g., TCP Hybla, TCP Noordwijk) or use TCP proxies (PEP — Performance Enhancing Proxy, RFC 3135). The two approaches are complementary: satellite-specific TCP (or even specific transport layers, e.g., SCTP) have better scalability, but they are limited to cases where the terminals are aware of the satellite network segment. If the endpoints cannot dynamically adjust the TCP flavor, PEPs are an interesting alternative.

Recent advances in reconfigurable signal processing devices also open the door to software defined radio technology in space. Hybrid computing architectures based on field programmable gate arrays, where radiation hardened and redundant schemes are jointly employed to provide a space tolerant functional blocks, allow the adoption of software configuration capabilities for payloads. The advantages are enormous in terms of flexibility and efficiency of the spacecraft that can evolve through its usually long lifetime.

Conclusions

Environmental changes are leading to ever more frequent natural events with regional (sometimes global) impact on peoples' lives. Space communications have great potential to help in emergency management, even though this potential is still not fully exploited due to limited or even nonexistent integration among different satellite systems. Two main contributions can be provided by satellite systems to the management of crises: dramatic reduction of latency in the delivery of alerts to the population, and efficient and robust exchange of information among emergency operators and authorities. Several ongoing activities have been reported, essentially conducted through research projects and standardization activities. The emergency context also represents a great research opportunity for scientists, whose aim is the improvement of quality, efficiency, and availability of emergency communications assisted by satellites.

References

[1] EC CORDIS — Community Research and Development Information Service; http://cordis.europa.eu

[2] ESA; http://www.esa.int/ESA

[3] ETSI, "Satellite Earth Stations and Systems (SES); Satellite Emergency Communications (SatEC); Device classes for Emergency Communication Cells over Satellite (ECCS) ," TS 103 284, Aug. 2014.

[4] ETSI, "Emergency Communications (EMTEL); Requirements for Communications from Authorities/Organizations to Individuals, Groups or the General Public During Emergencies," TS 102 182, July 2010.

[5] T. De Cola, J. Chaves, and C. Niebla, "Designing an Efficient Communications Protocol to Deliver Alert Messages to the Population During Crisis through GNSS," *Proc. Advanced Satellite Multimedia Systems Conf. and 12th Signal Processing for Space Commun. Wksp.*, 6 Sept. 2012, pp. 152–59.

[6] M. Berioli *et al.*, "Satellite Assisted Delivery of Alerts: A Standardization Activity within ETSI" *Proc. 7th Advanced Satellite Multimedia Systems Conf.*, Sept. 2014.

[7] ETSI, "Satellite Earth Stations and Systems (SES); Broadband Satellite Multimedia (BSM) Common air Interface Specification; Satellite Independent Service Access Point (SI-SAP)," TS 102 357 , May 2005; http://www.etsi.org/deliver/etsi_ts/102300_102399/102357/01.01.01_60/ts_102357v010101p.pdf

[8] L. Chisci, T. Pecorella, and R. Fantacci, "Dynamic Bandwidth Allocation in Geo Satellite Networks: A Predictive Control Approach," *Control Engineering Practice*, vol. 14, no. 9, 2006, pp. 1057–67.

[9] R. Yeung, "Network Coding: A Historical Perspective," *Proc. IEEE*, vol. 99, no. 3, Mar. 2011, pp. 366–71.

[10] R. Ahlswede *et al.*, "Network Information Flow," *IEEE Trans. Info. Theory*, vol. 46, no. 4, July 2000, pp. 1204–16.

[11] F. Chiti *et al.*, "Optimized Random Network Coding for Reliable Multicast Communications," *IEEE Commun. Lett.*, vol. 17, no. 8, Aug. 2013, pp. 1624–27.

[12] A. Eryilmaz *et al.*, "On the Delay and Throughput Gains of Coding in Unreliable Networks," *IEEE Trans. Info. Theory*, vol. 54, no. 12, Dec. 2008, pp. 5511–24.

[13] M. Esmaeilzadeh, N. Aboutorab, and P. Sadeghi, "Joint Optimization of Throughput and Packet Drop Rate for Delay Sensitive Applications in TDD Satellite Network Coded Systems," *IEEE Trans. Commun.*, vol. 62, no. 2, Feb. 2014, pp. 676–90.

[14] B. Nazer and M. Gastpar, "Reliable Physical Layer Network Coding," *Proc. IEEE*, vol. 99, no. 3, Mar. 2011, pp. 438–60.

[15] R. Alegre-Godoy and M. Vazquez-Castro, "Spatial Diversity with Network Coding for On/Off Satellite Channels," *IEEE Commun. Lett.*, vol. 17, no. 8, Aug. 2013, pp. 1612–15.

Biographies

TOMMASO PECORELLA [S'90, M'00] received his Dr.Ing. degree in electronics engineering from the University of Firenze, Italy, in 1996, and his Ph.D. degree in telecommunications engineering in 1999. In 2000 he joined CNIT, the Italian University Consortium for Telecommunications, as a scientific researcher. Since 2001 he has been an assistant professor at the University of Firenze. His research interests involve computer communications, mobile communication networks, QoS-enabled access schemes, satellite communication networks, queuing theory, wireless sensor networks, ad hoc routing, network simulations, network management, and security.

LUCA SIMONE RONGA (luca.ronga@cnit.it) [S'89, M'94, SM'04] received his M.S. degree in electronic engineering in 1994 and his Ph.D. degree in telecommunications in 1998 from the University of Florence, Italy. In 1997 joined the International Computer Science Institute of Berkeley, California, as a visiting scientist. In 1999 he joined CNIT, where he is currently head of the research area. He has been Editor of the *EURASIP Newsletter* for four years, a member of the ETSI SatEC working group, and a member of NATO Task Force on Cognitive Radio. His research interests span from satellite communications to cognitive radio, software defined radio, radio resource management, and wireless security. He has been principal investigator in several research projects and AN author of more than 90 papers in international journals and conference proceedings.

FRANCESCO CHITI [M'01, S'13] received his Ph.D. degrees in informatics and telecommunications engineering from the University of Florence in 2004, where he is presently working as an associate professor. His current research topics are devoted to link and network layer protocol design for ad hoc and sensor networks. He has taken part in several European research projects, such as REGPOT AgroSense, IP GoodFood, STREP DustBot, NoEs NEWCOM and CRUISE, GJU TWIST, ETSI STF179, and COST 289 action. He serves as Associate Editor of the *Wiley International Journal on Security and Communication Networks*. He is an author of three book chapters, more than 20 journal papers, and more than 50 conference papers.

SARA JAYOUSI received her Ph.D. degree in computer science, multimedia, and telecommunications in 2012 from the University of Florence. She has been with the Department of Information Engineering of the University of Florence since 2008. Her research activity is mainly focused on satellite communications for emergency, IP QoS network management in hybrid satellite/terrestrial networks, cooperative communications, and diversity algorithms in relaying systems. She is an author of several transactions and conference papers.

LAURENT FRANCK [SM] graduated in computer science (1994) and social sciences (1998) from Brussels University. In 2001, he received a Ph.D. degree in telecommunications from Télécom ParisTech and an Habilitation à Diriger des Recherches from the Institut National Polytechnique de Toulouse in 2009. Since 2007 he has been with Télécom Bretagne (Toulouse site) where he teaches and conducts research on satellite communications. His main research interest lies in the development of satellite-based emergency communications. Laurent is involved in ETSI standardization activities. He has also been a first aid worker and team trainer with the French Red Cross for 14 years.

Alerting over Satellite Navigation Systems: Lessons Learned and Future Challenges

Tomaso de Cola and Cristina Párraga Niebla

The authors are with DLR — German Aerospace Center

Abstract

Public warning in emergency situations has experienced great momentum in recent years due to the increasing need for suitable communication means to disseminate alerts due to the increasing number of natural and technological disasters, as well as the consequent damage created and number of people affected by them. This goes along with a number of activities that promote multi-channel approaches to make use of different communication channels to disseminate alerts. However, only a little work has been performed so far considering the use of satellite communications in the last mile within a multi-channel architecture for alerting purposes. This article addresses the feasibility and impact of deploying alert message services over satellite navigation systems.

Introduction

Public warning in emergency situations has gone through great momentum in recent years. On one hand, the number of natural and technological disasters, as well as the consequent damage created and number of people affected by them, are significantly increasing according to the International Disaster Database (http://www.emdat.be/database). On the other hand, the technologies and methods to alert and inform the population in such emergency situations were outdated compared to the rapidly evolving communications technologies of the last two decades. It is a paradox that a significant part of the population has access to powerful communication capabilities with potential that is not fully exploited for public warning purposes. Hence, several activities have been triggered in recent years to rethink public warning from its foundations and align with contemporary communications networks. However, public warning is typically a national or even regional matter, depending on the governance structure of each country. This yields solutions in different countries that are similar to some extent but uncoordinated and sometimes not fully interoperable.

The publication of the Common Alerting Protocol (CAP) specification [1] and the EU-Alert standard [2] have significantly contributed to draw a path toward interoperability in public warning, but still, the concrete solutions applied in different countries remain dedicated. EU-Alert provides guidance for the use of Cell-Broadcast for warning purposes, which has been deployed in some countries, such as the United States, the Netherlands, and the United Kingdom. In addition, the adoption of a multi-channel approach has become a trend in public warning. By using several communication channels to disseminate the same alert message, the probability that the message will be received by the citizens at risk in emergency situations increases. Some examples are the U.S. IPAWS, the CHORIST project, MoWaS in Germany, eVigilo in Israel, and the NL Alert System in the Netherlands.

A major concern is that these solutions largely rely on terrestrial communication systems that are sometimes affected (partly or completely unavailable) by major disasters. Indeed, satellite communication systems as additional alerting channels would increase the overall availability of a multi-channel public warning system and provide appealing features, such as broadcast capability and robustness against disasters. In this context, the use of satellite-based navigation systems for warning purposes has also been taken into consideration. This idea was first approached in the ALIVE project [3] and further investigated in the Safety-of-Life (SoL) initiative [4], with focus on safety of life applications, such as guided aircraft landing. Smart use of satellite navigation devices would in fact provide the advantages of satellite communication devices (robustness and broadcasting capability) with the added value of location information, an interesting feature that enables the alert message delivery through the device only if the recipient is in the area of risk. The portion of the population that could benefit

0163-6804/15/$25.00 © 2015 IEEE

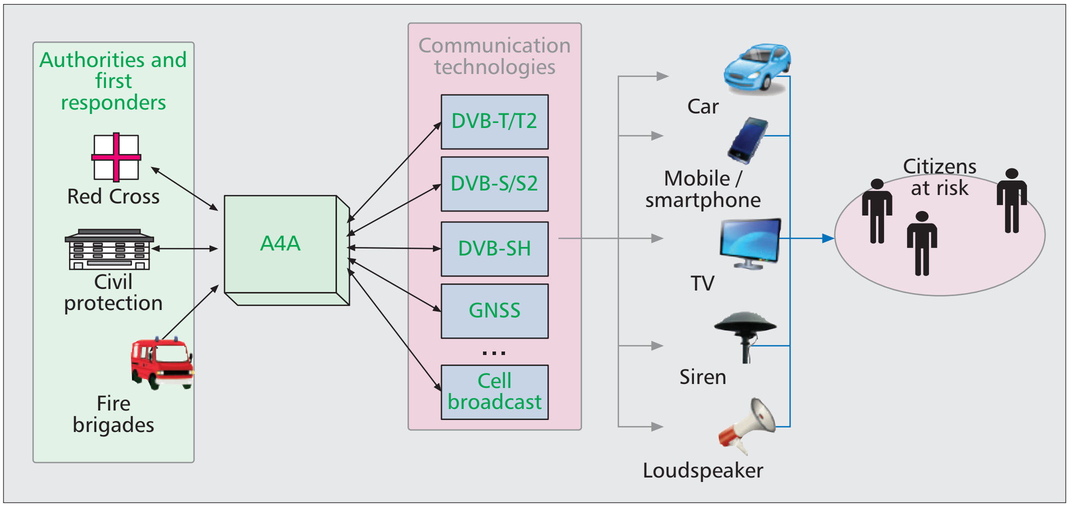


Figure 1. Alert4All integrated multi-channel public alert system.

The term public alert refers to authorities providing meaningful warning information to citizens threatened by a hazard to enable them to prepare, act appropriately, and in sufficient time to reduce the possibility of harm or loss. To this end, devices that are publicly available and personal devices should be envisaged to reach the population at risk.

from it is not negligible considering the increasing popularity of navigation-enabled devices in the automotive and mobile device markets, although receiver software upgrades would be required for the reception of warning messages.

The Alert4All project (A4A, http://www.alert4all.eu) designed, implemented and demonstrated a multi-channel public warning system that makes use of satellite and terrestrial components, as shown in Fig. 1. This article discusses the conditions in which public alert services could be deployed over satellite navigation systems as part of a multi-channel public warning system based on the lessons learned from the analysis made in the A4A project on the feasibility, design, and performance implications.

The rest of the article is organized as follows. The next section introduces the public alert context, while the third section provides an overview on relevant information about satellite navigation services and discusses the related challenges and opportunities. After that, we discuss how alerting services can be provided over navigation systems, considering architectural and traffic implications, which are the baseline to understand the performance simulation results shown later in the article. The final section draws our conclusions.

Public Alert Context

The term *public alert* refers to authorities providing meaningful warning information to citizens threatened by a hazard to enable them to prepare and act appropriately in sufficient time to reduce the possibility of harm or loss. To this end, devices that are publicly available and personal devices should be envisaged to reach the population at risk. As stated in [5], *"assets applied to alert the population in crisis situation shall be operational, robust, and available every minute of every day and tailored to the needs of a wide range of different threats and different user communities ... The warning lead times range from seconds to weeks (depending on the type of hazard) … where common needs exist, it may be possible to make use of the same communications systems for more than one type of warning information."* Indeed, a public alert system should be conceived to cope with any hazard, where the content of the message and the time criticality of the message delivery will depend on the hazard and the overall context. In this regard, it is important to remark that hazards requiring public alert in very short times (in the range of 10 s, e.g., tsunamis) shall also be covered by a public alert system.

A multi-channel approach, as shown in Fig. 1, is important to reach people at risk, regardless of where they are or what their communication habits are. Furthermore, as discussed in [6], different communication channels have different capabilities that can be seen as complementary in different ways, so receiving coherent messages through different channels can significantly improve the overall performance of a public alert system. This article will focus on the alert message dissemination over global satellite navigation systems (GNSS), intended as complementary channels to those also available (e.g., cell broadcast, TV broadcast, sirens, mobile networks) within the multi-channel approach.

Global Navigation Satellite Systems

The GNSS umbrella embraces three main technology classes: the existing and future satellite-based positioning systems (e.g., GPS and Galileo), satellite-based augmented services (SBAS, addressed in the next subsection), and ground-based augmented service (GBAS, out of the scope of the present article). The next generation GNSS system includes the Galileo constellation [7], which will be composed of 30 medium Earth orbit (MEO) satellites and is expected to be fully operational by 2018.

Candidates to Transport Alert Messages

The potential of GNSS [8] for alerting purposes has already been explored in past projects (e.g., ARCTIC, MLU) in the framework of the European GNSS Evolution Programme with the idea of exploiting the broad coverage and reliability together with messaging capabilities associated

Type	Message content	Max refresh time (s)
0	GEO information useless (SBAS test mode)	6
1	PRN mask	120
2-5	Fast corrections	6
6	Integrity information	6
7	Fast correction degradation factor	120
9	GEO ranging function parameters	120
10	Degradation parameters	120
12	SBAS network Time/UTC offset parameters	300
17	GEO satellite almanacs	300
18	Ionospheric grid point masks	300
24	Mixed fast corrections/long-term satellite error corrections	6
25	Long-term satellite error corrections	120
26	Ionospheric delay corrections	300
27	EGNOS service message	300
63	NULL message	NA
Others	Reserved	NA

Table 1. Summary of EGNOS messages.

with positioning services, originally conceived to increase positioning accuracy. Consequently, the means to enable messaging services associated with positioning services are candidates to transport alert messages.

Satellite-based augmented systems (SBAS) are state-of-the-art candidates for the transport of alert messages. SBAS have been designed to improve the geographical coverage and accuracy of the existing positioning services. Their main function is to provide GNSS receivers with corrections to the navigation data, thus making the overall position estimation three times more precise. National and international entities have designed different SBAS to accommodate such positioning function upgrades: Wide Area Augmentation System (WAAS) in the United States, European Geostationary Navigation Overlay Service (EGNOS) in Europe, System for Differential Correction and Monitoring (SDCM) in Russia, and Satellite Navigation Augmentation System (SNAS) in China, just to cite a few of them. Particular attention is devoted in this article to the EGNOS system, taken as our reference SBAS technology.

EGNOS [9] defines three types of services: open service (OS), safety of life (SoL), and EGNOS data service access (EDAS);[1] EGNOS messages are transmitted in the L1 frequency band (1575.42 MHz) with the same modulation scheme as GPS but at a five times higher rate (250 b/s) [9]. EGNOS supports 64 types of messages, whereas only the 17 message types indicated in Table 1 are actually used [10]. Each message is encapsulated in EGNOS frames (250 bits), composed of header (38 bits) and payload (212 bits), and sent periodically, with a maximum refresh time as indicated in Table 1.

Two Galileo services [7] are further near-term candidates to transport alert messages: the commercial service (CS) and public regulated service (PRS). CS [11] is intended to support commercial applications, providing transfer of navigation messages carrying up to 500 bits at a rate of 500 b/s. Encryption mechanisms, managed by service providers, are included to guarantee higher service quality and make the transmitted data accessible only for authorized users subscribed to a specific service. As to the latter, PRS service [7] is characterized by a data rate of 50 b/s and conceived for governmental services run by the appropriate authorities (e.g., police, fire departments), thus implementing sophisticated authentication and encryption mechanisms.

Opportunities and Challenges for Public Warning Applications

Current and future satellite navigation systems offer the capability to implement services that carry short messages for different purposes. This feature together with the robustness and broadcast capability make satellite navigation systems an appealing solution to deliver alert messages to citizens at risk. Nevertheless, messaging services in satellite navigation systems are significantly limited in capacity. At a first glance, this imposes either a significant limitation on the information that can be carried or large delivery delays, both undesired attributes in the context of public warning. Furthermore, the capacity available in satellite navigation systems for messaging services is to be shared by several services, many of critical importance. Hence, the operational conditions in which a public warning service could be deployed deserve special mention.

In state-of-the-art public warning systems, alert messages are typically encoded as CAP messages [1] according to XML or ASN.1 syntax, resulting in messages transporting an amount of data ranging from 500 to 2000 bytes when pure text messages are disseminated, or even larger if encryption and authentication are applied, hence making transport over GNSS challenging (GNSS frame size < 100 bytes).

In response to this problem, recent work in the area of public warning to achieve smart and harmonized formulation of alert messages based on alerting libraries [12] combined with a suitable communications protocol [13] provides a remarkable opportunity to translate alert messages into a code that can fit the capacity of an EGNOS frame. On one hand, lessons learned on best practices in the area of public warning conclude that an alert message must include information about the hazard type, location, time, intensity, and certainty, guidance on protective actions, and issuing source. While it is not realis-

[1] EDAS is a commercial service distributed to subscribers over terrestrial networks and not transmitted over the EGNOS space segment. Hence, it is not considered in the analysis reported in this article, since it can hardly be applied in emergency situations where the terrestrial infrastructure can be disrupted or unavailable.

tic to have a pre-recorded alert message for each possible situation, it is possible to envisage a limited dictionary for each of the information items above. By combining those with some syntax intelligence, any alert messages could be managed by transmitting only the relevant encoded keywords together with metadata. The syntax intelligence can be located at the receiver side to formulate the alert message out of the received encoded keywords and metadata. This concept is extensively developed in [12], proposing the light encoding of keywords and terms of these dictionaries to achieve lightweight transmission of alert messages. On the other hand, the A4A protocol has been developed to transport highly compressed alert messages (size < 50 bytes) applying the concept of alerting libraries in its design [13]. In practice, the A4A protocol header fields correspond to the information items identified above, and their values correspond to the alerting libraries values. By storing alerting libraries at the receiver devices and some intelligence to reassemble the "human readable" message out of the code embedded in the A4A protocol, the receiver device must be capable of delivering the alert message in any language and several delivery modes (e.g., text, speech, or sign language using avatars). This approach makes public warning friendly to people with special needs, even if the transmission is language-agnostic. Furthermore, combining these features with map-based information stored in navigation devices yields even more enriched alert messages that can be delivered indicating the area at risk in the map without transmitting pictures.

It is worth mentioning that both the A4A protocol and the concept of alerting libraries have been validated and demonstrated in the Alert4All project and are currently contributing to a number of new European Telecommunications Standards Institute (ETSI) standards, making tangible the opportunity to use these concepts to alert over EGNOS in an efficient manner.

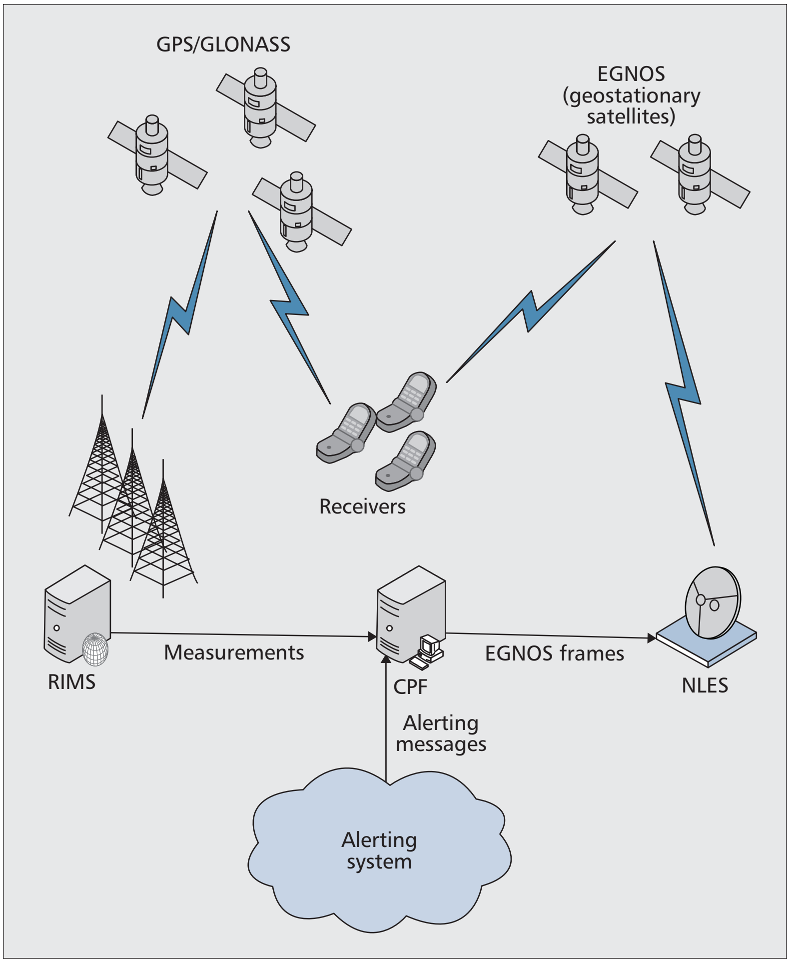


Figure 2. Integrated EGNOS-alerting system.

ALERTING OVER EGNOS

SYSTEM ARCHITECTURE

The EGNOS network (Fig. 2) is constituted by three main actors [10]: ranging and integrity monitoring stations (RIMS), central processing facilities (CPFs), and navigation land Earth stations (NLES). RIMS stations receive the GPS signal and extract the relevant information contained in the navigation message for further processing carried out by the CPF. The latter is actually the entity in charge of computing the differential correction with respect to GPS data, which will be transported in the EGNOS frame. Finally, the built EGNOS frames will be converted by the NLES into an EGNOS signal and transmitted accordingly over the EGNOS satellite system. The integration of EGNOS with public warning systems has to be performed in the CPF, which is the entity generating the EGNOS frames. Hence, the CPF should be in charge of encapsulating the incoming alert messages[2] into EGNOS frames. Fragmentation shall be provided in case the message does not fit in the EGNOS frame. However, assuming that the A4A protocol is applied to transport the alert messages, fragmentation would be only necessary if authentication were applied, resulting in the generation of 2-3 fragments (depending on the length of the digital signature).

EGNOS TRAFFIC CHARACTERIZATION

Particular attention has to be dedicated to the characteristics of EGNOS traffic and the corresponding transmission strategies. For the sake of completeness, the capacity of the EGNOS system has been assumed to be fully utilized in order to account for both open service and SoL applications. The overall EGNOS traffic has been characterized by means of synthetic time-series generated by the SISnet tool [14]. Post-processing of results helped compute the frequency of the various types of messages and the inter-arrival time between two consecutive frames belonging to the same EGNOS service.

According to the computed statistics summarized in Table 2, message types 2–4, 7–10, and 12–17 show very similar characteristics and can therefore be regrouped into larger classes. The other EGNOS frames exhibit very different frequency and average inter-arrival times, thus belonging to dedicated message classes. Furthermore, messages with higher frequency are assigned the smallest refresh time (6 s). This implies that the EGNOS transmitter must enforce specific priority policy to ensure that

[2] It is worth noting that CPFs may not accept alert messages in case of overly high traffic load on the EGNOS space segment to avoid degradation of the native EGNOS services. As such, CPFs should implement call admission control or traffic load policies, which, however, are not detailed here, being outside the scope of the article. In turn, the alerting system is in charge of distributing the alert messages through the other communication channels implemented according to the multichannel approach (TV broadcasting, mobile phones, sirens, etc.)

The performance is even more degraded when downgrading the alert service to class 3, resulting in much larger delivery delays. In this case, both navigation and alert services are not able to meet delay requirement, making this priority configuration not pursuable at any message arrival rate.

these messages (type 2–4) are transmitted within the maximum refresh time. As for the other messages, the constraints are more relaxed. A priority mechanism, however, turns out to also be desirable in this case to prevent other messages being transmitted beforehand, thus resulting in unacceptable overall delivery delays.

We model an EGNOS transmitter as a priority-queue system, where non pre-emptive actions are performed by the transmission scheduler. EGNOS traffic can be modeled as a Poisson arrival process and the service time as deterministic, since the access to the EGNOS channel is performed in time-division multiplexing (TDM) fashion with fixed frame length (250 bits) and data rate (250 b/s), resulting in a deterministic serving time of 1 s.

We define three priority classes according to the maximum refresh time in Table 1:

- Class 1: services with maximum refresh time 6 s
- Class 2: services with maximum refresh time 120 s
- Class 3: services with maximum refresh time 300 s

The transmission of alert messages has to be compliant with this priority strategy, requiring that a new message type be defined and assigned to class 1, 2, or 3. More importantly, the transmission of alert messages has to meet specific system requirements:

- It cannot significantly penalize the delivery of EGNOS service frames.
- Alert messages have to be delivered quickly enough to allow protective actions.

Delay Analysis

In order to prove the feasibility of transporting alert messages over EGNOS under different operating conditions, the queuing system described previously has been implemented in a MATLAB script,[3] taking into account the priority levels assigned to the three aforementioned classes of service [15]. The generation of alert messages is modeled as a Poisson process, with tunable average inter-arrival time. Finally, alert messages are assigned a fixed level of priority and multiplexed with the EGNOS frame belonging to the same class.

Special note must be made regarding the setting of the average inter-arrival time. The frequency of alert message is usually not very high because alert messages are repeated with a frequency on the order of minutes, and then new messages are transmitted. On the other hand, it is worth remarking that EGNOS (and GNSS in general) can be used to transport alert messages from multiple issuers (e.g., in cases where several emergency situations need to be addressed in different regions or countries in overlapping timeframes), thus implying that the overall alert message load can significantly vary. In order to have a complete understanding of the system dynamics and draw correct conclusions about the transport of alert messages over EGNOS, we considered the case of aggregate alert message traffic, wherein the average message arrival rate ranges between 0 and 25 messages/min (the latter describing the case of a large number of risk situations occurring at the same time).

Type	Frequency (%)	Average inter-arrival time (s)
2	26.14	3.82
3	24.88	4.01
4	24.88	4.01
6	0.06	1285.38
7	1.27	79.12
9	1.27	79.12
10	1.26	79.12
12	0.51	194.96
17	0.51	194.96
18	2.55	39.02
25	4.90	20.35
26	11.24	8.89
27	0.51	194.99

Table 2. Statistics of EGNOS messages.

The objective of the conducted simulation campaigns is to evaluate the delivery delay of the alert messages and the impact on the delivery delay of other EGNOS services in different operating conditions.

For the sake of statistical validation, runs consisting of the transmission of 5E5 EGNOS frames were carried out. The A4A protocol was taken as the reference to distribute alert messages, operating without and with security extensions (i.e., authentication).

A4A Protocol without Security Extensions

If the A4A protocol is used without security extensions, a pure-text alert message can fit in the EGNOS frame payload carrying up to 212 bits (~26 bytes).

As a first possible priority configuration, we considered the case where alert messages have the highest priority and are therefore multiplexed with the EGNOS services allocated to class 1. As a result, the traffic load is increased, possibly affecting the performance of the services allocated to the other two classes.

This intuition is confirmed in Fig. 3a, sketching the average sojourn time (delivery delay) of all EGNOS services. At first glance, it can be observed that the services allocated to a specific class present almost identical performance. The delivery delay increases for all services as a function of the alert message arrival rate. In particular, it can be noticed that services allocated to classes 2 and 3 are particularly affected, especial-

[3] The considered queuing system can in principle be modeled as an M/D/1 queue with priority (non-pre-emptive). However, a theoretical mathematical analysis is not possible since the traffic load is sufficiently large to make the system unstable.

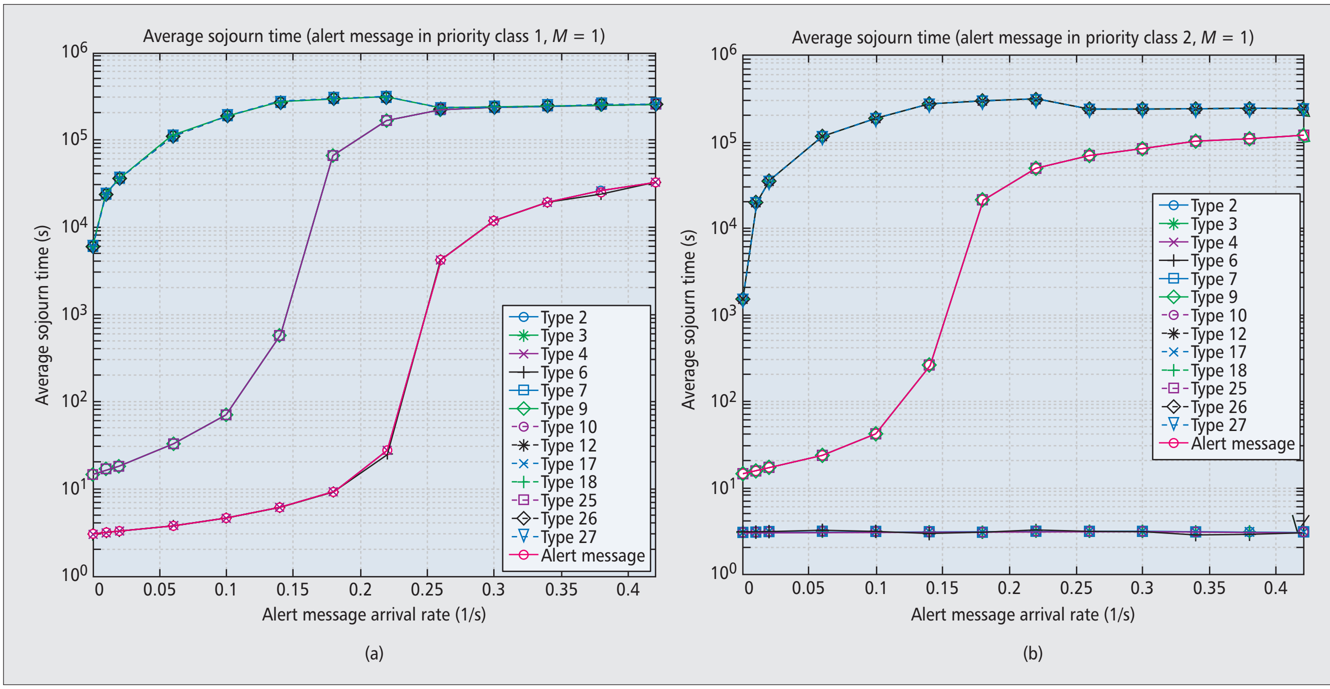


Figure 3. Sojourn time of EGNOS services: a) alert message priority 1; b) alert message priority 2.

ly when the alert message arrival rate exceeds 0.1 messages/s (=6 messages/min). In these conditions, the transfer delay of EGNOS frames jump from about 100 s (class 2) up to nearly 2105 s, actually making management of the traffic load unfeasible.

As far as the alert service is concerned, it can be noted that alert messages with stringent delivery delay (e.g., within 10 s) can easily be managed when the alert message arrival rate stays below 0.17 (~10 messages/min). After this point, the overall performance of class 1 is penalized, meaning that messages allocated to class 1 will be delivered on the order of thousands of seconds. However, it should also be noted that this configuration is able to offer a satisfactory alert service though preserving an acceptable level for native EGNOS service, provided that the alert traffic load is not excessive. In this regard, we can also remark that the identified threshold of 6 messages/min is quite high with respect to common alert management situations where the frequency of message is expected to be lower.

Classifying alert messages in priority class 2 (Fig. 3b) would certainly provide an advantage to EGNOS services with very low refresh time. Predictably, the performance of class 1 is not affected, whereas class 2 is already significantly penalized with moderate values of alert message traffic load (0.12, ~7 messages/min). Furthermore, alert services cannot be used in this configuration for time critical alerts as the delivery delay in the range of 10 s cannot be met, even for very low arrival rates. Nonetheless, it can also be pointed out that emergency services with more relaxed requirements (delivery delay < 300 s, ~ 5 min) can be still be accommodated provided that the overall load is less than 0.14 (~8 messages/min).

Obviously, the performance is even more degraded when downgrading the alert service to class 3, resulting in much larger delivery delays. In this case, both navigation and alert services are not able to meet the delay requirement, making this priority configuration not pursuable at any message arrival rate.

A4A Protocol with Security Extensions

When the A4A protocol is configured with security extensions, the generated alert messages cannot fit into the EGNOS frame payload; hence, fragmentation functions must be invoked. In particular, when source authentication is requested, the A4A protocol message size ranges between 356 and 484 bits (> 212 bits, EGNOS payload size), depending on the length of the specific digital signature being applied (128 and 256 bits, respectively). Upon fragmentation, the original alert messages can be fragmented accordingly into two or three fragments to be transported by EGNOS frames.

In the case of an alert message split into two fragments (Fig. 4a), it is worth considering only the high priority case (class 1), as lower priorities were shown to already penalize the delivery delay of 1 fragment. It can be observed from Fig. 4a that a stringent delivery delay requirement (e.g., less than 10 s) can be satisfied as long as the alert message arrival rate is below 0.04 (~2 messages/min). If the alert message load reaches 0.1 messages/s, the delivery delay increases to 20 s, which can still be acceptable in some emergency scenarios. However, it can be noticed that in this case, the delivery delay of the other EGNOS services is severely affected (> 100 s for class 2), imposing a significant penalty.

In the case of three fragments (Fig. 4b), the performance of all services is severely impaired because of the larger traffic load injected into the EGNOS system. In more detail, stringent delay requirements (< 10 s) cannot be satisfied even at very low traffic load. In the presence of

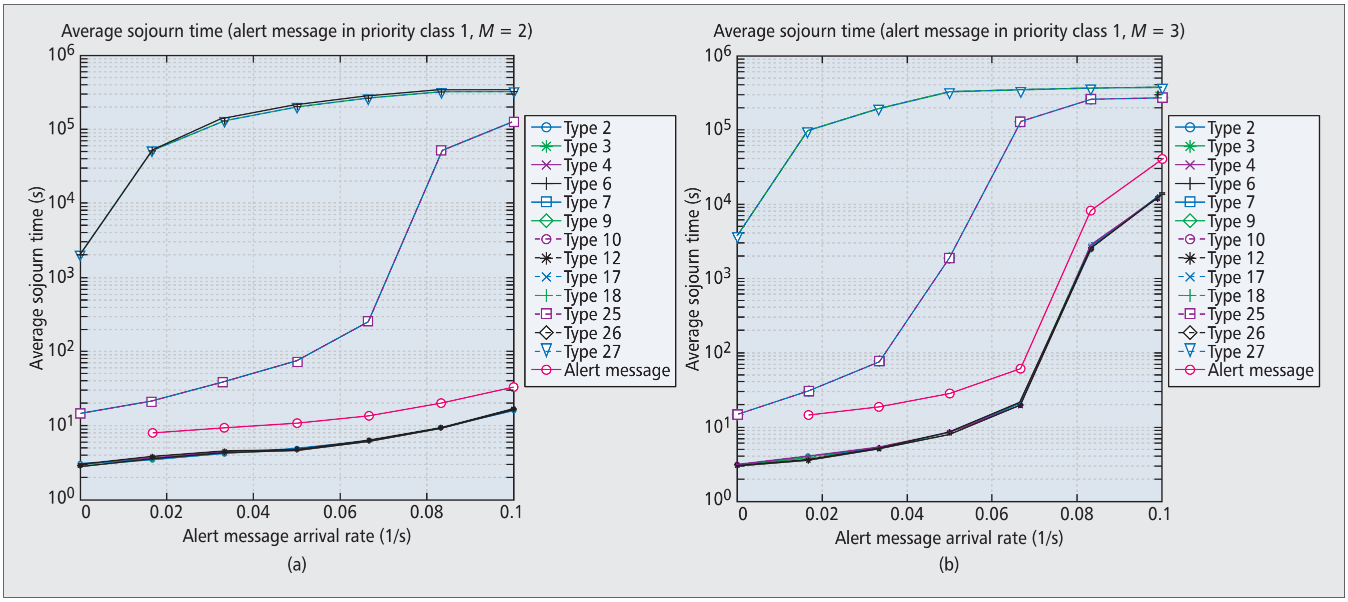


Figure 4. Sojourn time of EGNOS services: a) alert message priority 1, 2 EGNOS frames; b): alert message priority 1, 3 EGNOS frames.

less critical emergency applications, delivery delay lower than 100 s can still be achieved as long as the traffic load stays below 0.065 (~4 messages/min). For higher traffic load, instead, the delivery delay becomes much higher (e.g., thousands of seconds); in this case, both EGNOS and alert message services are severely affected.

Conclusions and Lessons Learned

This article focuses on the potential of satellite navigation systems to disseminate alert messages in the context of a multi-channel approach where different channels complement each other in terms of their coverage, robustness, capabilities, and features to reach the maximum people at risk independent of their location and habits in the use of technologies. This trend has already been explored by numerous projects and currently exploited in standardization frameworks.

In order to validate this concept, we analyze the feasibility of transporting alert messages over EGNOS. The conducted simulation campaigns prove the suitability of EGNOS in that respect in specific operational conditions, provided that efficient alerting protocols (i.e., A4A protocol) are used to transfer alert messages. In particular, it is shown that stringent delay requirements (<100 s) can easily be met when no security measure has to be taken provided that the alert message traffic stays below 7 messages/min (for classes 1 and 2).

On the other hand, the application of authentication measures makes the fragmentation of alert messages necessary, thus penalizing the performance of the overall system. In particular, numerical results show that only the case of a message composed of two fragments and very low alert message load can guarantee stringent delivery delay requirements (<10 s), whereas more relaxed requirements (< 100 s) can also be met in the case of three fragments at lower traffic load.

These technological challenges encourage the enhancement of current navigation systems to efficiently support alert services. In particular, the increase of capacity currently allocated to EGNOS would allow a larger service rate (250 b/s), which will certainly mitigate the problem of the delivery delay increase as long as the alert message load is not too high. This is actually one of the key targets addressed in the evolution of the EGNOS system, currently underway, with the aim to also become more suitable to transport safety of life services.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. [261732].

References

[1] OASIS, "Common Alerting Protocol (CAP) Version 1.2," July 2010.
[2] ETSI TS 102 900 v1.1.1 (2010-10), "Emergency Communications (EMTEL); European Public Warning System (EU-ALERT) Using the Cell Broadcast Services."
[3] A. R. Mathur *et al.*, "Provision of Emergency Communication Messages through SBAS: The ESA ALIVE Concept," *FIG Int'l. Federation of Surveyors*, Feb. 2007.
[4] "Safety of Life" Service Definition Document, EGN-SDD SoL, v. 2.0, June 2013, http://ec.europa.eu/enterprise/policies/satnav/egnos/files/egnos-sol-sdd-v2-0_en.pdf
[5] United Nations, Global Survey of Early Warning Systems, "An Assessment of Capacities, Gaps and Opportunities Toward Building a Comprehensive Global Early Warning System for All Natural Hazards," Sept. 2006.
[6] J. Mulero Chaves and C. Párraga Niebla "Effectiveness Evaluation Model for Public Alert Systems," in Disaster Advances, vol. 7, no. 8, Aug. 2014, pp. 1–11.
[7] http://ec.europa.eu/enterprise/policies/satnav/galileo/files/brochures-leaflets/galileo-the-prs-brochure_en.pdf
[8] C. Fernandez-Prades, L. L. Presti, and E. Falletti, "Satellite Radiolocalization From GPS to GNSS and Beyond: Novel Technologies and Applications for Civil Mass Market," *Proc. IEEE*, vol. 99, no. 11, Nov. 2011, pp. 1882–1904.

[9] Statement of Work EGNOS V3 Definition Phase, ESA-DTEN-NF-SoW/01281, issue 1.0, 08/07/2010, ESA
[10] User Guide for EGNOS Application Developers, ED 2.0, http://www.esa.int/Our_Activities/Navigation/The_present_-_EGNOS/What_is_EGNOS
[11] Galileo Open Service Signal In Space Interface Control Document (OS SIS ICD) issue 1.2, Sept. 2014, draft version.
[12] C. Párraga Niebla *et al*., "The Benefits of Alerting System Based on Standardised Libraries," *Planet@Risk*, vol. 2, no. 2, Apr. 2014, pp. 89–93.
[13] T. De Cola, J. Mulero Chaves, and C. Prraga Niebla, "Designing an Efficient Communications Protocol to Deliver Alert Messages to the Population During Crisis through GNSS," *Proc. 6th Advanced Satellite Multimedia Systems Conf. and 12th Signal Processing for Space Communications Workshop*, 5–7 Sept., 2012, Baiona, Spain.
[14] A. R. Mathur *et al.*, "The European Space Agency Free SBAS Software Developments: Learning, Practising and Accessing the EGNOS Performances in Real-Time," *Proc. Satellite Navigation Technologies and Euro. Workshop on GNSS Signals and Signal Processing*, 5–7 Dec. 2012, ESTEC, The Netherlands.
[15] T. de Cola and C. Párraga Niebla, "Design Aspects for Efficient Distribution of *Alert Messages over GNSS," Proc. IEEE GLOBECOM*, 9–13 Dec. 2013, Atlanta, GA.

Biographies

TOMASO DE COLA [M] received a Laurea degree (with honors) in telecommunication engineering in 2001, a Qualification degree as a Professional Engineer in 2002, and a Ph. D. degree in electronic and computer engineering, robotics and telecommunications in 2010 from the University of Genoa, Italy. From 2002 until 2007, he worked with the Italian Consortium of Telecommunications (CNIT), University of Genoa Research Unit, as a scientist researcher. Since 2008, he has been with the German Aerospace Centre (DLR), where he is involved in different European projects focusing on different aspects of DVB standards, CCSDS protocols, and testbed design. He is a co-author of more than 50 papers, including international conferences and journals. His main research activity concerns TCP/IP protocols, satellite networks, transport protocols for wireless links, interplanetary networks, as well as delay tolerant networks. He has served on the Technical Program Committees of several IEEE international conferences.

CRISTINA PÁRRAGA NIEBLA graduated in telecommunications engineering from Universidad Politcnica de Catalunya, Barcelona, Spain, in July 2002. She joined the German Aerospace Centre (DLR) as a research scientist in 2002. Her research activities dealt with satellite-based communications system design and architectures for different application areas, such as air traffic management and public protection and disaster relief, participating in several ESA and EU-funded projects. She is a co-author of over 50 conference and journal publications, and an active participant ETSI. Since 2008, her main activities deal with project management in the area of satellite communications for civil protection.

Flexible Heterogeneous Satellite-Based Architecture for Enhanced Quality of Life Applications

Enrico Del Re, Simone Morosi, Luca Simone Ronga, Sara Jayousi, and Alessio Martinelli

Abstract

The future pervasive communication ecosystem can improve and enhance the quality of our everyday lives by increasing comfort and safety perception, and improving interaction and conditions of activities: remarkable quality and performance will be afforded by the synergistic use of communication, positioning, and monitoring techniques, provided by means of meshed heterogeneous architectures based on satellite and terrestrial segments. In such a scenario satellite communication will represent the missing link of the evolution of services. This article aims to describe a flexible architecture that could be adopted in the future system and discussing the open points in the development of the system as well as of its single building blocks.

Introduction

Looking ahead at the coming years, a pervasive communication *ecosystem* [1] will be realized by integrating the main systems. In this scenario users will interact with services more than with different technologies in a continuous search for high capacity, high reliability, flexibility, adaptivity, and innovative applications. Eventually based on location and context awareness, this will be fully supported.

The improvement of everyday life will be the result of a many-fold set of actions involving several different aspects of human behavior. Information and communication technology (ICT) can ease this process by allowing the growth of information exchange: in order to follow users during their daily movements, satellites will represent the missing link of the evolution of services.

The "big picture" of how telecommunications can improve our lives in the near future is represented in Fig. 1. There are five main location contexts where we spend the majority of our life. At home, a smart environment can increase our comfort, assist elderly people, and increase our perception of safety when sick [2, 3]. In smart connected workplaces, collaboration tools and information exchange increase both productivity and working conditions [4]. Entertainment applications, automated driving, and road assistance [5] can dramatically increase the quality of our time spent commuting. Social activities can also benefit from new entertainment applications characterized by real-time interactions and immersive experiences. Even the global environment, considered as the place where we spend all our life, can be improved by technologies that are able to help us identify our location, to constantly monitor the health status of Earth's surface, and identify potential risks or hazards.

Citizens are able to share and receive information if they are well equipped with advanced mobile and web-based communications technologies (e.g., smartphones able to send messages, high-resolution images, and high-definition videos). Past crises have clearly demonstrated intensive use of short message service (SMS), which requires lower capacity than voice, and Internet-based messaging (e.g., posts on social networks, text messages), which is boosted by cellular services or the availability of WiFi access points (APs). This trend highlights citizens' empowerment as central information providers because they become the first *in situ* sensors, able to detect and monitor incidents. The use of innovative smart technologies (platforms, devices, services, and applications) could significantly improve situational awareness for public safety organizations (PSOs) and citizens and, as a consequence, provide actionable intelligence and bolster effectiveness and performance (e.g., act faster, react better, and build a safer and more secure society).

This article aims to describe a future flexible architecture that could be adopted in future systems with the goal of providing innovative services. The article is organized as follows. The next section reports some example application scenarios. Then a current satellite-based architecture, which can be considered as a basic pillar for the definition of the future platform, is introduced, and the future flexible proposed architecture is discussed. Next, we focus on open issues and outlooks that need to be considered for the development of such an architecture. Finally, conclusions are drawn in the final section.

Enrico Del Re, Simone Morosi, and Alessio Martinelli are with the Department of Information Engineering (DINFO), University of Florence.

Luca Simone Ronga and Sara Jayousi are with CNIT, University of Florence.

 0163-6804/15/$25.00 © 2015 IEEE

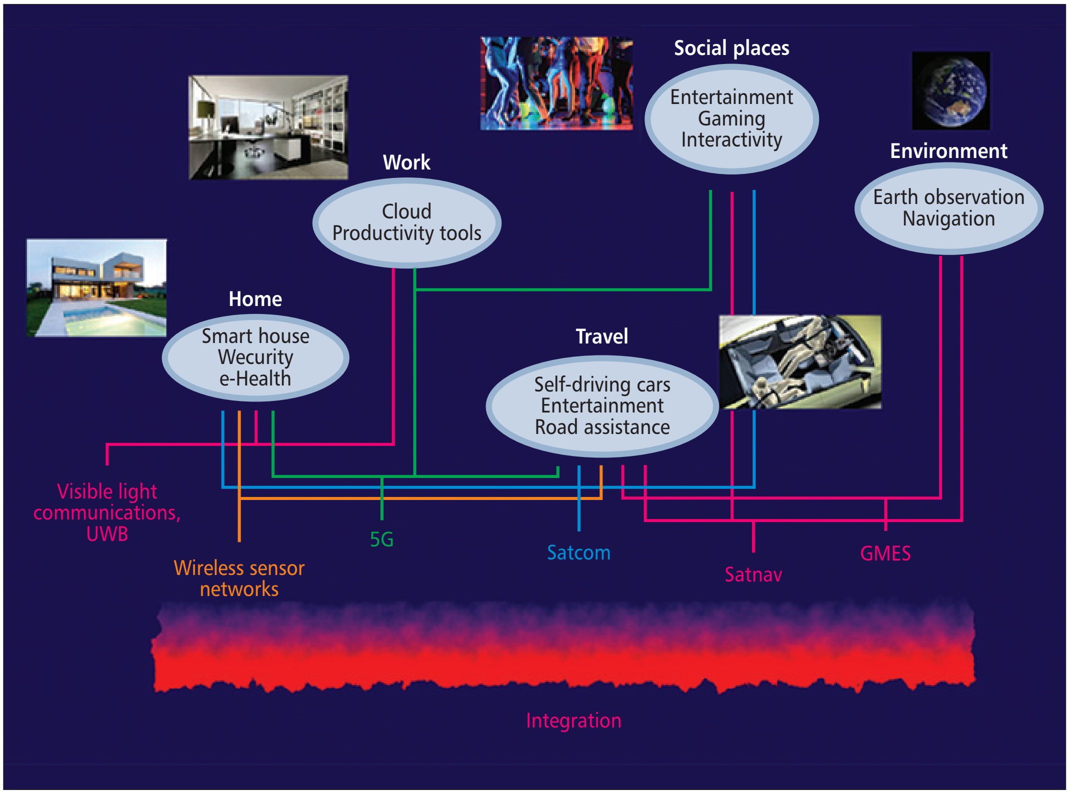


Figure 1. The future pervasive ecosystem.

The flexibility and reconfigurability features of the proposed architecture can support a large variety of services for different potential application contexts. Although already existing services may not be able to exploit all the functionalities provided by the architecture, new advanced services can be developed relying on the integration of them.

Potential Application Scenarios

The flexibility and reconfigurability features of the proposed architecture, described in the next section, can support a large variety of services for different potential application contexts. Although already existing services may not be able to exploit all the functionalities provided by the architecture, new advanced services can be developed relying on their integration. Some examples are reported in the following.

e-Health and Well Being

The proposed architecture can provide solutions to some of the main problems related to the management of many patients in a care environment (hospital, rehabilitation structure, etc.) during all the different care phases and to the remote care of convalescent or home patients (for diseases such as diabetes and hypertension). The patient can be assisted "transparently" at any time through adaptive interactions between smart objects (sensors, drugs, doctors' terminals, patients, relatives, and pharmacies) that have received suitable localization and context information: the monitoring of some main parameters allows immediate assistance in case of necessity or if access to specified building areas is aided or blocked based on established relations helping the necessary patient mobility. The transparency to the user, flexibility, and cooperation between the patient, the objects, and the network are of utmost importance in the disease context and can be specifically tailored to the assistance of patients affected by specific problems or pathologies (e.g., wandering dementia). The specific operational modalities of the platform (user-oriented, object-driven, and network/context-enabled) can provide suitable solutions to the different applications.

Another widespread field of interest that can be addressed by the proposed architecture is well-being systems and sport activities: in particular, people practicing a sport can share information about their activities and experiences while also being connected and monitored by means of wearable devices.

The specific services for these scenarios are summarized in Fig. 2.

Public Safety

Every member of a rescue or safety team is equipped with a portable radio transceiver, with advanced integrated navigation/ communication (NAV/COM) capabilities. Navigation capabilities are necessary to determine the terminal position, using both Global Navigation Satellite System (GNSS) services and, in the absence of a satellite radio link, terrestrial network-based positioning methods. The requirements of radio communication systems for emergency rescue applications [6] shall ensure that the required information is available to the correct person at the appropriate time: communications must be timely, relevant, and accurate since this implies the efficiency of the emergency operations among several authorized emergency personnel teams.

Mobility

In this case, a driver obtains information on traffic conditions not only from a public or private communication network and from other roadside objects, but also from other drivers and

The proposed heterogeneous network architecture can be seen as a general system that is able to self-adapt to the available network resources, to guarantee the adequate quality of service level required by the target applications and to enable new pervasive advanced services.

Figure 2. Enhanced quality of life scenario.

vehicles by means of the establishment of relationships among the cars that usually cover the same track and the consequent exchange of information; hence, by optimizing the travel time and fuel consumption, the best paths can also be chosen in the presence of traffic congestion. Moreover, this scenario also encompasses the exchange of information for traffic, parking, detours, accidents, and so on.

Satellite-Based Architectures

As a matter of fact, technological and societal changes push toward a future in which telecommunication entities and architectures will be integrated with heterogeneous systems with localization and sensing capabilities, and people and intelligent objects will cooperate as in social networks with the goal of providing services. This trend is also enforced by the ongoing Internet of Things (IoT) development, in which pervasive connection of the environment will facilitate positioning, sensing, communication, and interaction capabilities never experienced before.

The required degree of interaction and interoperability among the various communication components is enabled by forging a new integration substratum, which will play a key role in the definition of the next architectures. These components are highlighted in the lower part of Fig. 1. Particularly, satellite communication, navigation, and monitoring systems are envisaged to play a crucial role in the definition of specific functionalities in the framework of the provision of security, transportation, enhanced quality of life, and institutional services.

As for the integration of localization and sensing capabilities, terrestrial and satellite communication subsystems and intelligent objects and networks will have to cooperate in order to provide enhanced value location-based services.

The proposed heterogeneous network architecture can be seen as a general system that is able to self-adapt to the available network resources, to guarantee the adequate quality of service level required by the target applications and to enable new pervasive advanced services. The high degree of flexibility allows the identification of a large variety of application contexts, where the different components of the described architecture may be selected from the different technology domains and be used.

Current Advanced Heterogeneous Architectures

As a starting point, some examples of current satellite-based advanced architectures are briefly described, highlighting their relevant features. In particular, the e-health and emergency contexts are considered, since they represent two significant scenarios where network availability and quality of service (QoS) have to be guaranteed even in critical situations.

E-Health System Architecture — In the e-health context a heterogeneous satellite-based system architecture allows distributed medical competence to be shared and the integration of clinical information, contributing to the quality of medical care for both professional and self-care purposes. In this framework it is worth mentioning, as an example, the KosmoMed project [7]: the system architecture is designed for the provision of professional medical services to allow specialists, health care operators, and patients to access an interactive platform that is able to support different kind of applications such as real-time tele-diagnosis and tele-consultation services. Particularly, the proposed system is oriented to the development of a reliable e-health platform, which is supported by the implementation of QoS control over a commercial satellite broadband network. The system includes the integration of machine-to-machine communication agents in Linux-embedded devices for QoS information exchange and supporting a satellite bandwidth reservation mechanism to improve the user quality of experience (QoE).

This reconfigurable architecture shows the integration of communication and monitoring functionalities that allow an appropriate infrastructure to support the provision of multimedia services to different kinds of users regardless of their access technology (either satellite or terrestrial).

Emergency Heterogeneous Architectures — In the whole cycle of emergency situation management, and particularly during the response and mitigation phases, the availability of a system architecture that integrates localization, communication, and sensing capabilities leads to remarkable gains in terms of speed of response, and completeness and effectiveness of intervention. The exploitation of these benefits can be favored by the adoption of reconfigurable and interoperable devices. The baseline network architecture and proposed solutions for the integration of localization and communication functionalities, which have been studied in the framework of the SALICE project [8], represent an interesting starting point for both the definition of a general system architecture and more sophisticated devices to be used in critical operation conditions. In detail, the definition and testbed implementation of the reconfigurable NAV/COM terminal based on software defined radio (SDR) technology aimed at providing emergency operators (first responders) means to achieve localization information even in light indoor conditions, which are characterized by partial visibility of satellite navigation systems. This result represents an interesting application example of the integration of two of the three main pillars of the architecture proposed in this article. As a result, the functionalities provided by a flexible and heterogeneous architecture can be exploited in the context of emergency situation management. The adoption of satellite is essential for both communication and navigation purposes: mainly thanks to its resilience to terrestrial damage, its broadcast capability over a large area, and its multiple functionalities, the use of satellite is particularly suitable for supporting emergency management activities.

Future Flexible Architecture

In this section the archetypical architecture to be adopted in future heterogeneous networks is discussed.

As shown in the lower part of Fig. 1, the implementation of applications is currently bound to the available access technology for each specific context. The result is that the integration of services is only possible at the client level (i.e., smartphones, car computers, navigation devices), with heavy limitations on potential services. The evolutionary path requires a deeper level of integration of the satellite components as well in order to move services over networks, rather than moving clients over services. This approach is aligned with the recent trend of separation of the control plane from data transport in software defined networks (SDNs) [9]. The migration toward network virtualization functions allows fine-grained definition of available services on a per-user basis, introducing new cross-tier interactions among information domains (user position, user context, user environment, etc.).

In order to support the development of new advanced services that also exploit multiple combinations of existing ones, the proposed system architecture is based on integration of the aforementioned functionalities: localization, communication, and sensing/monitoring. This result can be achieved by the definition of a flexible, scalable, and reliable heterogeneous architecture, which includes the satellite component as a key element. Satellite GNSS, communication, and monitoring systems are all parts of the proposed architecture together with different terrestrial components. This meshed architecture enables cooperation among satellite systems and terrestrial ones with the goal of benefitting from the capabilities of each system. In particular, focusing on localization, communication, and sensing/monitoring functionalities, the capabilities of each single subsystem can be improved by exploitation of the features of the others, also in critical contexts. As an example the use of terrestrial communication technology can help the positioning of a GNSS system in non-line-of-sight (NLOS) visibility conditions (cooperative localization techniques). A synergistic use of the single functionalities can be afforded by the use of the most recent cooperative, cognitive, opportunistic, and context- and location-aware technologies. Moreover, a full interoperable architecture will be enabled by the distributed system intelligence and suitable interfaces.

This concept is depicted in Fig. 3, which shows the baseline system architecture, where the enabling technologies and systems of each of the three main functionalities to be integrated are highlighted (left of the figure):

Communication: The integration of the terrestrial (long- and short-range networks) and space (satellite, unmanned aerial vehicle — UAV) segments can guarantee the availability of network resources in several operational conditions, which include critical situations.

Localization: The adoption of both satellite and terrestrial localization systems is considered for supporting different location-based services even when the line of sight to the satellite is obstructed. This includes the use of the Galileo system together with its new services.

Sensing/monitoring: Environmental sensing and monitoring systems based on integration of information coming from satellite systems and terrestrial sensor networks or sensors installed in UAV platforms can be used for different purposes, such as safety and security applications.

The proposed architecture aims at integrating already existing and future network technologies and standards, exploiting their complementary features (e.g., in terms of coverage extension,

In order to support the development of new advanced services that also exploit multiple combinations of existing ones, the proposed system architecture is based on integration of localization, communication, and sensing/monitoring.

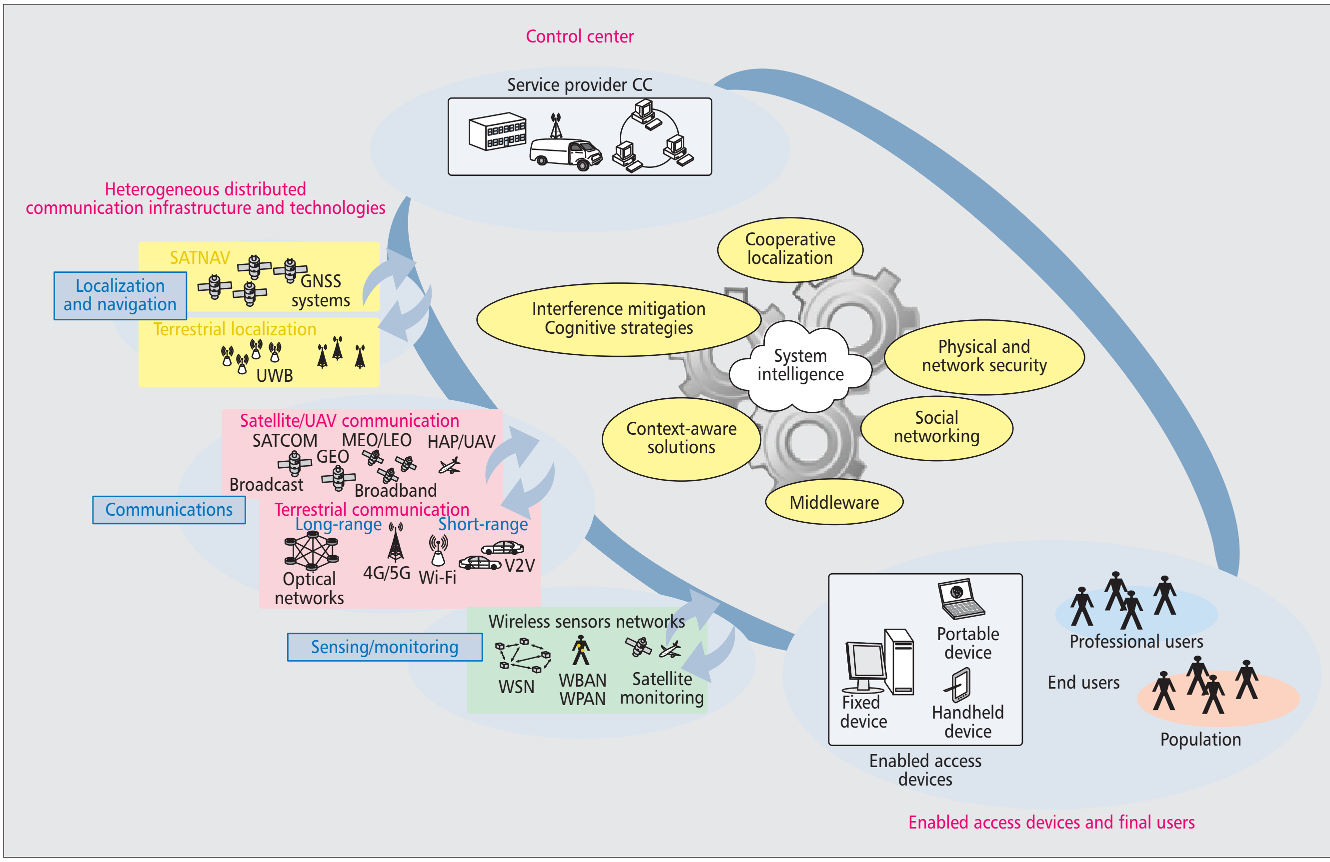


Figure 3. Proposed baseline architecture.

capacity, broadcast/broadband capabilities, resilience, and reliability). It enables users to access (bottom of the figure) multiple services managed by the service provider control center (top of the figure) by using any kind of enabled devices (i.e., fixed, portable, and handheld ones) as the interface with the distributed infrastructure.

The interactions among the different network entities of the overall architecture are supervised by the system intelligence, which, acting under the paradigm of machine-to-machine communications, manages data gathering, exchange, and control by implementing advanced solutions. Some of them are:

- Cognitive and cooperative systems for improving physical and network layer operations
- Security mechanisms
- A middleware architecture for the exchange and management of data in a distributed applications context
- Social networking solutions for supporting the adoption of distributed computing algorithms based on the usage of existing and new technologies for social interaction

In the architecture advanced and assisted localization techniques based on the use of GNSS are supposed to exploit additional information by means of a cooperative approach among users. System intelligence will support the interoperability of the sensor networks with the communication systems, so allowing context-aware capabilities of the system. In this general framework of reconfigurability, cognitive and cooperative paradigms are key elements for achieving the benefits which are guaranteed both at the transmission layer and at the network layer: particularly, the cognitive approach shall be focused on both distributed sensing and the definition of a cross-layer solution with QoS. The use of the cooperative paradigm can act as a booster to provide general and ubiquitous services.

It is worth highlighting that the proposed architecture shall not support only the integration of the information coming by different heterogeneous systems/services (e.g., navigation, monitoring), but will aim at the integration of the capabilities of each subsystem for the development of innovative and advanced techniques to be used for system performance improvement in all situational contexts.

Open Issues

In this section, some important themes and the relative open points for the development of the future heterogeneous ecosystem as well as its single building blocks are reviewed. Moreover, some possible countermeasures are proposed together with a description of the main trends and growth outlook.

Localization

Focusing on localization, an important support to enhanced-value location-based services can be identified in the integration of GNSS and Dead

Reckoning [10] techniques. In particular, GNSS involves several systems such as GPS, GLONASS, and the upcoming GALILEO, plus other regional navigation systems such as QZSS and BeiDou, thus increasing service availability in outdoor environments. On the other hand, dead reckoning allows user motion to be estimated in either indoor or outdoor scenarios without any external infrastructure assistance, relying only on a single terminal and its internal sensors. Nonetheless, this technique is often integrated with GNSS to improve the position accuracy when satellite signals are not available for a short time. In the presence of hostile conditions, that is, when the LOS to the satellites is partially or totally obstructed for a long time, the received signal strength might be too weak for appropriate processing, leading to GNSS-based localization degradation or even failure; in this context, standalone techniques may not afford the location services requirements.

Conversely, in these contests, peer-to-peer cooperative positioning (P2P-CP) [11] is able to improve GNSS receiver performance in terms of availability, accuracy, and time to first fix (TTFF) and can be an effective solution. As shown in Fig. 4, in the P2P-CP context a terrestrial wireless communication link is required between cooperating users to aid information exchange; to this end, short-range communication systems such as WiFi and Bluetooth can be used.

In the case of a navigation system relying on external infrastructure, many solutions can be foreseen, spanning from long- and medium-range radio navigation systems, exploiting cellular networks and pseudo-lites, to short-range ones relying on ultra-wideband, RFID, WiFi, and Bluetooth technologies. An emerging positioning technique is environmental feature matching, which considers different matching sources such as terrain-referenced, imaged-based, and map-based. Aiming at seamless outdoor/indoor localization, effective integration of the aforementioned techniques is a challenging research topic that may involve multi-sensor integration [10].

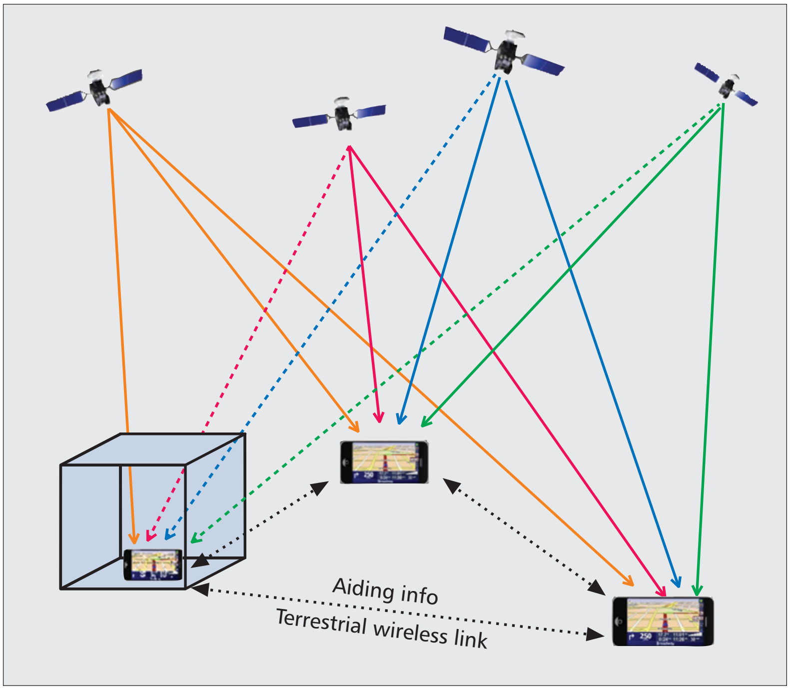


Figure 4. Cooperative localization scenario.

Interference Mitigation and Adaptive Strategies

The scarcity of frequency spectrum and the ever growing demand for data throughput has increased attention to resource sharing and interference management in today's satellite communication systems. In particular, the interference caused by spectrum sharing is deemed harmful, and its suppression by means of orthogonal transmissions in time, space, or frequency is not viable, especially in future heterogeneous systems. This is even more challenging if massive use of multispot antennas for the satellite forward link and cell densification are envisaged: these trends are likely to be adopted in the near future if the forecasts of traffic data growth in the coming years are confirmed.

In recent years, users of professional broadcast applications such as content contribution, distribution, and professional data services have requested more spectrum-efficient solutions. It is expected that in the near future the introduction of higher-resolution video services will drive the need for more efficient transmission technologies in Ku- and wideband Ka-band transponders even further. The delivery of contents will also be based on cache servers close to the point of consumption. This solution can reduce the service-level latency by ensuring that the requested content is available from the local cache.

On the other hand, the advent of efficient transmission, detection, and decoding techniques for satellite communications has already helped to approach the theoretical transmission rates of a conventional point-to-point channel in which the interference is treated as noise. Within this framework, only incremental improvements in data transmission rates can be envisaged. As a consequence, physical layer network coding and interference alignment are now being considered as possible technical solutions for the interference issues.

Context Awareness and Reconfigurability

In order to provide a flexible architecture supporting heterogeneous services, one of the main issue to be addressed is user-transparent network reconfigurability based on context-aware information. Context awareness is the ability to detect and estimate a system state or parameter, either global or concerning one of its components, in a radio system for enhancing performance at the physical, network, and application layers. Moreover, the awareness of the context can also indicate device features (capacity, resolution, etc.), users' interests, places, and locations; the knowledge of this information can contribute to improving QoS and network utilization.

In the proposed future architecture, the possibility to establish the best connection to the Internet will be based on context awareness. Therefore, the services that will be afforded by

Different profiles can be defined according to the specific requirement of each service and consequently a different charging and billing can be applied. The charging can be either volume-based or time-based, enlarging the variety of services to be supported by this model.

this radio access, such as heterogeneous content delivery, will be ideally provided anywhere on the planet.

Focusing on the end user side, the user shall be able to access the required service everywhere and independent of the available network resources. In the last few years the availability of heterogeneous network access technologies and new advanced multimedia services have led to the development of multiple network interface devices. Consequently, vertical handover among different technologies and smart routing mechanisms for dynamic interface management have been proposed. IEEE 802.21 is focused on handover and interoperability among heterogeneous networks, while the Multipath Transmission Control Protocol [12] addresses the simultaneous use of several interfaces aimed at increasing link bandwidth or always on connectivity. Cognitive algorithms, which allow the identification of the best network technology or path to reach the destination according to network parameters analysis, could be defined with the aim of increasing the provided QoS and therefore the QoE perceived by the end user. Moreover, the inclusion of the satellite as one of the interfaces that are available on the receiving device could be extremely useful in those areas where it represents the only network access technology (e.g., remote areas, critical contexts characterized by the unavailability of terrestrial infrastructure). New portable and handheld products have recently been launched in the market to address user mobility, such as Thuraya SatSleeve [13], which transforms a smartphone into a satellite smartphone, guaranteeing access in 161 countries across Thuraya's coverage network.

Security and Privacy

The proposed context-aware service abstraction also provides unprecedented security issues that have to be carefully addressed. In the proposed framework, sensible information flows across paths that are characterized by different reliability grades; hence, weaker segments impact overall reliability and security. Current end-to-end cyphering technologies may be difficult to adopt when multi-homed connections are established between end users and a service cloud. A solution is the enforcement of privacy rules among trusted peers, the composition of which creates the complete service path. IPsec is a candidate for internal core segments for its scalability features and strong privacy provision. On radio access networks, physical layer security techniques like Noise-Loop [14] will efficiently provide the required privacy stratum allowing a secure key exchange mechanism with no a priori shared secrets. Biometrics [15] will also provide an easy and reliable identity verification mechanism to remote peers without the need for complex human-machine interactions.

Adoption of Social Networking Concepts

Citizens may have important information about an event or an intervention scene, and by means of their mobile phones (with HD cameras and text messaging) may decide to share this information with others. Due to the availability of the advanced connectivity concepts and techniques that will be provided by future architectures, the adoption of social networking concepts within the network can be the basis for the introduction of applications allowing PSOs and citizens to share information efficiently (e.g., short messages) in rich formats (e.g., images) over a variety of devices. Hence, individuals may engage in content and social exchanges (e.g., request help, generate an alert) between self-formed groups. The applications will be context-, content-, social- and citizen-aware, and can be selected by end users according to their preferences and needs. As a result, the envisaged network and devices will need to meet the needs of the conception and creation of innovative end-user applications that support enhanced quality of life, mobility, security, and emergency activities.

New Economical Models for Integrated Systems

In the proposed context, the main benefits of using satellite technology as a complementary, alternative, or in some cases the only broadband access solution include technical, contractual, management, and deployment aspects. From the point of view of a service provider, considering the contractual issues, the possibility of contracting with a single operator over different sites, spread over a large area, allows many advantages to be achieved, decreasing the complexity of both deployment and management of services to devise a global system.

Moreover, focusing on the provisioning of new cost-effective commercial solutions and taking into account the flexibility of the proposed architecture, innovative satellite communication services should include bandwidth reservation systems based on pay-per-use services. The flexibility of the satellite bandwidth allocation as well the possibility to hold down the network related costs enable the development of interesting commercial solutions. Different profiles can be defined according to the specific requirements of each service, and consequently a different charging and billing model can be applied. Charging can be either volume-based or time-based, enlarging the variety of services supported by this model. An example of a satellite bandwidth reservation mechanism integrated with e-health services is provided in [7].

Conclusion

Thanks to satellite segments, the integration of communication, positioning, and monitoring functionalities by means of heterogeneous networks will be enabled, allowing the definition of the future pervasive communication ecosystem. This fundamental step will permit the provision of a diversity of services and applications that will improve and enhance the quality of everyday life. In this article services related to the telehealth and security fields have been reviewed, and the future flexible proposed architecture is discussed. Finally, a detailed analysis of the open issues and trends to be taken into account is provided for the development of such architecture as well as of its single building blocks.

REFERENCES

[1] K. U. R. Laghari, N. Crespi, and K. Connelly, "Toward Total Quality of Experience: A QoE Model in a Communication Ecosystem," *IEEE Commun. Mag.*, vol. 50, no. 4, Apr. 2012, pp. 58–65.
[2] A. Honka *et al.*, "Rethinking Health: ICT-Enabled Services to Empower People to Manage Their Health," *IEEE Reviews Biomed. Eng.*, vol. 4, 2011, pp. 119–39.
[3] A. Solanas et al., "Smart Health: A Context-Aware Health Paradigm within Smart Cities," *IEEE Commun. Mag.*, vol. 52, no. 8, Aug. 2013, pp. 74–81.
[4] C. E. A. Mulligan and M. Olsson, "Architectural Implications of Smart City Business Models," *IEEE Commun. Mag.*, vol. 51, no. 6, Jun. 2013, pp. 80–85.
[5] IEEE 1609.3, "Wireless Access in Vehicular Environments (Wave) Networking Services," Dec. 2010.
[6] ETSI, "Emergency Communications (EMTEL) Requirements for Communication between Authorities/Organizations During Emergencies," TS 102 181, V.1.1.1, 2005.
[7] L. S. Ronga *et al.*, "KosmoMed: A Novel Paradigm for Satellite Professional Medical Services," *Proc. Joint 19th Ka and Broadband Commun., Navigation and Earth Observation Conf.*, Florence, Italy, Oct. 2013.
[8] E. Del Re *et al.*, "SALICE Project: Satellite-Assisted Localization and Communication Systems for Emergency Services," *IEEE Aerospace Electr. Sys. Mag.*, vol. 99, no. 9, Sept. 2013, pp. 4–15.
[9] H. Kim and N. Feamster, "Improving Network Management with Software Defined Networking," *IEEE Commun. Mag.*, vol. 51, n. 2, Feb. 2013, pp. 114–19.
[10] P.D. Groves, *Principles of GNSS, Inertial, and Multisensor Integrated Navigation Systems*, 2nd ed., 2013.
[11] S. Morosi, E. Del Re, and A. Martinelli, "Cooperative GPS Positioning with Peer-to-Peer Time Assistance," *Wireless Vitae*, 2014, Aalborg, Denmark.
[12] https://datatracker.ietf.org/wg/mptcp
[13] http://www.thuraya.com/satsleeveandroid
[14] L. Mucchi, L. S. Ronga, and E. Del Re, "Physical Layer Cryptography and Cognitive Networks, *Wireless Pers. Commun.*, vol. 58, no. 1, 2011, pp. 95–109.
[15] A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Trans. Circuits Sys., Video Tech.*, vol.14, no. 1, Jan. 2004, pp.4–20.

BIOGRAPHIES

ENRICO DEL RE is a full professor at the University of Florence, Italy. Presently he is director of the Department of Information Engineering (DINFO) of the University of Florence and president of the Italian Interuniversity Consortium for Telecommunications (CNIT), having served before as director. He is head of the Signal Processing and Communications Laboratory of DINFO, University of Florence.

SIMONE MOROSI (simone.morosi@unifi.it) received his Ph. D. degree in information and telecommunication engineering from the University of Florence in 2000. He is currently an assistant professor at the same university. His present research interests focus on communication systems for emergency applications, green ICT and energy-efficient wireless communications, and future wireless communication systems. He is an author of more than 100 papers in international journals and conference proceedings.

LUCA SIMONE RONGA received his Ph.D. degree in telecommunications in 1998 from the University of Florence. In 1997 he joined the International Computer Science Institute of Berkeley, California, as a visiting scientist. In 1999 he joined the Italian National Consortium for Telecommunications, where he is currently head of the research area. His research interests span from satellite communications to cognitive radio, software defined radio, radio resource management, and wireless security.He has been principal investigator of several research projects and an author of more than 90 papers in international journals and conference proceedings.

SARA JAYOUSI received her Ph.D. degree in computer science, multimedia and telecommunications in 2012 from the University of Florence. She has been with DINFO, University of Florence, since 2008. Her research activity is mainly focused on satellite communications for emergency, IP QoS network management in hybrid satellite/terrestrial networks, cooperative communications, and diversity algorithms in relaying systems. She is an author of several transactions and conference papers.

ALESSIO MARTINELLI is currently a Ph. D. student in information engineering at the University of Florence. In May 2013 he joined the Italian National Consortium for Telecommunications, where he cooperates in several research projects. His main research interests are in the fields of navigation, cooperative positioning, signal processing and classification for activity recognition, dead reckoning algorithms, and their application on smartphones.

Thanks to satellite segments, the integration of communication, positioning, and monitoring functionalities by means of heterogeneous networks will be enabled, allowing the definition of the future pervasive communication ecosystem. This fundamental step will permit the provision of a diversity of services and applications.

COMMUNICATIONS EDUCATION AND TRAINING: INDUSTRY CERTIFICATION AND UNIVERSITY ACCREDITATION


David G. Michelson


Wen Tong


Larry Reeves

Student design competitions have emerged as a popular method for both professional associations and companies to engage engineering students and encourage them to join with like-minded colleagues to develop and test their engineering design and project management skills. By their nature, such competitions encourage a level and intensity of activity that is difficult to replicate in a design or project course. For many engineering students, participation can be a career-defining event.

Student design competitions comprise a diverse set. In some competitions, teams are given a problem to solve during the event, and can only prepare by practicing and training to solve similar problems. In other cases, the teams are given months to develop the solution on their own, and come together with other teams only for demonstration and judging. Some competitions are conducted as an activity associated with an engineering society's annual conference. In other cases, the competition is conducted as a standalone event. The prizes offered range from modest to exceptional.

Despite the significant effort and expenditure devoted to student design competitions in recent years, there are surprising few contributions to the literature on the subject. Manseur [1] and Schuster *et al.* [2] address the benefits and challenges of organizing and conducting student design competitions, while Wankat [3] assessed institutional factors that contribute to student team success. Here, we took a different route, and invited both organizer and student participants to reflect on their experience with the Canadian Satellite Design Challenge (CSDC), a national competition that encourages Canadian university students to form teams to design, build, and test small satellites called cubesats, and compete for both prestige and a possible opportunity to see their design lofted into orbit as a secondary payload.

In "The Canadian Satellite Design Challenge: Building Future Engineering Capability in Canadian Universities," Larry Reeves describes the organization and structure of the CSDC competition. Unique aspects of the competition include the manner in which the student teams from across the country are brought together for workshops and professional visits, the mid-term feedback provided through the Critical Design Review process, and the requirement that each team deliver Educational Outreach presentations to their local communities. As noted above, the prize awarded to the winning team is also unique: a possible opportunity to see their design turned into flight hardware and launched into orbit.

Some of the student projects in the 2012–2014 CSDC focused on contribution to space technology. In "Pioneering the Application of Diamagnetic Materials for Spacecraft Attitude Control," Justin Curran and Cass Hussman from the University of Victoria describe their team's efforts to develop a cubesat that will demonstrate a unique method for realizing an attitude control actuator. In "Testing a Self-Healing Material in Microgravity using a 3U CubeSat," Mehdi Sabzalian, Justin Whatley, Nathalie Parmentier, and Ali Elawad of Concordia University describe their team's efforts to develop a cubesat that will test the properties of a self-healing carbon composite material in space.

Other student projects focused on remote sensing using cubesats. In "Obtaining Infrared Spectral Imagery of the Upper Atmosphere Using a Cubesat," Keith Menezes and Tremayne Gomes of York University describe their team's efforts to develop a cubesat that will use a specialized infrared camera to analyze greenhouse gases in the upper atmosphere and thereby contribute to climate change research. In "Monitoring Glaciers from Space Using a Cubesat," Constance Fodé and colleagues from École Polytechnique de Montréal and the University of Bologna describe their team's efforts to develop a cubesat that will return imagery of Arctic glaciers in support of research being conducted by the University of Montreal's Cold Regions Geomorphology and Geotechnical Laboratory.

Still other student projects focused on study of a unique payload. In "Biological Investigations Using a Cubesat,"Chantelle Dubois, Pawel Glowacki, and Ahmed Byagowi from the University of Manitoba described their team's efforts to develop a cubesat that will expose tardigrades — segmented micro-animals that have been described as nature's toughest animal — to the extreme conditions of space and assess their ability to survive.

In each article, the students share their perspectives on team formation and organization, project management, the impact of CSDC on their schools, and the overall legacy of their involvement. We hope that this rare opportunity to hear the voices of both the competition organizers and participants will stimulate much useful discussion concerning best practices and future opportunities for such events.

REFERENCES

[1] R. Manseur, "Hardware Competitions in Engineering Education," *Proc. 30th ASEE/IEEE Front. Educ. Conf.*, Kansas City, MO, 18–21 Oct. 2000.

[2] P. C. Wankat, "Undergraduate Student Competitions," *J. Eng. Educ.*, vol. 94, no. 3, July 2005, pp. 343–47.

[3] P. Schuster, A. Davol, and J. Mello, "Student Competitions — The Benefits and Challenges," *Proc. 2006 ASEE Annual Conf.*, Chicago, IL, 18 June 2006.

BIOGRAPHIES

DAVID G. MICHELSON [M] (davem@ece.ubc.ca) is with the Department of Electrical and Computer Engineering at the University of British Columbia. His research interests include antenna design, channel modeling, and radar remote sensing. He is a member of the Boards of Governors of the IEEE Communications and Vehicular Technology Societies, and past Director of Education and Training for ComSoc. He previously served as a visiting professor in the Space Physical Sciences Department at the International Space University.

WEN TONG (tongwen@huawei.com) is the vice president and head of wireless research at Huawei Technologies and a Huawei Fellow. He leads one of the largest wireless research organizations in the industry with more than 700 research experts. During a 20-year career, he has pioneered fundamental technologies in wireless with 210 granted U.S. patents. He also serves on the Board of Directors of the WiFi Alliance and the Board of Directors of the Green Touch Consortium.

LARRY REEVES (LReeves@CSDCMS.ca) is a senior member of the Systems Engineering Group at UrtheCast in Vancouver, British Columbia, Canada, where he is responsible for the analysis, design, and development of space and ground segment systems. He holds a Master of Space Studies from the International Space University and has worked in the Canadian space industry for 15 years. He is President of the Canadian Satellite Design Challenge Management Society.

The Canadian Satellite Design Challenge: Building Future Engineering Capability in Canadian Universities

Lawrence A. Reeves

ABSTRACT

The CSDC is a Canadian university competition to design and build a small satellite, called a "cubesat," which will conduct a science mission. The CSDC simulates a real-world engineering project, requiring the multi-disciplinary teams to prepare mandatory documentation, and give a detailed design review in front of industry experts. The CSDC also includes mentored "hands-on" workshops, providing a valuable career-development educational experience.

INTRODUCTION

The Canadian Satellite Design Challenge (CSDC) is a Canada-wide competition for teams of university students to design and build a cubesat satellite, with the goal of launching the winning satellite into orbit in order to conduct its science mission.

The CSDC has been structured to present a significant learning and career development opportunity for the participants, by challenging them to excel in a technologically-advanced domain, and to learn interdisciplinary technical, teamwork, leadership, and management skills.

In participating in the CSDC, due to the unique structure of the competition, students gain knowledge and experience about satellites and space missions, and how to apply their engineering and management educations to a representative real-world project. These skills are valuable not only in the space domain; they are transferable to other advanced technical project domains, and will help them to achieve success in their post-university careers.

The CSDC is managed by the CSDC Management Society Inc. (CSDCMS), a registered not-for-profit organization in Canada, and is entirely volunteer-managed. The first two offerings have each had budgets of approximately CAD $60,000, but have benefitted from significant in-kind contributions of time and use of facilities from several organizations.

The author is with the Canadian Satellite Design Challenge Management Society.

CSDC ACADEMIC BENEFITS

The CSDC is modelled after a number of existing university-level engineering competitions, such as those offered by the Society for Automotive Engineers (SAE). However, there are a number of additional objectives and benefits that have been incorporated in order to enhance the educational experience:

"REAL WORLD" PROJECT STRUCTURE

The CSDC gives participating students a simulated experience of proposing, managing, and working on a real space mission under realistic conditions. A simplified project process is shown in Fig. 1. The CSDC management assumes the role of a customer, and each team is a prime contractor. With a project structure such as a critical design review and required deliverable documents, the CSDC provides a foundation in project management that is rarely encountered in the university curriculum. This experience benefits the students following graduation, as well as the companies that hire them.

WORKSHOPS AND PROFESSIONAL VISITS

The CSDC includes a number of workshops that expose the students to supervised "hands-on" experience in spacecraft assembly techniques and processes, the quality assurance process, mission and spacecraft analysis using leading-edge commercial software packages, and experience working in spacecraft environmental testing facilities. An example of students at work during a workshop is shown in Fig. 2. The workshops have been held at facilities of leading Canadian space companies, such as MDA and Neptec, and at the Canadian Space Agency's David Florida Laboratory spacecraft testing facility. The students appreciate the mentorship, as well as the exposure to potential future employers.

CRITICAL DESIGN REVIEWS

Each team is required to give a comprehensive 2-1/2-hour critical design review (CDR) presentation about the design of their satellite. This

 0163-6804/15/$25.00

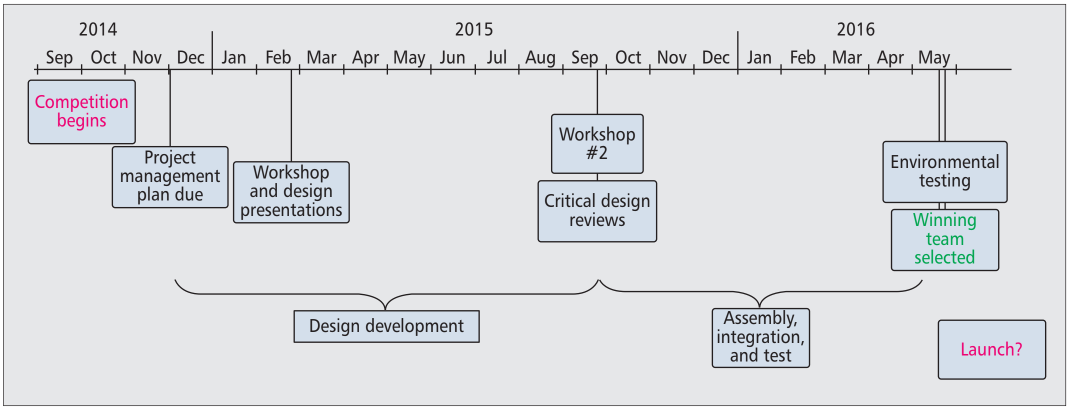


Figure 1. The CSDC schedule is a simplified version of a typical space mission project.

CDR resembles a real mission CDR by a mission prime contractor to its customer, with four competition judges conducting the reviews. The judges are experienced professionals from the Canadian space industry (MDA, Microsat Systems Canada) and government agencies (Defence Research and Development Canada, Canadian Space Agency), and they give detailed review comments about the teams' designs, as well as a grading toward the final determination of the winning team. The CDR presentations are valuable for the teams, all of which have expressed sincere appreciation for the thorough review, and for providing the "voice of experience" that helped them improve their understanding of their spacecraft and mission design process.

Multi-Disciplinary Teamwork

The CSDC gives students experience in working in a multi-disciplinary environment. The more successful teams include students from engineering, science, and business or management departments, as well as occasionally arts and education faculties. This more closely resembles the composition of industry project teams, an enhanced experience compared to uni-departmental projects on which students typically work.

Educational Outreach

An important additional element of the CSDC is the educational outreach. As part of their participation in the CSDC, each team is required to give presentations about space, space science and engineering, their CSDC projects, and/or space applications, to at least five different audience types (elementary school, secondary school, their university, a professional group or sponsoring company, and the general public) each year.

The objectives of the educational outreach component are to:

- Motivate and inspire pre-university students to pursue science and engineering educations.
- Increase public awareness and interest in the applications of satellite missions, and the importance of satellites and space for Canadians.

The educational outreach component of the CSDC is intended to help begin to build younger students into a generation, even further in the future, of capable and inspired leaders. An additional benefit of the educational outreach is to improve the presentation and communication abilities of the team members. These are essential skills required in industry, and are often considered to be lacking in science and engineering students. Figure 3 illustrates a typical classroom presentation given by a CSDC team member.

Figure 2. Students participating in a solar panel assembly workshop, hosted by MDA Corporation in their satellite payload integration cleanroom.

The CSDCMS is grateful to UrtheCast Corporation for providing financial and in-kind contributions in order to create an Educational Outreach Award to recognize a team that demonstrates outstanding excellence in giving presentations or outreach events specifically intended for elementary and secondary school students.

Judging Process

The judging process has evolved since the first competition, and continues to do so. Currently, the winning team is determined by a combination of judges' scores from the critical design review (40 percent) and the final environmental qualification testing (60 percent). Points are allocated for sub-system design elements and the ability to meet requirements, as well as the ability to survive the qualification testing.

Although the winning teams in the first two competitions were quite evident following the

Figure 3. Educational outreach to elementary and secondary school students is a critical developmental step toward improved future capability.

final testing, the disparity between first and second place is becoming increasingly difficult to discern as the teams' capabilities increase. Although some evaluation points can be relatively objectively allocated, other elements are difficult to compare objectively between projects, such as different types of on-board science instruments.

The challenge is to develop evaluation criteria that are clear and objective and eliminate any potential for exploitation of "loop-holes," particularly when the prize for the winning team is hoped and intended to be a launch into orbit. It is likely that the judging process will continue to be iteratively refined through future competitions.

CSDC Summary and Results

To date, 14 Canadian universities and two foreign universities (partnering with Canadian teams) have participated in the CSDC. The perspectives of some of those participating teams are presented in additional articles in this issue.

The most notable result of the CSDC has been the fact that a lack of prior spacecraft mission experience or engineering curriculum at a university is *not* a barrier to successful participation in the competition. Indeed, despite there having been three participating universities with spacecraft design courses in their engineering departments, the first two winning teams in the CSDC were universities at which there was no prior space mission or spacecraft engineering academic content. These first two winning teams were both characterized by excellent team leadership, cohesiveness, energy, and enthusiasm, which more than compensated for any perceived disadvantages.

In only two offerings, the CSDC has created several career and educational opportunities for participating students that might otherwise not have existed:

- During the first CSDC, three teams submitted proposals to the Mission Ideas Contest, hosted by Japan's University Space Engineering Consortium and the University of Tokyo. The University of Alberta's proposal was selected as a finalist. A team representative was invited to Japan to present their proposal, and received an award from the International Academy of Astronautics for the best environmental-focused mission.
- The University of Alberta, which did not have cubesat design experience prior to the first CSDC, is now participating in the European Space Agency's QB-50 mission.
- Several students who have participated in the CSDC have gone on to space-related graduate studies work, including at California Polytechnic State University, where the cubesat standard originated, and the University of Toronto's Institute for Aerospace Studies, Canada's pre-eminent space mission university which has launched several nano-satellite missions.
- The Royal Military College of Canada, which participated in the first CSDC, was invited to include its experimental payload on-board a Canadian research satellite, which is now nearing completion.
- Concordia University, the winning team from the first CSDC, subsequently was one of six university cubesat projects selected for the European Space Agency's "Fly Your Satellite" program (although difficulties arising from requirements changes resulted in the team having to withdraw from the current launch opportunity).

The CSDC has endeavored to offer a unique, interesting, and exciting academic challenge for university students, and it has fostered exceptional results, dedication, and innovation in those who have participated in it. In less than four years, despite having little or no prior space engineering content at many of the participating universities, the teams have grown their universities' space engineering capabilities to the point of having advanced science research satellites that are qualified to be launched.

The Canadian Satellite Design Challenge offers a unique, interesting, and exciting academic challenge for university students, and has fostered exceptional results, dedication, and innovation in those who have participated in it. We believe that the CSDC is a model of a low-cost initiative that can help increase the capability of future engineers, scientists, managers, and teachers and those who learn and benefit from them.

Acknowledgments

The CSDCMS has only been able to achieve its goals thus far with immense support, both financial and in-kind, from many organizations. The author would like to acknowledge and thank the following for their invaluable contributions: Boeing, MDA, UrtheCast, Defence Research & Development Canada, Magellan Aerospace, and the Canadian Space Agency.

Biography

Lawrence (Larry) Reeves (LReeves@csdcms.ca) received his B.Sc. in computer science from the University of Victoria (BC), his M.Sc. in computer science from Queen's University (Kingston, ON), and his Master of Space Studies degree from the International Space University in Strasbourg, France. He has worked for over 15 years in the Canadian space industry, and is currently a senior member of the engineering department at UrtheCast in Vancouver.

CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE
TOWARD AUTONOMOUS DRIVING: ADVANCES IN V2X CONNECTIVITY

BACKGROUND

The research area of intervehicle and vehicle-to-infrastructure (V2X) networking and respective cooperative driving applications has been growing in the last few decades. Now it is clear that wireless technology will be a communication baseline for many promising cooperative automotive applications, which will make driving safer, more energy-efficient, and more comfortable.

Autonomous driving is the next step in what is considered a strategic direction by many vehicle manufacturers. Although there is a long way to go before fully autonomous vehicles will be introduced massively in ubiquitous city environments, it is already practically feasible to consider fully automatic operation of vehicles in restricted areas (harbor, parking lot, dedicated public transport lanes). Autonomous cooperative driving enabled by V2X communications have highly demanding operating conditions and generate delay-sensitive data traffic with requirements for high reliability. On the way toward purely autonomous vehicles, platooning is a state-of-the-art emergent application, where a caravan of vehicles on the highway automatically follow the leading one. Although a high level of automation is inherent in platooning applications, the leading vehicle itself is still controlled by a human.

To reflect the above aspects of V2X vehicular networking, this Feature Topic calls for original manuscripts with contributions on all aspects of highly automated and fully autonomous vehicles, including, but not limited to, the following topics:

- V2X vehicular networking
- Cooperative adaptive cruise control
- Platooning
- Networking applications and services
- Security and privacy
- Simulation and performance evaluation
- Experimental systems, testbeds, and field trials

SUBMISSIONS

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words (from introduction through conclusions). Figures and tables should be limited to a combined total of six. The number of references is recommended not to exceed 15. In some rare cases, more mathematical equations, figures, and tables may be allowed if well justified. In general, however, mathematics should be avoided; instead, references to papers containing the relevant mathematics should be provided. Complete guidelines for preparation of the manuscripts are posted at http://www.comsoc.org/commag/paper-submission-guidelines. Please send a pdf (preferred) or MS Word formatted paper via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to Author Center. Follow the instructions there. Select "December 2015/Toward Autonomous Driving: Advances in V2X Connectivity" as the Feature Topic category for your submission.

SCHEDULE FOR SUBMISSIONS

- Manuscript Submission Due: June 1, 2015
- Decision Notification: August 1, 2015
- Final manuscript due: October 1, 2015
- Publication Date: December 2015

GUEST EDITORS

Alexey Vinel (Halmstad University, Sweden), alexey.vinel@hh.se
Lin Lan (Hitachi, France), lan.lin@hitachi-eu.com
Onur Altintas (Toyota InfoTechnology Center, Japan), onur@jp.toyota-itc.com
Henrik Pettersson (Scania, Sweden), henrik_x.pettersson@scania.com
Oleg Gusikhin (Ford, United States), ogusikhi@ford.com

Pioneering the Application of Diamagnetic Materials for Spacecraft Attitude Control

Justin Curran and Cass Hussman

ABSTRACT

Developed by a student team at the University of Victoria, the ECOSat2 cube satellite is a demonstration platform for communications and attitude control technologies useful in small satellites. Its scientific payload is an experimental attitude control actuator that uses a diamagnetic material called pyrolytic graphite. We seek to demonstrate that this diamagnetic effect can be used to control the attitude of small satellites in Earth orbit, and thereby inspire research into developing materials with stronger diamagnetic properties and further their use in spacecraft control.

INTRODUCTION

The ECOSat2 cube satellite developed by the University of Victoria is a demonstration platform for communications and attitude control technologies. Its scientific payload is an experimental attitude control actuator that uses a diamagnetic material called pyrolytic graphite. As a diamagnetic material, pyrolytic graphite seeks a magnetic neutral point, that is, it is repelled from both north and south magnetic fields. Through optical excitation of the material using lasers, causing localized heating, it is possible to modify the magnetic susceptibility of the material [1]. This creates an unbalanced magnetic force in the material that interacts with Earth's magnetic field. Our goal is to demonstrate that this diamagnetic effect can be used to control the attitude of small satellites in Earth orbit, and to inspire research into developing materials with stronger diamagnetic properties, furthering their use in spacecraft control.

SATELLITE SYSTEMS

ECOSat2's primary attitude control system features an all-solid-state configuration with no moving parts. It is based on strong electro-magnets called magnetorquers. Magnetorquers are typically used as a secondary actuator in conjunction with momentum wheels in order to counteract the torque moment created when de-saturating the rotational speed of momentum wheels. Magnetorquers, however, have a much higher reliability as they do not contain moving parts, such as bearings, that momentum wheels require. The ECOSat2 magnetorquers were designed, manufactured, wound, and tested in house, boasting a current density output seven to ten times that of leading cubesat manufacturers, in a similar size [2].

The communication system operates in the 2m and 70cm amateur radio frequency bands, and is implemented on an in house developed software defined radio (SDR). The platform for this is based on Simulink in Matlab combined with code generation for Texas Instrument C series DSP microprocessors. The use of SDR allows for enhanced flexibility of the satellite communications and the ability to reconfigure the implementation in orbit. Additionally, the flexibility of SDR has allowed for multiple functionality of the communication system hardware with a single radio frequency and intermediate frequency front end. The communication system implemented by our team contains three bands: a linear transponder, QPSK data link, and a CW beacon. This flexibility allows for amateur radio enthusiasts to use the satellite as a relay as well as download telemetry data. In addition, it provides the ability to change the modulation scheme dynamically in orbit.

TEAM ORGANIZATION

The ECOSat team was formed as an engineering student club at the start of the first round of the Canadian Satellite Design Challenge in January 2011. The team typically consists of 20 to 30 students from different backgrounds, primarily electrical, computer, and mechanical engineering, but also computer science, physics, math, and business. The mandatory co-op program at the University of Victoria results in a high team member turnover rate each term.

The team has seven core members filling leadership roles to help organize the team and track individual tasks and deadlines of different aspects of the satellite. There are two primary management roles: the chief engineer deals with overall system design and integration as well as

The authors are with the University of Victoria.

mitigating all technical arbitrations; the project manager is responsible for maintaining timelines, allocating resources, interfacing with administration, and acting as the first point of contact. Working below them are five positions that focus on individual aspects of the satellite: electrical, mechanical, payload, communications, and software. The leaders for these individual teams oversee their members, co-ordinate resources and information, and help troubleshoot problems prior to speaking with the chief engineer or the general team forum.

Project Management

The leaders of the sub-teams do the majority of the work on the satellite, with smaller design projects flowing down to core members. Once per week, the team meets to discuss system progress, design changes, difficulties encountered, and required resources. This provides an open forum to discuss systems changes, troubleshoot roadblocks to design progress, as well as disseminate system information to team members. This helps to eliminate confusion about system operations and the repetition of information.

By the end of the second round of the Canadian Satellite Design Challenge, all initial revisions of systems were complete and well under way toward assembly, integration, and testing. The entire attitude determination and control system, power system, on-board computer, communications system, and payload were developed by the team in-house, with only the GPS being purchased as a completed system. The fully assembled and integrated satellite was then packaged and transported with the team members to Ottawa to conduct ground vibration testing at the Canadian Space Agency's David Florida Lab. The team also brought a full complement of hand tools, cleaners, safety equipment, and debug hardware to perform on-site troubleshooting.

Legacy

The Canadian Satellite Design Challenge and the start of the ECOSat team has provided opportunities for students passionate about space technology to be able to work on a small satellite and gain expertise and contacts within the field previously inaccessible to them. Several current and previous team members have obtained both post-graduation positions and co-op placements in aerospace companies due to their involvement in the competition. Originating from the work conducted in the club, many members have taken aspects of the satellite to further explore systems as senior design projects, class projects, honors theses, and graduate studies. Whenever possible the team management tries to work with faculty members to incorporate system designs into class projects. This helps provide additional incentive for team members, and helps to provide the students with a reduction in workload to facilitate the large time commitments required to successfully develop satellite system designs.

The University of Victoria has dedicated two rooms for developing satellite systems; the primary office provides 24-hour access for all team members. The office is fitted with a complement of computers for programming, CAD, and simulation work. The office is also equipped with some basic electronic equipment for developing and troubleshooting systems. The second room functions as a lab space for in depth testing and assembly of systems; this room also features a small office in which graduate students can work. The team has also procured a third room to house the ground station for satellite communication. There have not been any permanent changes to courses or curricula. However, professors have been very accommodating by allowing class projects to include aspects from the satellite design, as well as promoting satellite system projects to third and fourth year engineering design project classes.

There have not been any permanent changes to courses or curricula. However, professors have been very accommodating by allowing class projects to include aspects from the satellite design, as well as promoting satellite system projects to 3rd and 4th year engineering design project classes.

References

[1] K. Masayuki and A. Jiro, "Optical Motion Control of Maglev Graphite," *J. American Chemical Society*, vol. 134, no. 51, 2012, pp. 20593–96.

[2] Innovative Solutions in Space, "CubeSat Magnetorquer Rod," Cubesatshop, 2015; available: http://www.cubesatshop.com/index.php?page=shop.product_details&flypage=flypage.tpl&product_id=75&category_id=7&option=com_virtuemart&Itemid=69. [Accessed 15 1 2015].

Biographies

Justin Curran (justintcurran@gmail.com) completed his technology diploma program at Camosun College, then bridged to the University of Victoria and now holds a degree in electrical engineering with a specialization in electronics (2012), and a Masters of Applied Science in mechanical engineering with a focus on aerospace (2014) from the University of Victoria. His graduate work stems from the involvement in the satellite project at the university. He also currently holds the chief engineer position on the satellite project, which is a continuation of his work as an undergrad. He has maintained this position since the project's inception in January 2011. He currently plans to actively work on the project until the completion of the satellite and on through the pre-launch and ground station operation phases. Having volunteered for the position at the initial team meeting, he has been enthralled with the project and delighted with all the opportunities it has garnered for him. He strongly encourages others on a continuing basis to join and get involved with not only the satellite project and the Canadian Satellite Design Challenge, but with any extra-curricular team. He is an instructor at Camosun College, where he now teaches electronic and computer engineering technology as well as its related courses, topics, and fields.

Cass Hussmann completed his undergraduate degree in computer engineering specialized in digital and embedded systems along with mechatronics at the end of 2013 from the University of Victoria. After joining the Ecosat team, He found a passion for the aerospace industry, leading him to pursue projects and topics outside of the computer engineering curriculum, including an honours thesis focusing on preventing single event effects in digital systems. He is currently pursuing a Masters of applied science in mechanical engineering with a focus on aerospace at the University of Victoria. His graduate work focuses on reliable design and reliability analysis of embedded systems on the satellite project at the university. He has worked as the electrical lead since 2013, with his work focusing on command and data handling of all on-board systems, the on-board computer design and layout, and general management duties. He intends to continue work with the University of Victoria satellite project until its launch and operation while pursuing a career in space technology.

Testing a Self-Healing Material in Microgravity Using a 3U CubeSat

Mehdi Sabzalian, Justin Whatley, Nathalie Parmentier and Ali Elawad

The authors are with Space Concordia–Spacecraft Division, Montreal, Quebec, Canada

ABSTRACT

Space Concordia, a student society from Concordia University, has designed and built a 3U CubeSat named Aleksandr as their entry into the 2014 CSDC. The satellite, once deployed, will test the properties of a self-healing carbon composite by performing a three-point bending test in space, and thereby assess its viability for use in space applications. The engineering model of the satellite payload met all CSDC testing requirements and placed second in the competition. Space Concordia now aims to expand and refine Aleksandr to participate in the upcoming CSDC.

INTRODUCTION

Spacecraft frequently experience collisions with micrometeoroid and orbital debris (MMOD). For instance, NASA's Long Duration Exposure Facility reported approximately 3439 impacts over almost six years, where more than 90 percent of impacts were micrometeoroids smaller than 1 mm [1]. Damage and expensive repairs caused by MMOD impacts could be avoided by using a protective layer of self-healing material. For example, this material might form a layer in a multi-layer shield, such as the Whipple shield, which could recover from MMOD impacts, thus improving the longevity of a spacecraft. The addition of a self-healing material formed of carbon composite, which could be made compact and lightweight without compromising durability, would be ideal for spacecraft [2].

The mission of Space Concordia's 3U CubeSat, named Aleksandr (previously ConSat-2, engineering model shown in Fig. 1) is to study the behavior in microgravity of a carbon composite self-healing material, produced by Dr. Suong V. Hoa at Concordia University [3]. Following deployment of Aleksandr, Space Concordia aims to compare the healing efficiency of the composite in microgravity relative to ground tests. If successful, this experiment will give flight heritage to the composite and enable future spacecraft to incorporate this self-healing material into their structures.

TEAM ORGANIZATION

Aleksandr is being developed by the student society Space Concordia, operating under the Engineering and Computer Science Association at Concordia University. The society was formed in November 2010 to participate in the Canadian Satellite Design Challenge (CSDC). Since its inception, Space Concordia has developed significantly, becoming one of the largest engineering societies at Concordia University with more than 100 members participating in four independent divisions: spacecraft, rocketry, robotics, and ground station.

Space Concordia's ConSat-1 team won CSDC in September 2012 based upon a detailed assessment of the engineering model of their satellite. Consequently, when the second edition of CSDC was announced, members of Space Concordia were eager to participate again. Recruitment consisted of an open call to engineering students at the university, inviting them to a team formation meeting. Students from mechanical and industrial engineering, computer and electrical engineering, software engineering, and physics would form the new team, the majority of whom were in either their first or second year of undergraduate studies. Throughout the competition, every effort was made to recruit new engineering and non-engineering students with an interest in space science. Initially, a few members from the ConSat-1 team helped organize the new ConSat-2 team to maintain a communication link between the two teams. This allowed for the new team to benefit from the previous team's knowledge and satellite design experience.

Similar to ConSat-1, the team behind Aleksandr was partitioned into seven sub-systems: attitude control system, command and data handling, communications, mechanical (structural and thermal), payload, power, and software. Sub-system leads were responsible for most of the technical work and important decisions made for their sub-system. Sub-system teams met frequently when everyone was available, while the sub-system leads had regular weekly meetings supervised by the team lead. The meetings helped ensure that each sub-system was on schedule, and that all sub-systems remained compatible with each other. In the final year of the competition, the ConSat-2 team worked for eight hours every Sunday, a sig-

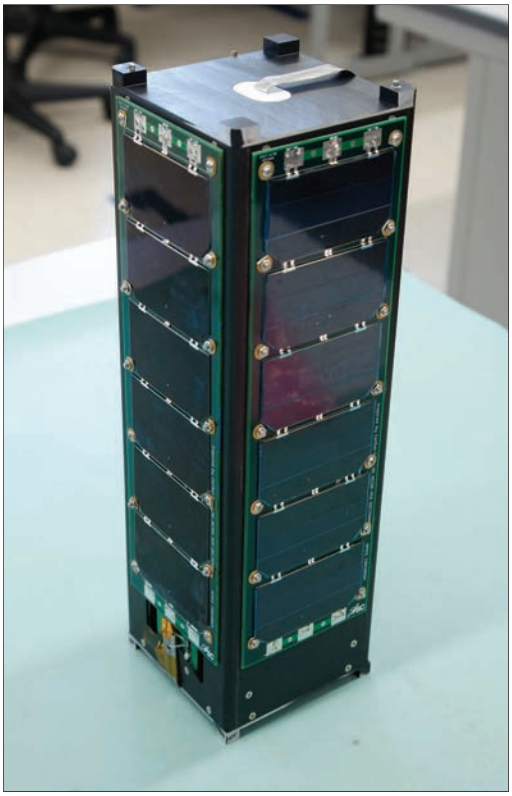


Figure 1. The engineering model of the Aleksandr CubeSat.

Figure 2. The Aleksandr team during the final environmental testing phase at the Canadian Space Agency's David Florida Laboratory in May 2014 (Photo Credit: Larry Reeves, CSDC).

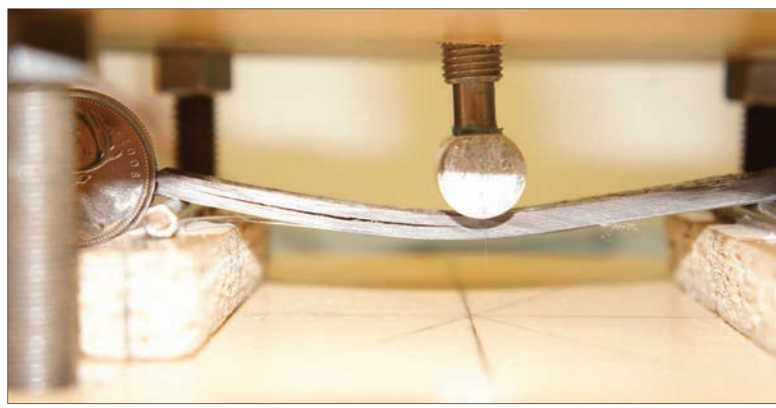


Figure 3. Broken composite sample from a payload prototype with a Canadian quarter to the left side for scale.

nificant time commitment outside of the regular demands of their full-time student schedules.

This project could not have been completed alone. Funding for Aleksandr was provided by many generous external sponsors. To raise additional funds, the team also launched a crowdfunding initiative, a Kickstarter campaign, which received support from national and international donors. Throughout development, the Aleksandr team also sought support from engineers in industry with advice or help. Among others, the team had the great opportunity to use the testing facilities at MPB Technologies Inc. The support of company employees was both greatly appreciated and critical to the success of this project.

Primary Experiment

The engineering model of Aleksandr was delivered to CSDC in May 2014, as shown in Fig. 2. In designing the satellite's primary experiment, the goal was to make it as simple and reliable as possible, while also providing scientific value. For the experimental portion of the satellite, the payload was designed to perform a three-point bending test on a sample of self-healing composite. This self-healing material is a fiber-reinforced carbon composite containing an epoxy resin matrix [3]. Following damage, an autonomic polymer healing process is triggered that regenerates the damaged area, thus stopping the propagation of a crack [2]. Organizations such as the Canadian Space Agency, Concordia University, and MPB Technologies Inc. are among those supporting the development of this self-healing composite for space applications [4].

The experiment follows a proposed standard test by the American Society for Testing and Materials (ASTM): "Recommendations for an ASTM Standardized Test for Determining GIIc of Unidirectional Laminated Polymeric Matrix Composites" [5]. This particular test was chosen to conform to the experimental design used in ground tests for the self-healing composite currently being conducted in the Concordia Centre for Composites. To accomplish the ASTM standard test, the sample had to be subjected to a large force, while also being restrained in touching points, as shown in Fig. 3, and without being clamped. Fulfilling these testing requirements, while still securing the sample enough to prevent damage from harsh launch vibrations, was one of the greatest engineering challenges the team faced. However, the limitation of only three touching points is necessary to perform a three-point bending test in orbit. At the Canadian Space Agency's David Florida Laboratory, the team was gratified to discover that the payload did, in fact, meet the team's predictions and passed the vibra-

Further incentivized by the passion and enthusiasm demonstrated by Space Concordia students for Aerospace Engineering, Concordia University will be creating an undergraduate Aerospace Engineering program that is expected to begin in September 2016.

tion tests. During these vibration tests, the Aleksandr team also successfully verified the structure of the satellite, which they designed, modeled, analyzed, and manufactured themselves.

SUPPORT SYSTEMS

When designing Aleksandr, the team had to consider more than just the payload. The challenge of creating a satellite that could be fully functional in space required extensive research, several designs and redesigns, and the creation of various support systems. In order to support the payload, Aleksandr included several components. Some of the most significant accomplishments were:

- Creating an antenna system that was manufactured and tested in-house.
- Perfecting a procedure to make and assemble solar panels.
- Developing a passive attitude control system.

The Aleksandr team also designed their own four-layer power PCB to charge the CubeSat batteries as efficiently as possible, regulate the on-board power, and dissipate the generated heat properly. The on-board computer, a BeagleBone Black, runs flight software that is fully developed by Concordia students. One component of this flight software, called Operational Frames 2nd Generation (OF2G), ensures the reliability of data transmission between the satellite and the ground station. This feature was demonstrated during the functional testing at the end of CSDC.

LEGACY AND CONCLUSION

The ConSat-1 and Aleksandr projects have been incredibly enriching for everyone involved. Indeed, some members of the 2014 Aleksandr team have continued to work on the satellite as a Capstone Project, improving its design to make it more customizable. For their project, these students intend to build a modular assembly satellite standard (MASS). In the spirit of collaboration, these MASS designs will be made open-source so that other amateur satellite builders can use them to their advantage. Further, they are working on creating an in-house vibration testing facility that will provide empirical validation for future CubeSat testing at Concordia. The challenge of designing, building, and testing satellites has even inspired some to continue their pursuit in space engineering through graduate studies or full-time jobs in the space industry.

For the current CSDC, the Aleksandr project is being refined and expanded. Following tradition, a new group of students have been given the opportunity to try their hands at space engineering. The goal of the new team is to prepare Aleksandr's flight model for the final flight qualification tests in May 2016 during the environmental testing phase for CSDC. With an influx of new recruits, Space Concordia hopes to encourage collaboration and make inexpensive space technology accessible around the world. Further incentivized by the passion and enthusiasm demonstrated by Space Concordia students for Aerospace Engineering, Concordia University will be creating an undergraduate Aerospace Engineering program that is expected to begin in September 2016.

ACKNOWLEDGMENTS

Team Members: Justin Bunker, Jan Clarence Dee, Ali Elawad, Eric Heiligenstein, Ahmed Ibraheem, Daniel Kabagema, Rami Kandela, Mohammad Asgar Khan, Michael Khor, Yaz Khoury, Nirojan Mahendrarajah, Amine Mhedhbi, Sari Morcos, Olivier Nguyen, Mats Powlowski, Jean-Robert Harvey, Mehdi Sabzalian, Svetlana Spitsina, Anthony Tran, and Nicholas Velenosi

External Support: Brahim Aissa, Nadia Bhuyian, Tyson Boer, Shawn Bulger, Scott Gleason, Guy Gosselin, Emile Haddad, Suong V. Hoa, Teodor-Alexandru "T-Rex" Ionita, Ivan Ivanov, Robert Jakubowicz, Tiago Leao, Alex MacPherson, Matthias Martineau, Alfred Ng, Dominic Ng, Alex Potapov, Michael Rembacz, Shawn Stoute, Nick Sweet, and Kamel Tagziria

Sponsors: Concordia Center for Composites, Concordia Council on Student Life, Concordia Institute of Aerospace Design & Innovation, Concordia University Alumni Association, Concordia University's Engineering and Computer Science Association, Concordia University's Department of Mechanical and Industrial Engineering, Fonex, Google, MDA, MPB Technologies, Northforge, Pulsar Circuits, and Kickstarter supporters around the world.

REFERENCES

[1] T. H. See *et al.*, "LDEF Meteoroid & Debris Special Investigation Group Investigations and Activities at the Johnson Space Center," *LDEF 69 Months in Space: Third Post-Retrieval Symposium, part 1*, NASA CP 3275, Feb. 1995, pp. 257–73.

[2] P. Merle *et al.*, "Self Healing Composite Material and Method of Manufacturing Same," U.S. Patent 20090036568 A1, Feb. 2009.

[3] G. Thatte *et al.*, "Development and Characterization of Self-Healing Epoxy Systems for Space Applications," *Proc. 1st International Conference on Self Healing Materials*, Noordwijk aan Zee, Netherlands, 2007.

[4] B. Assa *et al.*, "The Self-Healing Capability of Carbon Fibre Composite Structures Subjected to Hypervelocity Impacts Simulating Orbital Space Debris," *ISRN Nanomaterials*, vol. 2012, Sep. 2012.

[5] B. D. Davidson and S. S. Teller, "Recommendations for an ASTM Standardized Test for Determining GIIc of Unidirectional Laminated Polymeric Matrix Composites," *J. ASTM Int'l.*, vol. 7, no. 2, Jan. 2010, pp. 1–11.

BIOGRAPHIES

MEHDI SABZALIAN (mehdi@spaceconcordia.ca) is the president of Space Concordia, former Aleksandr team lead, and a senior mechanical engineering student at Concordia University. He is deeply interested in space engineering, a subject he will be pursuing at a graduate level.

JUSTIN WHATLEY is the executive director of Space Concordia. He has a bachelors degree in behavioral neuroscience from Concordia University, and is currently studying computer science and software systems. Some of his aspirations include making contributions to space engineering and technologies.

NATHALIE PARMENTIER is an English student at Concordia University. While pursuing her studies, she has developed a keen interest in the sciences and humanities, and enjoys reading and writing in both fields.

ALI ELAWAD is a member of Space Concordia, where he worked on Aleksandr's thermal and vibration analysis, and testing. His main interest is innovation and developing sustainable technologies. He hopes to bring affordable manufacturing to developing countries in order to promote self-sustainability.

Obtaining Infrared Spectral Imagery of the Upper Atmosphere Using a Cubesat

Keith Menezes and Tremayne Gomes

The authors are with York University.

ABSTRACT

The mission of the Canadian Satellite Design Team at York University is to develop a Cubesat payload that will use the AURORA line scan camera to obtain infrared spectral imagery of the upper atmosphere. The imagery will be used to assess the concentration of greenhouse gases in the atmosphere and thereby advance our understanding of climate change. The team comprises a diverse group of students with a variety of backgrounds. The manner in which our group has been organized and the facilities that we have obtained access to will be a critical factor in our ultimate success.

INTRODUCTION

The primary goal of the POLARIS (Polar Orbit, Lower Atmosphere Research Instrument using Spectroscopy) cubesat is to provide infrared spectral imagery from the Earth's upper atmosphere using the cost-efficient Aurora multipixel camera made by Thoth Technologies. The captured images will allow for analysis of numerous greenhouse gases in the atmosphere that is useful for understanding climate change (Fig. 1). This mission will demonstrate the radiation balance of the Earth by understanding the upwelling radiation and help to quantify the effects of clouds, aerosols, water vapor, and carbon dioxide on the climate (Fig. 2). Our secondary objectives are to:

- Promote space exploration to our local community and university.
- Demonstrate in a cost-effective manner the extensive capabilities of nano-satellites.
- Provide a hands-on learning experience for our engineering and science students.

In order to achieve these goals we needed a great team to ensure they could be met.

TEAM ORGANIZATION

We established the Canadian Satellite Design Challenge (CSDC) team in order to improve the university experience of students with an interest in space science and exploration. In a short period of time, we found quite a few like-minded students. Part of the attraction was the manner in which CSDC integrates many components of space engineering, computer, software, electrical and mechanical engineering, as well as the space and atmospheric science curriculum. Through software analysis of orbits, space hardware, control systems, payload design, remote sensing, and space mission design, CSDC allows the students to apply what they are learning in class to a realistic project.

By establishing the club under the York Federation of Students, we were able to obtain basic support from the university. Our team includes engineering and science students, as well as business majors and liberal arts students. Finance students help with money management, Human Resources resolve any conflicts between the subsystem teams, and the Liberal Arts majors are responsible for public relations and communications. To attract the most dedicated members, we created videos and text messages that were sent to each college at York University. As a result, we found various members that were willing to devote their extracurricular time for the challenge. This is why we have such a diverse and committed group. The outcome of these efforts has been a strong turnout and a significant increase in the number of participants.

DEVELOPMENT AND TESTING

Our CSDC team has been working diligently on designing our satellite and creating technical documentation. Early on in the competition we prepared a project management plan, the first required deliverable document of the competition. This document outlined how the team will be structured and managed. We also prepared a presentation explaining the basic characteristic of our satellite and its purpose for the CSDC workshop that was held at Magellan Aerospace in Winnipeg, MB in February 2015.

The CSDC team at York is moving forward into the next couple of months by obtaining more facilities and training for the team. We are arranging our expenses for the year so we can purchase materials for developing our satellite engineering model after the competition's preliminary design review. Team meetings are held every week in the Space Engineering Laboratory of the Petrie Science and Engineering Building.

 0163-6804/15/$25.00

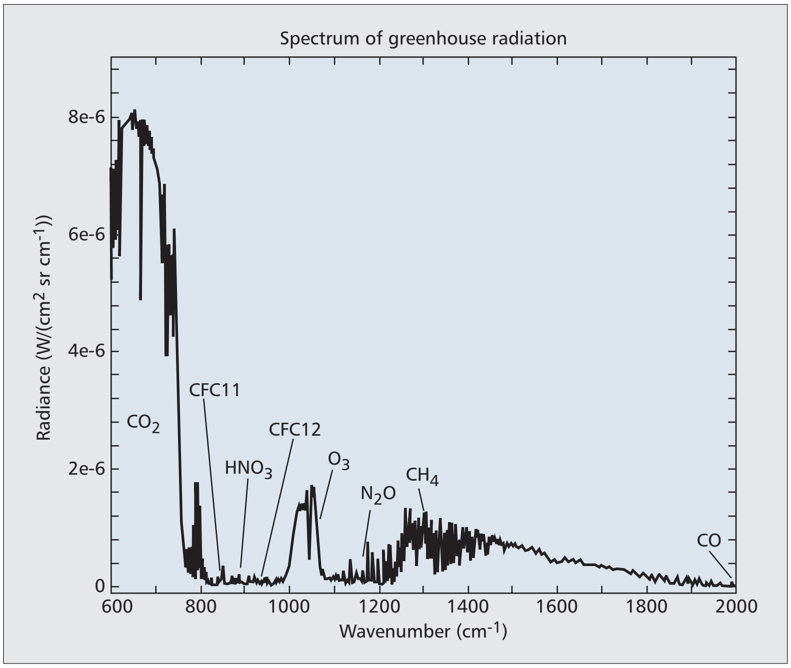


Figure 1. Typical infrared spectrograms [1].

We adhere to a strict format where meetings are kept under an hour and our project manager oversees the discussions to keep the meetings on schedule. The format of our meetings begins with general announcements for the entire team. After this, members are branched into sub-teams (which are divided by interest, specialized major, year of study, and critical importance of the sub-system) where they can conduct their own briefing.

Our members have access to numerous experimental facilities at York University, including a space instrumentation laboratory, spacecraft assembly clean room, electronics and machining facilities, spacecraft testing vacuum chambers, and vibration tables, which can all be found inside the Petrie Science and Engineering Building at The Centre for Research in Earth and Space Science.

NEXT STEPS

For the current CSDC competition our team plans to continue refining our preliminary nano-satellite design. Our team is looking to involve other educational institutions and amateur radio operators from around the world in an effort to make communication with the satellite possible at all times, as opposed to only when the satellite is passing over the operations and support center. It is hoped that by doing so we will not only allow for more data to be received and analyzed, but that people worldwide can experience communicating with a satellite in orbit. These users will not have access to the spacecraft software or sub-systems. However, they will be able to download data from the spacecraft and relay it via the Internet to the operations and support center.

We plan to demonstrate the practicality and flexibility of nano-satellites and how they can be used to produce useful data. Furthermore, our team is determined to seek expert opinions and advice from industry professionals and academics on any new technology that we can create using software or hardware. Due to our budget we may have to create our own elec-

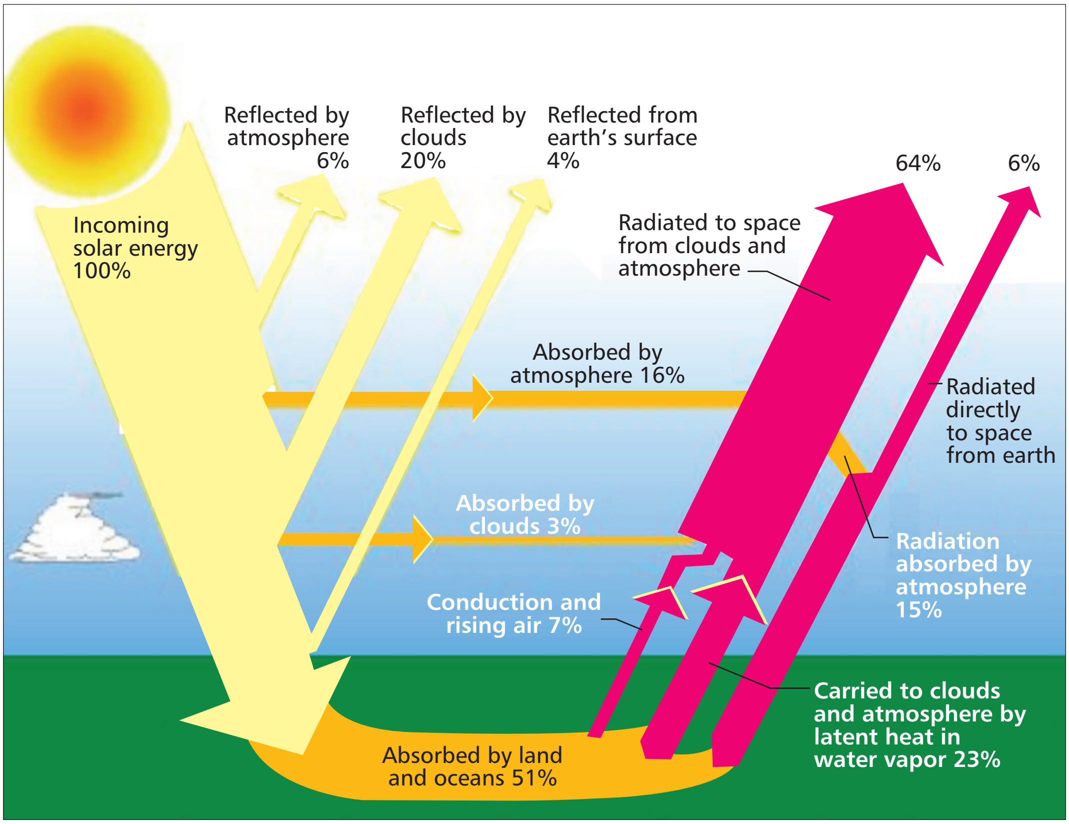


Figure 2. Earth's energy budget [2].

tronics, which is something we would like to minimize. There will also be some creation of our own products, such as software or hardware, that will undergo rigorous testing. With more funding, we hope to purchase more components with a high technology readiness level (TRL) that are considered 'off-the-shelf' commercial items that are already space-qualified like our payload.

Conclusion

Foremost, our team at York University will benefit through participation in the Canadian Satellite Design Challenge by learning the life-cycle design, manufacturing, building, and testing of a nano-satellite. Moreover, the practical experience the students will gain from being involved in space mission design and understanding how to work in a multidisciplinary project with a broader spectrum of skilled individuals in areas such as business personnel, human resources, and communications is invaluable. Furthermore, the new opening of the The Bergeron Centre for Engineering Excellence by the Lassonde School of Engineering is where we will have a satellite ground-station on the roof for establishing reliable communications, and it will definitely be a great assistance to our team (Fig. 3)

Figure 3. The Canadian Satellite Design Team at York University; L to R by last row: Manik Gupta, Ishfaaq Muhammad Jookun, Alex Bocaud, Tremayne Gomes, Thomas Giles, Gleb Sitigun; Saquib Ansan, Keith Menezes, Lucien Devon, Tetiana Sitiguina, Shamil Samigulin, Peter Dunsworth; Elisa Mesaroli, Sonal Ranjit, Bobby Ingino, Arthur Wong, Mohammed Kagawala; Hugh Chesser.

References

[1] W. F. J. Evans, Spectrum of Greenhouse Radiation, Digital Image, North West Research Associates, 30 Jan. 2006, Web, 01 Mar. 2015.

[2] J. Medigan, Earth's Energy Budget, Digital Image, National Aeronautics and Space Administration, 28 Jan. 2014, Web, 03 Feb. 2015.

Biographies

Tremayne Gomes (tremaynegomes1@gmail.com) is a fourth year English undergraduate major at York University. During his time at York he was involved in student activities, being a member of the York University Model United Nations and also the Canadian Satellite Design Challenge Team at York University. For the latter he was the director of public relations, which was his first experience in the field. For the club he did things such as updating social media, assigning tasks to assistants, constructing ideas for campaign videos, creating sponsorship packages, and constructing ideas for pitches. His goal is to work in a public relations firm.

Keith Menezes is a third year space engineering undergraduate major at the Lassonde School of Engineering at York University. As an undergraduate, he has balanced a rigorous course load with a number of extracurricular activities to enhance his skills for future career aspirations in the space industry. Keith is passionate about unmanned aerial vehicles, satellites, and sustainable energy. As president he led the team, drove growth within the organization, presented and facilitated all executive team meetings while addressing any concerns that were raised. As part of the engineering and science team he helped define system requirements, define project scope, identify risks to the project, and enforce the completion of satellite subsystem tasks.

Monitoring Glaciers from Space Using a Cubesat

Constance Fodé, Jacopo Panerati, Prescilia Desroches, Marcello Valdatta, and Giovanni Beltrame

ABSTRACT

The École Polytechnique de Montréal and the University of Bologna recently collaborated to develop a nanosatellite mission in the context of the 2012–2014 iteration of the Canadian Satellite Design Challenge, an inter-university competition intended for the development of space expertise among graduate and undergraduate students. The mission comprised two different scientific payloads: one aiming at monitoring climate changes in the Arctic, and the other addressing the need to reduce space debris. Here we report the organizational and technical challenges we faced, as well as the lessons we learned.

INTRODUCTION

Since 2012, the École Polytechnique de Montréal (EPM) in Québec, Canada, and the University of Bologna, in Italy, have been collaborating to design and build a nanosatellite for the Canadian Satellite Design Challenge (CSDC), specifically a triple-cubesat, or "3U" cubesat, that measures 10x10x35cm and weighs less than 4kg [1]. The cubesat's primary payload is an imaging system designed to support the University of Montreal's Geocryolab (Laboratory of Geotechnics and Geomorphology of Cold Regions) in studying the changes of the periglacial ecosystem of the Bylot and Baffin Islands, in Canada's Arctic. For this, it has a twofold objective: to create a photographic database to help estimate the reduction of the ice cover on the Bylot Island, and to acquire high-resolution images of selected Baffin Island's glaciers at specific times of the year. The cubesat's secondary payload is a passive deorbiting device based on a drag sail, and it is motivated by the growing concerns for the presence of space debris in low Earth orbit [2–4]. This payload occupies less than 0.4U of the 3U cubesat, and can be integrated onto any structure. It is a stand-alone subsystem with its own flight computer and power system, and it can work even in the case of failure of the other subsystems of the satellite. The deorbiting system respects the Inter-Agency Space Debris Coordination Committee's (IADC) 25-year guidelines concerning post-mission disposal [5]. Our objective is to test the satellite's capability to survive the launch phase and operate in the space environment.

Constance Fodé, Prescillia Desroches, Jacopo Panerati, and Giovanni Beltrame are with École Polytechnique de Montréal.

Marcello Valdatta is with University of Bologna.

ORIGINS

When PolyOrbite was founded in 2012, it was the first student society aiming at developing space applications at EPM. Prior to the start of the development of this satellite, very few PolyOrbite's members had taken any class on space technologies. The PolyOrbite's team (Fig. 1) was composed of undergraduates as well as graduate students in several engineering disciplines (including aerospace, mechanical, physical, computer, and electrical engineering). Professor Giovanni Beltrame from the Department of Computer Engineering supervised the team. At the beginning of the project, PolyOrbite's founders structured the organization into five major teams: structure, telemetry and data handling, power, telecommunication, and mission. It was only after the first year that a team was created to conceive an attitude control prototype. The assignment of leadership positions within each team is usually based on the motivation of the members rather than on the degree of previous expertise. Thus, a team leader has the responsibility to select competent students and meet with them once a week. On the other hand, PolyOrbite's management is composed of a project leader and a financial director who meet with the team leaders every two weeks to ensure the progress of all the satellite's subsystems. As for the external help that PolyOrbite required for its first time being entered in the CSDC, we worked with MDA, a Canadian space missions company in Montreal, in order to prepare for our critical design review (CDR), but we mostly found the resources we needed by asking professors working at Polytechnique.

COLLABORATION WITH THE UNIVERSITY OF BOLOGNA

After attending a presentation of PolyOrbite, the director of the CSDC program, Larry Reeves, suggested that the students might like to participate in the CSDC, and benefit from the hands-on learning opportunity. The director also proposed to the University of Bologna students that they join the CSDC program, to exploit their experience on educational space missions (in particular, on sounding rockets, space debris

0163-6804/15/$25.00 © 2015 IEEE

Figure 1. The team leaders of PolyOrbite for the new iteration of CSDC 2014-2016. From left to right: Q. Michaud, J. Chann, J-R. Roy, A. Guay, M. Lalonde, P. Desroches, C. Fodé, and Z. Barrou Dummont.

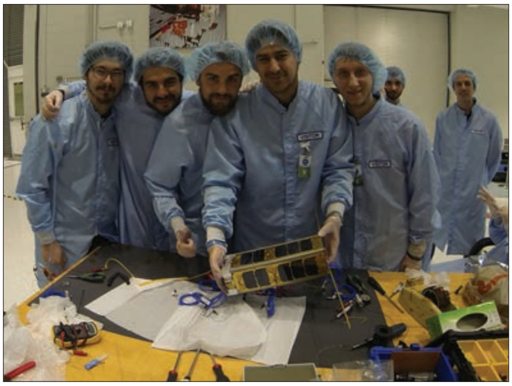


Figure 2. Students from the University of Bologna with the cubesat at the David Florida Laboratory spacecraft test facility. From left to right: A. Locarini, S. Naldi, N. Bellini, M. Valdatta, D. Rastelli.

Figure 3. Members of EPM, University of Bologna and the jury of CSDC for awards ceremony (Ottawa).

research, and cubesat subsystems), and to help PolyOrbite with its first experience in satellite design and manufacturing. Thus, the collaboration between EPM and the University of Bologna was born. The team from Bologna (Fig. 2) was also supported by the Italian company NPC (New Production Concept).

Development and Testing

Space represents a challenging environment for electronic computers. Therefore, while designing the command and data-handling subsystem of our cubesat, we knew that in order to deliver a reliable system, we had to mitigate the effects of particle radiation. For our primary payload, the imaging system, we developed a distributed solution composed of a central space-proof processing element (acquired from Pumpkin, Inc.), running the multi-tasked real-time operating system "Salvo" [6], and a secondary commercial off-the-shelf (COTS) board hosting FPGA fabric to perform computational-intensive, but non-critical, payload tasks (i.e. the image processing). The secondary board was designed to allocate FPGAs based on different technologies to also compare their resilience to the effects of radiation. The secondary payload, the deorbiting system, was composed of a deployable structure containing the drag sail, the opening system actuated by a hot wire cutting system, the on-board computer of the payload composed of three real-time clocks (for redundancy), and a microcontroller to determine the activation time of the system. The deorbiting system was also provided with its own battery pack to test the stand-alone philosophy. In the summer of 2014, at the David Florida Laboratory (DFL) in Ottawa, the final cubesat design underwent, and passed, the vibration and functional tests to simulate the launch environment.

Space represents a challenging environment for electronic computers. Therefore, while designing the command and data-handling subsystem of our cubesat, we knew that, in order to deliver a reliable system, we had to mitigate the effects of particle radiation.

Conclusions and Legacy

PolyOrbite is a now a major asset of EPM, as it explores such a novel and promising field of study. Moreover, because the secondary payload was developed in collaboration with the University of Bologna, one of PolyOrbite's proudest accomplishments is to have designed and built a satellite with students on the other side of the world. Students at both universities have benefited from this international collaboration, learning how to work together despite the language (English was everyone's second language), cultural, and time discrepancies (Fig. 3). The CSDC was especially useful in establishing new connections and gaining practical experience. Thanks to the CSDC, it was possible to attend several workshops, to present our work at international conferences, and to visit important Canadian space companies and advanced laboratories such as the DFL and those at MDA. It was possible to experience all the aspects of a space mission, from the managing to the test and production, learning how to work with the stringent constraints of a space mission. Eventually, the Italian team had the benefit of being hired by their sponsor company NPC to continue the project as a private developer.

Acknowledgments

We would like to express our gratitude to all the students from the University of Bologna and from École Polytechnique de Montréal involved in the technical society PolyOrbite, as well as the organizers of the Canadian Satellite Design Challenge. Moreover, we would like to thank all of our sponsors, without whom we would not have been able to achieve our goal: Pumpkin, INC, Fondation ASéQ, FAIE, Devcorp, MDA, Desjardins, Fasken Martineau, École Polytechnique de Montréal, AGI (Software for Space, Defense and Intelligence), NPC New Production Concept, and the University of Bologna.

REFERENCES

[1] The Canadian Satellite Design Challenge, Rules and Requirements; available: http://www.geocentrix.ca/images/CSDCFiles/CSDCRules2a.pdf.

[2] J.-C. Liou, N. L. Johnson, and N. M. Hill, "Controlling the Growth of Future LEO Debris Populations with Active Debris Removal," *Acta Astronautica*, vol. 66, nos. 5–6, March–April 2010, pp. 648–53.

[3] K. T. Nock, K. M. Aaron, and D. McKnight, "Removing Orbital Debris with Less Risk," *J. Spacecraft and Rockets*, vol. 50, no. 2, March 2013, pp. 365–79.

[4] D. J. Kessler and B. G. Cour-Palais, "Collision Frequency of Artificial Satellites: The Creation of a Debris Belt," *J. Geophysical Research*, vol. 83, no. A6, June 1978, pp. 2637–46.

[5] IADC Space Debris Mitigation Guidelines, IADC-02-01, Revision 1, Sept. 2007; available: http://orbitaldebris.jsc.nasa.gov/library/IADCMitigationGuidelinesRev1Sep07.pdf.

[6] Products Salvo; available: http://www.pumpkininc.com/.

BIOGRAPHIES

CONSTANCE FODÉ (constance.fode@gmail.com) received her Technical University degree in material science and engineering from the University of Blois (Blois, France) in 2013. Currently she is a mechanical engineering student at École Polytechnique de Montréal, Canada, where she is the project manager of PolyOrbite for its second iteration.

JACOPO PANERATI received his M.Sc. degree in computer science from the University of Illinois at Chicago (Chicago, IL) in 2012, the Laurea degree in computer engineering from Politecnico di Milano (Milan, Italy) in 2006, and the Laurea Specialistica degree in computer engineering again from Politecnico di Milano in 2011. Currently he is a Ph.D. candidate and Vanier scholar in computer engineering at the École Polytechnique de Montréal (Montréal, QC, Canada). His research interests include machine learning, artificial intelligence, and real-time and embedded systems.

PRESCILIA DESROCHES will graduate from aerospace engineering at École Polytechnique de Montréal, Canada in 2016. She has been the financial director of PolyOrbite since 2014. Currently she is in a student exchange program in mechanical engineering at the University of Compiègne (Compiègne, France) for a semester.

MARCELLO VALDATTA is an aerospace engineering student at the University of Bologna, a member of the former Space Robotic Group of University, and involved in various projects including the ESA Rexus/Bexus programme (Rexus12 with Redemption experiment). Thanks to the Erasmus Placement program he worked at the Space Science Technology Institute of Lithuania on UAVs technologies. He took part in different editions of the Unisec Mission Idea Contest, winning the 3rd edition. Currently he is working at NPC on cubesat subsystems and technologies.

GIOVANNI BELTRAME received the M.Sc. degree in electrical engineering and computer science from the University of Illinois, Chicago in 2001, the Laurea degree in computer engineering from the Politecnico di Milano, Italy in 2002, the M.S. degree in information technology from CEFRIEL, Milan in 2002, and the Ph.D. degree in computer engineering from the Politecnico di Milano in 2006. He worked as an engineer at the European Space Agency until 2010, and he is currently an assistant professor at École Polytechnique de Montréal, Canada, where he directs the MIST Laboratory. He has published more than 40 papers in international conferences and journals, he is on the organizing committee of several international conferences, and he is principal investigator on multiple projects funded by government and industry.

Biological Investigations Using a Triple-Cubesat

Chantelle Dubois, Pawel Glowacki, and Ahmad Byagowi

ABSTRACT

The University of Manitoba Space Applications and Technology Society is a student group participating in the Canadian Satellite Design Challenge (CSDC). For the 2012-14 iteration of the CSDC, a diverse team of undergraduate and graduate students worked together to develop T-SAT2, a triple-cube nanosatellite weighing under 4 kg. This satellite contains a biological payload with the purpose of exploring the survivability of tardigrades in low Earth orbit. Through participation in the CSDC, student members developed skills not included in their respective curriculums and became better prepared for their careers in industry and academia.

INTRODUCTION

The University of Manitoba Space Applications and Technology Society (UMSATS) is a student group dedicated to designing, building, testing, and eventually launching a triple-cubesat into low earth orbit (LEO). UMSATS set out to develop an experiment involving tardigrades (Fig. 1), a micro-animal, that may hopefully build evidence for the theory of Panspermia, the theory that the first life on earth originated from elsewhere in the cosmos [1]. Tardigrades' robustness makes them an ideal candidate for this experiment [2]. By attempting to test the hypothesis of whether tardigrades can survive interplanetary travel (the basis of Panspermia), UMSATS hopes to provide insight into whether it is possible that life may have originated elsewhere.

The experiment developed by UMSATS will send tardigrades, genetically modified with enhanced green fluorescent proteins, in a cryptobiotic state into LEO inside a purpose-built chamber. Once in orbit, the chamber will be flooded with water, rehydrating the tardigrades and their food sources. An onboard camera with a custom lens designed by Acryl Designs will monitor the tardigrades for signs of life (moving toward food sources, reproducing, etc). Image processing and statistical analysis will be used to determine whether the tardigrades are flourishing while exposed to the vacuum of space.

The authors are with University of Manitoba.

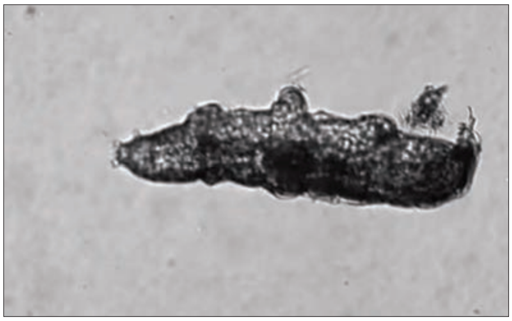


Figure 1. A tardigrade (by UMSATS payload team).

ORIGIN

UMSATS was founded as a student group shortly after the announcement of the Canadian Satellite Design Challenge (CSDC) in 2010. The student group began small, with a membership of 10 students, who then began recruiting efforts. UMSATS has grown rapidly, and has had more than 200 students involved in some capacity with the project. Team members have been recruited from almost every faculty on campus, including engineering, science, business, education, arts, and fine arts. Members range from first year university students in undergraduate programs to Ph.D. candidates. UMSATS emphasizes the need for diversity within the team, as a broad variety of skill sets are essential to make the project successful. UMSATS makes an effort to ensure that any student who wishes to be involved has an opportunity to contribute. Students were recruited by word of mouth from friends involved in the project, information tables set up on campus, open-house events, and through social media.

DEVELOPMENT AND TESTING

From the beginning of the CSDC, it was understood by the UMSATS team that designing a nanosatellite with a complex experimental payload, such as the tardigrade experiment, would be challenging. In preparation for the challenge, UMSATS members organized themselves into a hierarchal structure consisting of an advisory

Aside from the general complexity most space projects have, the nanosatellite format is an emerging field of satellite technology, and many of our intended goals have not previously been accomplished on this technology.

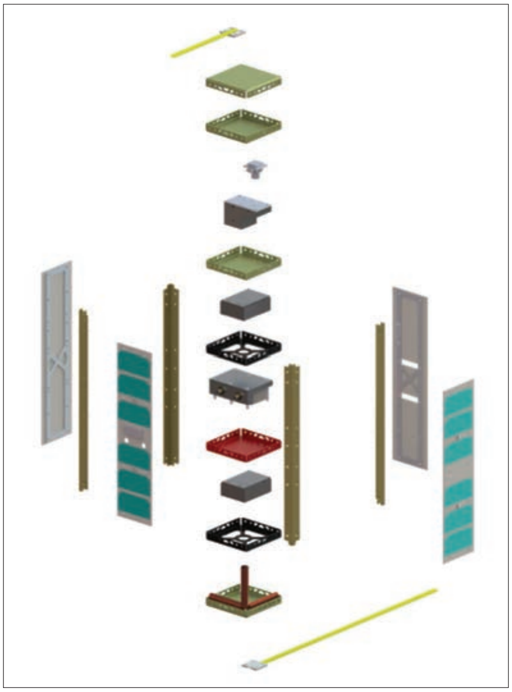


Figure 2. Exploded diagram of T-SAT2.

committee, an executive committee, subsystem leads, and general group members. These groups included the following subsystems: power design, communications design, mechanical design, payload, command and data handling, and attitude determination and control. Each group member was responsible for the delivery of different critical components of the satellite, including different technical design aspects of the satellite, project management, and fundraising. Additionally, UMSATS was also given access to resources including full memberships to Winnipeg's makerspace Assentworks, professional advice from senior engineers at Magellan Aerospace, as well as support from faculties such as the Faculty of Engineering and the Faculty of Science.

At the time of the formation of the team in 2010, none of the student members had experience with developing a satellite. The process of the development of the T-SAT1 satellite for the 2012 CSDC environmental testing was completed in parallel with learning and researching the design process of space systems. By 2012, some members of the team remained and their experience was carried forward for the development of T-SAT2 (Fig 2). The result was a more ambitious and sophisticated satellite design.

In May 2014, UMSATS travelled to Ottawa, Ontario for environmental testing at the David Florida Labs (DFL) (Fig. 3). The team members representing UMSATS were given a tour of the facility where many of Canada's satellites have undergone testing, and went through vibrations testing with T-SAT2. During the experience in DFL, the team learned about the importance of assembly, integration, and testing procedures in order to limit risks and ensure that the assembly is completed in a timely manner.

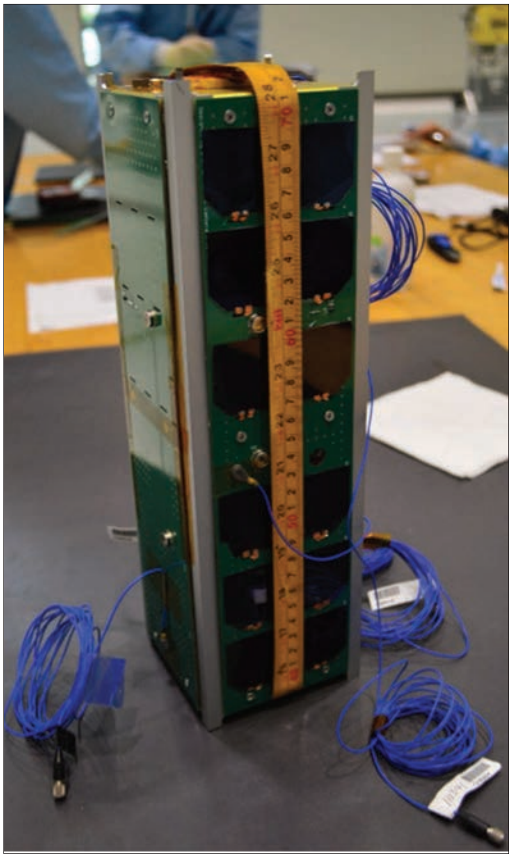


Figure 3. UMSATS' T-SAT2 being prepared for testing at the David Florida Lab.

Legacy

The University of Manitoba has benefited from the creation of UMSATS and participation in the CSDC by creating an on-campus space-program and diversifying opportunities for students. Many students who participate in the project get to see first-hand how a space project is developed from the initial stage to the delivery of the final product. This includes learning about design, prototyping, management, and cost-benefit analysis. Participation in the CSDC also benefits industry by producing newly graduated engineering students with the experience of working on a complex engineering project. The CSDC, like many competitions, is an excellent motivator for students to push themselves and use their creativity and technical skills to overcome engineering problems.

Aside from the general complexity most space projects have, the nanosatellite format is an emerging field of satellite technology, and many of our intended goals have not previously been accomplished on this technology. For the second iteration of the CSDC, the UMSATS team was the only competing team to have a biological payload. As a result of the UMSATS tardigrade experiment, the Faculty of Science has used the experiment in a special topics biology course in which students spend the semester carrying out research for the project. As of January 2015, these students have injected tardigrades with enhanced green fluorescent protein and enhanced red fluorescent proteins.

Additionally, the Department of Entomology has agreed to study tardigrade behavior and habits to help UMSATS assess their environmental needs for when they are launched into LEO. UMSATS is currently a participant in the third iteration of the CSDC, and has participated in all iterations of the competition to date.

CONCLUSION

The CSDC has been a valuable experience for participating universities. The competition has challenged students across Canada by engaging them in a complex project, providing an opportunity for them to apply their knowledge, and use creativity to develop a nanosatellite. For the University of Manitoba Space Applications and Technology Society, students have gained experience outside of their regular curriculums by working on an interdisciplinary team in both leadership and team member roles.

As the result of the development of T-SAT2, UMSATS members are now developing T-SAT3 with an even more in-depth understanding of the design process for space systems. So far, the design process for T-SAT3 has been more ambitious and focused than previous iterations, showing that the students are learning and passing on their knowledge.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contribution of the team members: Alex Gerry, Andre Marchildon, Andrew Bredin, Brent Schellenberg, Bryan Drobot, Bryce O'Donnel, Cass Erdelyan, Chantelle Dubois, Chang Zhou, Chris Rutley, Curtis Einarson, Dario Schor, Darwin Froese, David Stewart, Denise Geiskkovitch, Devon Hudson, Elias Klassen, Erik Johnson, Graham Rigby, Greg Linton, Ian Swintak, Joel Gamache, Joshua Jung, Joshua Pruden, Katrina Mae Soriano, Kenton Neustaeter, Majid Ostadrahimi, Maryna Tsybulska, Mathias Pielahn, Matt Driedger, Meiting Li, Michael Kleiber, Mike Lambeta, Nico Richard, Patrick Giesbrecht, Pawel Glowacki, Qian (Bill) Song, Robin Swanson, Rodin Raju, Ryan Wieler, Sailen Govind Kara, Sakib Rahman, Sean Iftody, Serge Broeska, Tong Duan, Valeria Pevtsov, Viridiana Urena Ramirez, Witold Kinsner.

In addition, the authors would like to acknowledge the sponsors of the team: Magellan Aerospace, Analytical Graphics Inc., Texas Instruments, SolidWorks, Altium, University of Manitoba Faculty of Engineering, University of Manitoba Faculty of Science, University of Manitoba Engineering Endowment Fund, Friends of Engineering, StandardAero, Assentworks, Princess Auto, Red River College, Parker Hannifin Corporation Canada, IEEE Canadian Foundation, University of Manitoba Student Union, Acryl Design Limited, University of Manitoba Engineering Society, Younes Sleep Technologies.

REFERENCES

[1] H. Kamminga, "Life from Space: A History of Panspermia," *Vistas in Astronomy*, vol. 26, no. 2, 1983, pp. 67–86.

[2] D. D. Horikawa *et al.*, "Analysis of DNA Repair and Protection in the Tardigrade Ramazzottius varieornatus and Hypsibius dujardini after Exposure to UVC Radiation," June 6, 2013, 8(6); avaliable: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0064793.

BIOGRAPHIES

AHMAD BYAGOWI received his Ph.D. from the Institute of Handling Robots and Technology (IHRT) at Vienna University of Technology (Austria). For his post-doc he was invited to the Biomedical Engineering Lab (UMBME) at the University of Manitoba. In 2011 he was introduced to the student group UMSATS. Currently he is the team lead.

PAWEL GLOWACKI completed his bachelors degree in electrical engineering from the University of Manitoba in 2011. He is currently a part-time masters student in electrical and computer engineering at the University of Manitoba. His thesis is related to designing and building an infrared spectrometer using MEMS for incipient fault detection in high voltage equipment. Pawel was a key member of the power subsystem for the UMSATS TSAT1 mission, and the team documentation lead and ADCS member for the TSAT2 and TSAT3 missions.

CHANTELLE DUBOIS (chantelle.dubois@umsats.ca) is an undergraduate electrical engineering student at the University of Manitoba. She joined UMSATS in 2013 as the outreach lead, and has been the business lead since 2014.

As the result of the development of T-SAT2, UMSATS members are now developing T-SAT3 with an even more in-depth understanding of the design process for space systems. So far, the design process for T-SAT3 has been more ambitious and focused than previous iterations, showing that the students are learning and passing on their knowledge.

GREEN COMMUNICATIONS AND COMPUTING NETWORKS


Jinsong Wu


John Thompson


Honggang Zhang


Daniel C. Kilper

Environmental issues are increasingly becoming one of central concerns in global societies. The production, application, and disposal of information and communication technology (ICT) platforms and relevant systems may introduce significant environmental problems. These include increasing environmental pollution, the exhaustion of natural resources and energy sources, which motivate us to reduce the negative environmental impact of ICT. On the other hand, ICT may also provide strong tools to solve, mitigate, or reduce environmental problems introduced by a wide range of factors. These topics may be considered as the core contents of "green ICT." Green ICT may profoundly impact the research, development, manufacturing, usage, and disposal processes of ICT systems and applications. The main motivation of green ICT is to support environmental sustainability.

The May 2015, second, issue of the IEEE Series on Green Communications and Computing Networks is a continual effort to support green ICT, and includes six relevant articles addressing both non-energy and energy related green topics.

The article "Cyber-Physical Systems for Water Sustainability: Challenges and Opportunities," written by Z. Wang and *et al.*, provides an overview on an extremely important non-energy green topic: the challenges for ICT to support water resources. This article provides an overview of water cyber-physical systems for sustainability from four critical aspects: sensing and instrumentation, communications and networking, computing and control, and opportunities and design challenges of relevant techniques.

The article "Macro-Assisted Data-Only Carrier for 5G Green Cellular Systems," written by X. Zhang and *et al.*, advocates a macro-assisted data-only carrier system for small cell enhancement. This article proposes a macro-assisted data-only carrier (DoC) for future 5G networks from a green perspective, and the relevant approach is verified using a complete system and link-level simulation platform, achieving significant throughput improvement and energy efficiency gain.

The article "Is Green Networking Beneficial in Terms of Device Lifetime?," written by L. Chiaraviglio *et al.*, discusses one often overlooked aspect of green networking, which is that the additional impact of green methodologies on cost, reliability, and even the overall life cycle analysis may go into the carbon footprint. This article considers the case of sleep mode power cycles and their impact on the lifetime of network devices, including both optical and cellular network elements. The impact is shown to depend on both the hardware details as well as the algorithms and parameters that are used in the green network design.

The article "Energy-Efficient Infrastructure Sharing in Multi-Operator Mobile Networks," written by A. Antonopoulos *et al.*, describes energy saving approaches in wireless networks by switching off base stations. In particular, the authors focus on the enhanced energy savings that are possible through dynamic infrastructure sharing among multiple mobile operators that collaborate. Initial performance results are presented, and practical implementation issues are discussed.

The article "A Low-Cost Methodology for Profiling the Power Consumption of Network Equipment," written by A. Francini *et al.*, describes a novel approach for modeling the power consumption of network switches and routers. The approach requires minimal equipment and uses linear models to approximate the true behavior of devices under test. The initial model results in the article highlight the importance of future hardware platforms including packet-timescale rate adaptation.

The article "Cost-Aware Green Cellular Networks with Energy and Communication Cooperation," written by J. Xu *et al.*, discusses two cooperative approaches, energy cooperation and communication cooperation, to reduce energy costs and reliably supply time- and space-varying wireless traffic over cellular networks, and proposes joint energy and communication cooperation among the base stations for cellular networks.

ACKNOWLEDGMENTS

We would like to acknowledge the great support and help from Osman S. Gebizlioglu and Sean Moore, the new and

former Editors-in-Chief of *IEEE Communications Magazine*, S. Charis Scoggins, Managing Editor of IEEE Communications Society Magazines, Jennifer Porcello, Production Specialist, 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) of the IEEE Communications Society.

BIOGRAPHIES

JINSONG WU [SM] (wujs@ieee.org) is an associate professor in the Department of Electrical Engineering, Universidad de Chile, Santiago. He is the Founding Chair of the IEEE Technical Committee on Green Communications and Computing. He is an Editor of the *IEEE Journal on Selected Areas in Communications* Series on Green Communications and Networking. He was the leading Editor and co-author of the comprehensive book *Green Communications: Theoretical Fundamentals, Algorithms, and Applications* (CRC Press, 2012).

JOHN THOMPSON [SM] (john.thompson@ed.ac.uk) currently holds a personal chair in Signal Processing and Communications at the School of Engineering, University of Edinburgh, United Kingdom. He was deputy academic coordinator of the recent Mobile Virtual Centre of Excellence Green Radio project. He currently leads the European Marie Curie Training Network ADVANTAGE, which trains 13 Ph.D. students in the area of smart grid technology. He is also a Distinguished Lecturer on green topics for ComSoc in 2014–2015.

HONGGANG ZHANG [SM] (honggangzhang@zju.edu.cn) is co-director of the York-Zhejiang Lab for Cognitive Radio and Green Communications at Zhejiang University, China. He served as Chair of the Technical Committee on Cognitive Networks of ComSoc during 2011–2012. He was the lead Guest Editor of *IEEE Communications Magazine* Feature Topic issues on Green Communications. He was General Co-Chair of IEEE GreenCom 2010 and TPC Co-Chair of IEEE Online GreenComm 2014. He is the book co-editor/co-author of *Green Communications: Theoretical Fundamentals, Algorithms and Applications* (CRC Press, 2012).

DANIEL C. KILPER [SM] (dkilper@optics.arizona.edu) is with the University of Arizona. He served as the founding Technical Committee Chair of the GreenTouch Consortium, and was the Bell Labs Liaison Executive for the Center for Energy Efficient Telecommunications at the University of Melbourne, Australia. While at Bell Labs, he received the President's Gold Medal Award in 2004 and was a member of the President's Advisory Council on Research. He has served as the General Co-Chair of the IEEE Online Green Communications Conference 2014 and 2015.

Cyber-Physical Systems for Water Sustainability: Challenges and Opportunities

Zhaohui Wang, Houbing Song, David W. Watkins, Keat Ghee Ong, Pengfei Xue, Qing Yang, and Xianming Shi

ABSTRACT

Water plays a vital role in the proper functioning of the Earth's ecosystems, and practically all human activities, such as agriculture, manufacturing, transportation, and energy production. The proliferation of industrial and agricultural activities in modern society, however, poses threats to water resources in the form of chemical, biological, and thermal pollution. On the other hand, tremendous advancements in science and technology offer valuable tools to address water sustainability challenges. Key technologies, including sensing technology, wireless communications and networking, hydrodynamic modeling, data analysis, and control, enable intelligently wireless networked water cyber-physical systems (CPS) with embedded sensors, processors, and actuators that can sense and interact with the water environment. This article provides an overview of water CPS for sustainability from four critical aspects: sensing and instrumentation; communications and networking; computing; and control. The article also explores opportunities and design challenges of relevant techniques.

Zhaohui Wang, David W. Watkins, Keat Ghee Ong, and Pengfei Xue are with Michigan Technological University.

Houbing Song is with West Virginia University.

Qing Yang is with Montana State University.

Xianming Shi is with Washington State University.

[1] Here "green" indicates the application of ICT to address environmental sustainability challenges.

WATER SUSTAINABILITY

Water is the lifeblood of the planet. Safe and abundant water resources are critical to all three dimensions of sustainability: social, economic, and environmental. Given its close linkage to a number of key global challenges, including population growth, industrialization, and climate change, water sustainability is of critical importance for sustainable development in the modern world. Achieving water sustainability, in turn, requires achieving universal access to safe drinking water, sanitation, and hygiene; improving the efficiency of water use for economic purposes; strengthening equitable, participatory, and accountable water governance; improving wastewater management and protecting water quality; and reducing the risks of natural and human-induced water-related disasters. In industrialized nations, fresh water resources, i.e. ground water, lakes, rivers, and streams, must meet a number of (sometimes competing) water use needs, including domestic, commercial, industrial, agricultural, and energy generation. In addition, sustainability requires that water withdrawals and consumption for human uses do not compromise aquatic ecosystems.

Although water is a renewable resource, water availability and water quality may not be sufficient for desired water uses and ecosystem functioning. Deficits, or scarcity, may occur at seasonal time scales and at geographic scales ranging from community water systems to regional river basins and groundwater aquifers. In addition to overuse or scarcity, water resources are facing many other severe challenges, including contamination, aging infrastructure, lack of data for informed decision making, weak public awareness of water challenges, and inefficient water management strategies. Many of these challenges are due to global change issues, such as population growth, economic development, and climate change, which are expected to increase in the foreseeable future.

Given these critical and persistent water resource challenges, we advocate *"green[1]" applications of information and communications technology (ICT) to help achieve water sustainability* [1]. To this end, green ICT can provide many benefits, including improved water quality, more effective emergency response, and increased public awareness of environmental challenges. We envision technologies including communications and networking, sensor technology, hydrodynamic modeling, data analysis, and human-centered decision support systems to enable intelligently networked water cyber-physical systems (CPS). This article explores opportunities and design challenges of CPS for promoting water sustainability, including sensing and instrumentation, communications and networking, computing, and control. Given the diversity of water sustainability challenges, techniques discussed in this article could be further tailored to specific application requirements.

Water Cyber-Physical Systems

Cyber-physical systems (CPS) are intelligently networked systems with embedded sensors, processors, and actuators that are designed to sense and interact with the physical world (including human users), and support real-time, guaranteed performance in safety-critical applications, as defined in a CPS Vision Statement published in 2012 by the Federal Networking and Information Technology Research and Development (NITRD) Program's CPS Senior Steering Group. The interplay between the "cyber" and "physical" elements among the CPS is critical: sensing, networking, computing, and control need to be deeply integrated in every component of CPS, and the CPS components must be inter-operable with a concerted design. Just as the Internet transformed the way people interact with information, CPS is transforming the way people interact with engineered systems and promoting sustainability.

Water Cyber-Physical Systems (CPS)

A typical water CPS architecture is depicted in Fig. 1. The water CPS monitors water quality conditions in real-time, and detects water pollution events quickly. Hydrodyamic modeling is integrated with the real-time measurements from various water quality sensors to generate the predicted transport of pollutants in the water environment, which is fed back into the input of the water CPS for optimal decision making of reactive and proactive actions to respond to water contamination emergencies effectively, thus forming a close-loop water CPS. In addition, energy harvesting is incorporated for a sustainable water CPS.

To develop a water CPS, we need:

- Sensing, communications, and networking technologies to enable flexible, reliable, and high performance distributed networking within a water CPS, that provide an accurate and reliable model of the water infrastructure and enable time-aware and time-critical functionalities.
- Computing technologies such as computational modeling, data management, machine learning, and other tools to understand, address, and communicate water sustainability challenges.
- Adaptive and predictive hierarchical hybrid control technologies are critical to achieve tightly coordinated and synchronized actions and interactions in a water CPS that is intrinsically synchronous, distributed. and noisy.

Based on the CPS Vision Statement published by NITRD[2], Table 1 provides a comprehensive summary of key technologies for developing a water CPS.

Example Applications of CPS for Water Sustainability

The application of water CPS could increase the efficiency, reliability, security, and public confidence in different types of water systems, hence setting the path toward water sustainability. In recent years there have been tremendous advances in real-time water quality sensing, communications, and computing, which have great potential for ubiquitous environmental monitoring and interaction. Using these technologies, cost-effective infrastructures can collect *in-situ* data remotely and on a continuous basis, as well as store, communicate, analyze, and visualize it in real time. However, compared to other applications of CPS, there has been relatively limited research on CPS for sustainable water management. In this section we present two applications of CPS for promoting water sustainability.

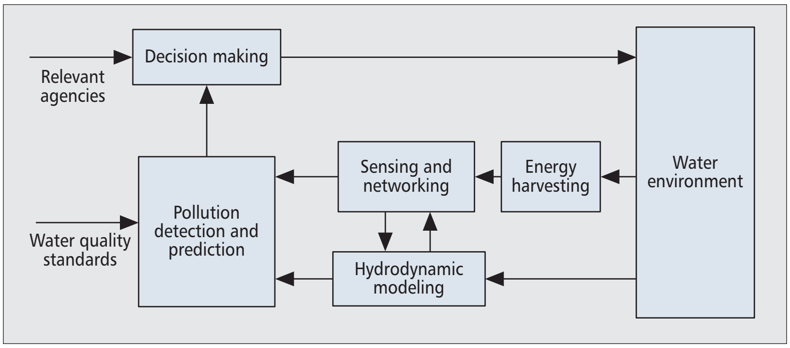


Figure 1. Architecture of a typical water CPS. Arrows indicate the flow of data, computing results, and control commands.

Water Distribution System Monitoring: Water distribution system monitoring is an important application of CPS in the water sector. In this application, the traditional real-time water quality monitoring performed at water treatment plants is extended to multiple locations within a water distribution system for contamination warning. The water CPS monitors baseline water quality conditions continuously in real-time such that a sudden change in water quality parameters can trigger a contamination warning. Benefits of this water CPS are improved water quality closer to the point-of-use, along with additional security for detecting intentional or unintentional contamination events within the system. A related CPS application is for control and mitigation of water losses in distribution systems (in the United States, there are an estimated 240,000 water main breaks per year). In this case, continuous monitoring of water pressure at various points in the system and automatic closure of valves can stop the flow of water to a broken section. Although the use of such on-line water distribution system monitoring and early warning systems is becoming common [2], water utilities are challenged by real-time data management and decision support, and there is a need for improved predictive models of contaminant transport [3].

Source Water Quality Monitoring: Source water quality monitoring can be critical for advance warning and emergency management. The water sector is vulnerable to a wide range of chemical, biological, and radiological releases. In the United States, there are nearly 14,000 oil spills reported per year, and such emergencies can affect large populations. For example, on January 9, 2014, crude 4-methylcyclohexane methanol (MCHM), a chemical primarily used

[2] Source: http://www.nitrd.gov/nitrdgroups/images/6/6a/Cyber_Physica_Systems_%28CPS%29_Vision_Statement.pdf

to clean coal, leaked from a storage tank near Charleston, West Virginia, and bled into a river upstream of a water-treatment plant. As a result, about 15 percent of the state's residents were advised not to drink the water.

Despite the need to safeguard water supplies and to protect public and ecosystem health, real-time monitoring of ambient environmental conditions is uncommon, primarily due to technological and cost limitations [4]. Some notable examples of ambient monitoring systems are a network of monitoring buoys on the Mississippi, Missouri, and Illinois Rivers developed by the National Great Rivers Research and Education Center (NGRREC); and the Intelligent River project pursued at Clemson University. Example projects monitoring aquatic ecosystem health (e.g. nutrients, invasive species) include the Jefferson project at Lake George, N.Y.; the Hudson River Environmental Conditions Observing System (HRECOS); and the River and Estuary Observatory Network (REON).

With remote, *in-situ*, and real-time water monitoring and warning systems in place, computing and communications technologies can facilitate the distilling of information from potentially vast amounts of data, as well as timely dissemination to decision makers and the public. Furthermore, hydrodynamic simulation modeling can provide forecasts of the scope of the emergency and contribute to effective response. Computing technologies can help keep track of numerous details in all stages of an emergency (forecast, warning, and response), helping people grasp the dynamic of a disaster and make quicker and better decisions.

In the following sections we will explore opportunities and design challenges of water CPS in four critical and complementary aspects, including sensing and instrumentation, communications and networking, computing, and control.

Key technologies	Role
Distributed sensing, ommunications and perception	Enable time-aware and time-critical functionality
Adaptive and predictive hierarchical hybrid control	Achieve tightly coordinated and synchronized actions and interactions in water CPS that is intrinsically synchronous, distributed, and noisy
Diagnostics and prognostics	Identify, predict, and prevent or recover from faults
Autonomy and human interaction	Facilitate model-based design of reactive water CPS that is used by humans
Validation, verification, and certification	Ensure high confidence in system safety and functionality
Abstractions, modularity, and composability	Enable water CPS system elements to be combined and reused while retaining safety, security, and reliability
Systems-engineering based architectures and standards	Enable efficient design and development of reliability systems while ensuring interoperability and integration with legacy systems
Integration of multi-physics models and models of software	Enable co-design of physical engineered and computational elements with predictable system behaviors
Cyber-security	Guarantee safety by guarding against malicious attacks

Table 1. Key technologies to develop a water CPS (based on the CPS Vision Statement by NITRD, 2012).

Sensing and Instrumentation

Sensing and Instrumentation

Advances in computer and electrical engineering in the past two decades have significantly reduced the size, cost, and power requirement of digital electronics. Low-cost, low-power devices with sufficient storage and data processing capabilities have led to the proliferation of wireless sensing in a wide range of applications. However, real-time and *in-situ* measuring of water quality and quantity has been progressing slowly, primary due to the harsh environment sensors and instruments have to endure.

Fostering water sustainability requires continuous monitoring of multiple parameters such as dissolved oxygen, flow rate, turbidity, conductivity, pH, and/or suspected chemical or biological pollutants. For an autonomous monitoring system, the sensors used in tracking these parameters need to be stable and accurate for a long period of time. Exposure to harsh weather and temperature changes can post a challenge for the long term durability of sensors, but proper sensor packaging and the use of high-performance electronic components can mitigate this issue. Another challenge is the biofouling effect, where aquatic fauna and flora aggregate on or around the sensors, preventing them from operating normally. Biofouling is the major limitation to longevity for many underwater sensors.

Many techniques have been deployed to reduce or limit biofouling effects. Physical cleaning such as scraping or wiping the sensors at fixed time intervals is effective but labor intensive. A common unmanned antifouling strategy is the long term deployment of biocides around the sensors. The biocides are usually embedded into surfaces of the sensor housings or coating layers protecting the sensor housings to be slowly leached out [5]. In addition, electro-mechanical principles have been employed to reduce biofouling [6]. Mechanical vibrations have been shown to remove biofouling, but the associated power requirement is high. Direct electrification of organisms has also been tested to remove fouling organisms from the sensor surface.

Low-cost sensing and computing devices have been a major driving force for a widespread use of wireless sensing techniques in the terrestrial radio environment. However, the cost of water sensing devices, such as underwater vehicles and acoustic communication modems, could be several orders of magnitude higher than their terrestrial radio counterparts, which significantly hinders the research progress and the broad use of underwater wireless sensing techniques. Research toward cost-effective instrumentation is essential for expanded use of remote, *in-situ*, real-time, and continuous water monitoring.

Energy Harvesting

To achieve a water CPS, it is desirable to use an array of sensors to sense environmental parameters in a wide water area and employ wireless devices to transmit water quality information back to a data collection or processing center. However, batteries become the bottle neck of such a system as they need to be replaced, after being exhausted, in a relatively short period of time. Replacing batteries for a large amount of sensors is time- and labor-consuming and likely to miss events that occur during replacement of batteries. Therefore, using energy harvested from water becomes the key component in achieving a sustainable water CPS.

An energy harvesting process could be used in a water CPS to convert energy from ambient sources in water to power a set of small wireless devices. However, traditional energy sources, e.g. solar power, thermal energy, and kinetic energy, may not be easily accessed in a water environment. For example, solar power and thermal energy are very limited in deep ocean, and kinetic energy may not exist in a lake.

We advocate an alternative energy harvesting technique, which harvests biological energy via microbial fuel cells (MFCs) that generate electricity through electrochemical reactions with a type of common and safe bacteria (manganese oxidizing microorganisms) ubiquitous in water [7]. As illustrated in Fig. 2, biocathode MFCs that harness the native population of manganese oxidizing microorganisms (MOM) abundant in natural waters will be made self-sufficient simply with the pretreated electrodes and nutrients continuously available in the environment. They can deliver potential renewable power (voltage up to 1.0V and current up to 1.2mA) in the aqueous water environment.

Due to the fluctuation of energy generated from MFCs, a power management system is needed to provide a stable and accurate DC output to power sensors. Unlike conventional or commercially available power converters for low voltage applications, the desired power convertors need to provide a stable and accurate output to sensors even when the voltage of the MFCs fluctuates in a wide range or drops to a very low level such as below 0.5V. Innovative and efficient design are essential to realize such a power management system.

Wireless Communications and Networking

Wireless Heterogenous Networks

Though direct access to the aquatic environment has been difficult, wireless communication technology makes unmanned water monitoring possible. Equipped with wireless (radio or acoustic) communication capabilities, sensing and actuation modules can be remotely controlled for specific tasks and transmit monitoring information to a centralized computer for analysis.

Especially about real-time monitoring of ambient aquatic conditions, the complexity of the aquatic environment calls for intelligently networked systems that could consist of several types of sub-networks. Figure 3 depicts a hybrid wireless networked system constituted by a surface buoy sub-network and an underwater sub-network with possibly mobile vehicles. The underwater sensing nodes can be anchored at the water bottom or float with water currents, and collect data samples at various water depths. Due to the large attenuation of electromagnetic waves in water, each underwater node has to rely on sound as the information carrier to communicate, and hence is often equipped with an acoustic modem. The surface buoys could have different types of sensors for collecting near-surface data samples as well as meteorological information, and be equipped with acoustic modems for communications with underwater nodes and radio-frequency (RF) modems for communications above the water surface with other buoys and a control center.

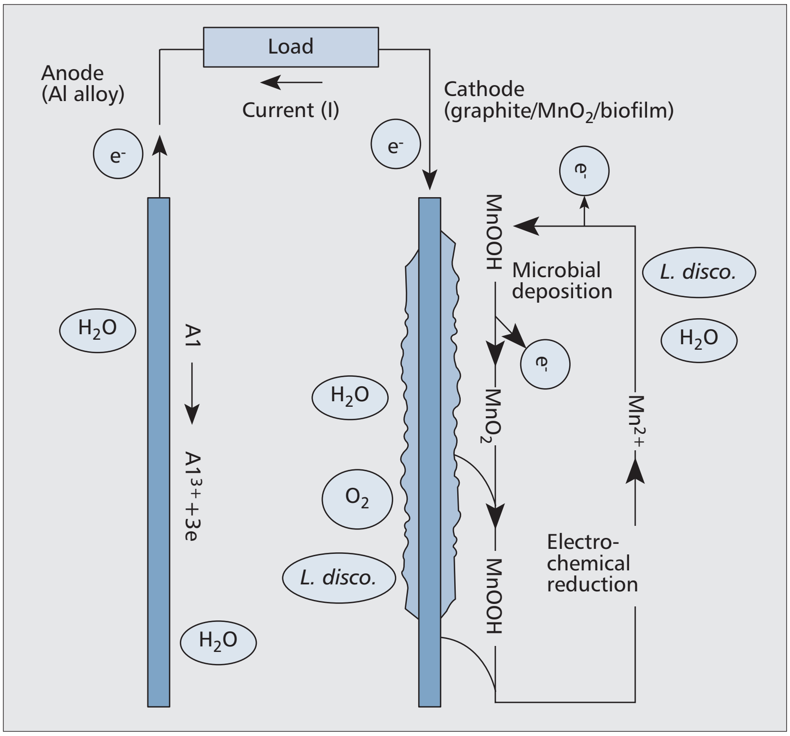


Figure 2. Schematic of the novel MFC, consisting of a sacrificial anode of aluminum alloy and a cathode of porous graphite covered by manganese dioxide and *L. discophora* biofilms [7].

Although data collection could be fulfilled by a surface buoy network through hanging sensors at different depths of the water column, incorporating an underwater acoustic sub-network yields several practical advantages. First, the underwater nodes do not disturb water surface activities (e.g. recreation and shipping). Second, they could maintain desired network functionalities in harsh weather conditions (e.g. stormy periods, and winter seasons with ice coverage). And finally, they reduce wireless network vulnerability to tampering or pilfering.

Challenges

Challenges in the above wireless heterogenous network design lie in a concerted network architecture and protocol suite under specific application requirements.

Hydrodynamic modeling can be brought to bear in several ways: to refine the grid size and time step of aquatic sensing, to predict future aquatic conditions, and to simulate multiple events or scenarios to improve decision making for water sustainability.

First, a majority of existing research on surface buoy networking adopts the Zigbee standard and operates in a clustered networking architecture. Data samples collected at each cluster head are sent to a remote monitoring center or the Internet via existing cellular/satellite networks. Despite intensive research in this area, long-range multi-hop data transmission along a river path or within a complex urban environment remains a challenging problem.

Second, compared with terrestrial radio networking, underwater acoustic networking is still in its early stages, primarily because of the distinct features of underwater acoustic channels:

- The underwater acoustic signal propagates at a speed (~ 1500 m/s) five orders of magnitude lower than the radio speed in air (~ 3×10^8 m/s), leading to very large signal propagation latency.
- Due to frequency-dependent sound attenuation, the underwater acoustic channel has a much lower bandwidth (~ kHz) than the surface RF channel (~ MHz).
- Underwater acoustic links often suffer large temporal variations due to environment dynamics.

Given the above differences between RF links and acoustic links, the RF connection among buoys can be regarded as a backbone for underwater acoustic networking, and multiple geographically distributed buoys could be deployed to reduce the end-to-end transmission latency and boost the end-to-end packet delivery performance. Leveraging the surface RF connection, many underwater networking protocols have been proposed; however, they lack sufficient experimental validation to demonstrate their performance in practical systems.

Third, the optimal deployment of heterogeneous nodes is another challenge in a water CPS. Depending on application scenarios (e.g. the water depth), the sensing nodes could be regarded as deployed on a 2-dimensional plane (e.g. water surface or bottom) or in a 3-dimensional water column [8]. Existing research often takes communication coverage, sensing coverage, end-to-end transmission latency, network reliability, and network resilience to node failures as performance metrics, and investigates the deployment of stationary networks with both surface buoys and underwater nodes, mobile underwater data collectors, or mobile surface buoys. Note that water parameters of interest often follow certain spatial-temporal distributions that can be simulated via hydrodynamic modeling. We anticipate more research on sensor node deployment that takes into account the spatial-temporal distribution of water parameters.

Computing Technologies

Computing technologies offer a wide range of tools and methodologies for understanding, addressing, and communicating water sustainability challenges. However, meeting these challenges will involve advances in a number of computing research fields. In this section we will examine the opportunities and challenges of computing technologies for water sustainability in two major areas.

Hydrodynamic Modeling

While real-time and *in-situ* water sensing could provide continuous measurements of water parameters of interest at specific sites, understanding their spatial and temporal distributions relies on computational modeling. Hydrodynamic modeling can be brought to bear in several ways: to refine the grid size and time step of aquatic sensing, to predict future aquatic conditions, and to simulate multiple events or scenarios to improve decision making for water sustainability.

Although hydrodynamic modeling has seen rapid development in the past three decades, it remains a challenge to resolve hydrodynamic processes at multiple scales. Coastal marine and fresh water systems typically consist of inner shelves, estuaries, and inland lakes, which are characterized by complex coastlines, series of islands and peninsulas, inlets, and extensive intertidal marshes. The irregular geometry has limited the ability of models to adequately resolve fine scale processes, which may be critical for predicting local conditions such as contamination of a water intake. Furthermore, the coupled dynamics among the hydrosphere, atmosphere, and lithosphere complicate the driving factors of the hydrodynamic simulation at a wide range of different scales. Nonetheless, driven by computing technology advances, such as adaptive mesh refinement, code parallelization, and GPU-accelerated computing [9, 10], hydrodynamic modeling has entered a period unprecedented toward resolving complex biophysical processes at scales appropriate for operational forecasting. Curvilinear and unstructured grid modeling can now resolve irregular geometries, and multi-grid approaches allow the resolution of small-scale features within a larger spatial domain [11].

Data-Driven Decision Making

In-situ aquatic sensing data and hydrodynamic modeling results can support decisions and subsequent actions to safeguard water sustainability. When appropriate, automatic decision making could be carried out through optimization algorithms; one example is automatic water pollution detection, warning, and response. Alternatively, decision support systems can help multiple parties, e.g. emergency managers, water utility companies, and policy makers and governments, distill information and coordinate with each other to make better-informed decisions.

Challenges in data-driven decision making are discussed in detail as follows.

Heterogeneity of Data Sets: Data sets from different observation systems could bear very different meanings and structures, and represent a wide range of spatiotemporal scales. For instance, in aquatic ecosystem monitoring, relevant data sets might include remotely sensed data from satellites, meteorological, and oceanographic information from surface observation networks, and data from underwater acoustic sensor networks.

Uncertainty of Hydrodynamic Models: Modeling real-world phenomena suffers from both aleatoric uncertainties (e.g. turbulent flows) and epistemic uncertainties (e.g. lack of physical measurements), which often lead to models performing worse with verification data sets than with calibration data sets. Though hydrodynamic modeling can be very accurate, many unknowns still exist, especially in complex hydraulic systems.

Sequential Decision Making: By definition, sustainable water systems are designed to operate over long time horizons. Sequential decision making is required based on continuous data samples, continuously improved models and sensor networks, and improved outcomes of decisions over time.

In the face of these challenges, machine learning and data mining techniques can be adopted for event detection and data-driven decision making. Machine learning techniques, including both supervised leaning (e.g. linear regression, neural networks, and decision trees) and unsupervised learning (e.g. clustering algorithms) have been developed for water quality event detection [12]. Data mining, data assimilation, and uncertainty estimation methods can be used with hydrodynamic models in real-time response as well as to continuously improve model calibration over time [13]. Probabilistic forecasting methods, now becoming common in water management at seasonal timescales, can be adapted for sequential decision making during contamination events [14]. For improved emergency response planning, scenario analysis and agent-based modeling hold promise for helping decision-makers account for complexity and hedge against contingencies [15].

Control Technologies

Control technologies, particularly feedback control and real-time control, are needed to design a water CPS with desired behaviors for sustainability. A CPS should have integrated protection, detection, and response mechanisms to be able to survive natural disasters, human error, and cyber attack without loss of function. This could be achieved with control technologies, which provide a systematic approach to designing feedback loops that are stable in that a CPS avoids wild oscillations, accurate in that a CPS achieves objectives such as target response times for service level management, and settle quickly to its steady state values.

Compared with feed-forward control, also called anticipative control, which is a control mechanism that predicts the effects of measured disturbances and takes corrective action to achieve the desired result, feedback control is a control mechanism that uses information from measurements to manipulate a variable to achieve the desired result and offers more advantages, including versatility and robustness. A water CPS featuring feedback loops, where physical processes affect computations and vice versa, is desirable. In a closed-loop CPS for real-time water quality monitoring and pollution detection, the water environment and hydrodynamic modeling affect the accuracy of pollutant transport prediction; the predicted pollutant transport affects the decision making about how to respond to water emergencies and the consequences of the water emergencies to the water environment.

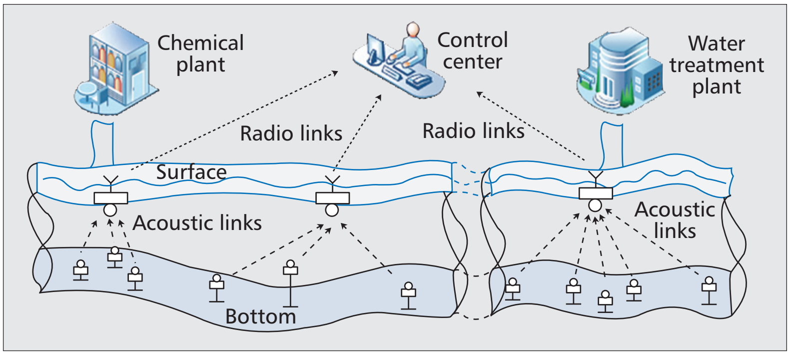


Figure 3. An illustration of a 3-dimensional hybrid wireless network architecture.

A water CPS for sustainability should also be a real-time control system in which its temporal properties are essential for reliability and correctness. Depending on the consequences that may occur because of a missed deadline, a real-time control system can be distinguished in three categories: hard, firm, or soft. In a water CPS, the correctness of the system behavior depends not only on the logical results of the computations, but also on the physical instant at which these results are produced. A missed deadline in a water CPS for real-time water quality monitoring and pollution detection is catastrophic, and a missed deadline in a water CPS for the control and mitigation of water losses can lead to a significant loss. Hence the predictability of the system behavior is the most important concern in a water CPS. The predictability is often achieved by either static or dynamic scheduling of real-time tasks to meet their deadlines. Static scheduling makes scheduling decisions at compile time and is off-line. Dynamic scheduling is online and uses schedulability tests to determine whether a set of tasks can meet their deadlines.

Conclusions

The world is facing severe challenges related to water sustainability. At the same time, information and communications technologies are rapidly advancing and are expected to capture and analyze data at a scale without precedent. With the potential to make remote sensing of the aquatic environment ubiquitous, cyber-physical systems can improve decision making with respect to many threats to water security, including the challenges discussed in this article (persistent water quality problems and emergency contamination events) and a host of others (e.g. availability and use of surface and groundwater, flood forecasting and response, and prediction of climate change impacts).

Real-time and *in-situ* data acquisition for better understanding of water sustainability challenges and identification of effective solutions requires crosscutting research. Given the pressing needs and the depth and complexity of the challenges, researchers must interact with water sector professionals to support improved decision making with the best-available science and technology, and water management agencies must continuously adapt their strategies as better tools become available.

Researchers must interact with water sector professionals to support improved decision making with the best-available science and technology, and water management agencies must continuously adapt their strategies as better tools become available.

References

[1] L. I. Millett and D. L. Estrin, *Computing Research for Sustainability*, Washington, DC: The National Academies Press, 2012.

[2] J. Hall *et al*., "On-Line Water Quality Parameters as Indicators of Distribution System Contamination," *J. Amer. Water Works Assoc*., vol. 99, no. 1, 2007, pp. 66–77.

[3] M. V. Storey, B. van der Gaag, and B. P. Burns, "Advances in On-Line Drinking Water Quality Monitoring and Early Warning Systems," *Water Research*, vol. 45, 2011, pp. 741–47.

[4] A. Jang *et al.*, "State-of-the-Art Lab Chip Sensors for Environmental Water Monitoring," *Meas. Sci. Technol*., 2011.

[5] J. P. Bearinger, S. Terrettaz, and R. Michel, "Sulfamer Block Copolymers Protection of Gold Surfaces," *Nature Materials*, no. 2, 2003, p. 257–64.

[6] C. H. Choi *et al*., "The Effect of Vibration Frequency and Amplitude on Biofouling Deterrence," *Biofouling*, vol. 29, no. 2, 2013, pp. 195–202.

[7] T. A. Nguyen *et al*., "Carbon and Steel Surfaces Modified by Leptothrix Discophora SP-6: Characterization and Implications," *Environmental Science & Technology*, vol. 41, no. 23, 2007, pp. 7987–96.

[8] D. Pompili, T. Melodia, and I. F. Akyildiz, "Deployment Analysis in Underwater Acoustic Wireless Sensor Networks," *Proc. ACM Intl. Workshop on Underwater Networks (WUWNet)*, Los Angeles, CA, USA, 2006.

[9] H. Y. Schive, U. H. Zhang, and T. Chiueh, "Directionally Unsplit Hydrodynamic Schemes with Hybrid MPI/OpenMP/GPU Parallelization in AMR," *Int'l. J. High Performance Computing Applications*, vol. 26, no. 4, 2012, pp. 367–77.

[10] Z. Shang, "Impact of Mesh Partitioning Methods in CFD for Large Scale Parallel Computing," *Computers & Fluids*, vol. 103, no. 1, 2014, pp. 1–5.

[11] C. Chen, H. Liu, and R. C. Beardsley, "An Unstructured Grid, Finite-Volume, Three-Dimensional, Primitive Equations Ocean Model: Application to Coastal Ocean and Estuaries," *J. Atmos. Oceanic Technol*., vol. 20, 2003, pp. 159–86.

[12] E. Brill, *Securing Water and Wastewater Systems*, Springer International Publishing, 2014, ch. 4: "Implementing Machine Learning Algorithms for Water Quality Event Detection: Theory and Practice," pp. 107–22.

[13] S. E. Haupt, G. S. Young, and C. T. Allen, "A Genetic Algorithm Method to Assimilate Sensor Data for a Toxic Contaminant Release," *J. Computers*, vol. 2, no. 6, 2007, pp. 85–93.

[14] E. Towler *et al*., "Incorporating Probabilistic Seasonal Climate Forecasts into River Management using a Risk-Based Framework," *Water Resources Research*, vol. 49, no. 8, 2013, pp. 4997–5008.

[15] E. M. Zechman, "Agent-Based Modeling to Simulate Contamination Events and Evaluate Threat Management Strategies in Water Distribution Systems," *Risk Analysis*, vol. 31, 2011, pp. 758–72.

Biographies

ZHAOHUI WANG [S'10-M'13] received a B.S. degree in 2006 from the Beijing University of Chemical Technology (BUCT), an M.S. degree in 2009 from the Institute of Acoustics, Chinese Academy of Sciences (IACAS), Beijing, China, and a Ph.D. degree from the University of Connecticut (UCONN), Storrs, all in electrical engineering. She has been with the Department of Electrical and Computer Engineering at the Michigan Technological University (Michigan Tech), Houghton, as an assistant professor since 2013. Her research interests lie in the areas of wireless communications, networking, and statistical signal processing, with a recent focus on wireless communications and networking in underwater acoustic environments.

HOUBING SONG [M'12-SM'14] received a Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, in 2012, and an M.S. degree in civil engineering from the University of Texas, El Paso, TX, in 2006. In 2012 he joined the Department of Electrical and Computer Engineering, West Virginia University, Montgomery, WV, where he is currently an assistant professor and the founding director of both the West Virginia Center of Excellence for Cyber-Physical Systems (WVCECPS), sponsored by the West Virginia Higher Education Policy Commission, and the Security and Optimization for Networked Globe Laboratory (SONG Lab). He worked with Texas A&M Transportation Institute (TTI), Texas A&M University System, as an engineering research associate in 2007. His research interests lie in the areas of cyber-physical systems, Internet of Things, intelligent transportation systems, wireless communications and networking, and optical communications and networking. His research has been supported by the West Virginia Higher Education Policy Commission.

DAVID WATKINS received B.S. degrees in civil engineering and engineering & public policy from Washington University in St. Louis in 1991. He received an M.S. degree in environmental health engineering in 1992 and a Ph.D. in civil engineering in 1997, both from the University of Texas at Austin. Following two years as a research hydraulic engineer for the U.S. Army Corps of Engineers, he joined the faculty of Civil and Environmental Engineering at Michigan Technological University in Houghton, Michigan, where he has been employed since 1999. His research focuses on modeling and analysis to support water resources sustainability, with recent projects addressing water scarcity in Florida and Texas and water pollution in the Great Lakes.

KEAT GHEE ONG received his B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Kentucky in 1997, 1998, and 2000, respectively. Currently, he is an associate professor in the department of Biomedical Engineering and director of Biotechnology Research Center at Michigan Technological University, Houghton, MI. His areas of expertise include wireless sensor networks, embedded sensors, implantable biosensors, and biological sensing materials and devices. He is also working on magnetism, RF, and acoustic-based sensors, magnetoelastic materials, and measurement technique and instrument automation. Ong is the associate editor of *Sensor Letters*, and has published over 100 publications in archival journals and conference proceedings.

PENGFEI XUE received his B.S. (2000) in applied mathematics from the East China Normal University, and his Ph.D. (2012) in marine and atmospheric system modeling and analysis (MASMA) from the University of Massachusetts. He is an assistant professor in the Department of Civil and Environmental Engineering at the Michigan Technological University (Michigan Tech). Prior to joining Michigan Tech in 2013, he worked as a postdoc associate in the department of Earth, Atmospheric and Planetary Sciences (EAPS), Massachusetts Institute of Technology (MIT) since 2012. His research is focused on the development and application of numerical models to problems in the coastal and regional oceans and the Great Lakes. His research interests include hydrodynamic modeling, dynamics of coupled ocean/lake-atmosphere models, ocean data assimilation, and estuary and coastal ocean circulation.

QING YANG [M'11], Ph.D, is a RightNow Technologies assistant professor in the Department of Computer Science, Montana State University (MSU). He received B.S. and M.S. degrees in computer science from Nankai University and Harbin Institute of Technology, China, in 2003 and 2005, respectively. He received the Ph.D. degree in computer science from Auburn University in 2011. His research interests lie in the areas of wireless sensor networks and trust assessment in online social networks. He serves as a TPC for several international conferences and on the editorial boards of several journals. His research has been funded by FHWA and DOT.

XIANMING SHI received a B.S. degree in 1993 from Beijing Institute of Chemical Technology, and M.S. degrees in 1996 from Tianjin University and in 2002 from Montana State University, respectively. He received his Ph.D. degree from the Institute of Chemistry, Chinese Academy of Sciences, Beijing, China in 1999. He is an associate professor in Civil & Environmental Engineering at Washington State University. From 2004 to 2014 he was a program manager and senior research scientist at the Western Transportation Institute and a research faculty at Montana State University. His research interests include smart and green infrastructure technologies, bio-energy harvesting, and sustainable materials.

Macro-Assisted Data-Only Carrier for 5G Green Cellular Systems

Xing Zhang, Jiaxin Zhang, Wenbo Wang, Yan Zhang, Chih-Lin I, Zhengang Pan, Gang Li, and Yami Chen

ABSTRACT

As the commercial operation of 4G systems is speeding up worldwide to meet the increasingly explosive growth of mobile Internet in the 2020 era, a great many R&D efforts targeting the next generation wireless systems (5G) have been launched. With densely deployed small cells in a heterogeneous network, the excessive signaling overhead of conventional carrier design greatly degrades system performance. A new carrier designed for small cells in the next generation system will be highly beneficial toward a green communication network. In this article, we propose a macro-assisted data-only carrier for future 5G networks from a green prospective. With the help of macro base stations, control channel overhead, and cell-specific reference signals for small cells can be minimized to achieve a pure-data carrier. Under this architecture, key procedures are designed including small cell identification and access, synchronization, handover, and small cell sleeping. Furthermore, to evaluate the potential of our proposed scheme, a complete system and link-level simulation platform according to the current 3GPP LTE standard is built. The simulation results show that our proposed systems, achieving significant performance enhancement of signaling overhead, can lead to more than 17 percent throughput improvement and 90 percent energy efficiency gain over the current LTE HetNet with on-off strategy implemented. The proposed scheme appears to be highly attractive as part of the future 5G green mobile networks.

Xing Zhang, Jiaxin Zhang, and Wenbo Wang are with Beijing University of Posts and Telecommunications.

Chih-Lin I, Zhengang Pan, Gang Li, and Yami Chen are with China Mobile, China.

Yan Zhang is with Simula Research Laboratory.

This work is supported by National Science Foundation of China (NSFC) under grant 61372114, by National 973 Program of China under grant 2012CB316005, by the Beijing Nova Program, by Beijing Higher Education Young Elite Teacher Project (No.YETP0434) and BUPT-China Mobile Joint Program.

INTRODUCTION

The fourth-generation (4G) mobile communication systems are commercially operated worldwide mainly based on Long Term Evolution (LTE) Release 8. As it evolved, LTE Release 9 was published in 2009 with the introduction of multimedia broadcast multicast services, home eNodeB (eNB), location-based services (LBS), and so on. The Third Generation Partnership Project (3GPP) Release 10, published in the beginning of 2011, made great progress through the introduction of carrier aggregation, multiple-input multiple-output (MIMO), and relaying. LTE Release 11 introduced the coordinated transmission and reception of base stations. LTE Release 12 (R12) mainly concentrated on small cell enhancement. However, there are still lots of open issues to be addressed. One of the most important topics in 5G networks is the ultra-dense network (UDN), where small cells will be ultra-densely deployed in hotspots. Thus, the downlink intercell interference will become much more serious, leading to huge energy consumption and intolerable delay. Although the concept of control and user plane (C/U) decoupling has been proposed for a long time, detailed approaches to splitting signals and channels are not discussed for the UDN scenario. In addition, existing new systems based on C/U splitting do not focus on the UDN scenario, and also face severe challenges in interference cancellation and energy saving. Thus, a new carrier designed for base stations in the ultra-dense scenario is vital for the next generation system to minimize interference and contribute to a "softer" and "greener" network [1].

In this article, based on the existed C/U split concept, we propose a novel macro-assisted data-only carrier (DoC) system, a set of procedures, and a new carrier toward a more flexible and energy-efficient network are designed for the first time. Furthermore, a complete system and link-level simulation platform is established, and simulation results show that the proposed scheme achieves significant performance improvement over the current LTE HetNet, helping to realize the user-centric "no more cell" network [1].

The rest of the article is organized as follows. In the next section the LTE/4G background and motivations to enhance the current LTE systems are illustrated. After that the concept of control and data decoupling under various architectures are analyzed. Then a novel macro-assisted system is proposed, and the channels and signalings are discussed carefully. Newly designed key procedures are introduced following that, and then we show the simulation results. Finally, we draw the conclusion in the last section.

 0163-6804/15/$25.00

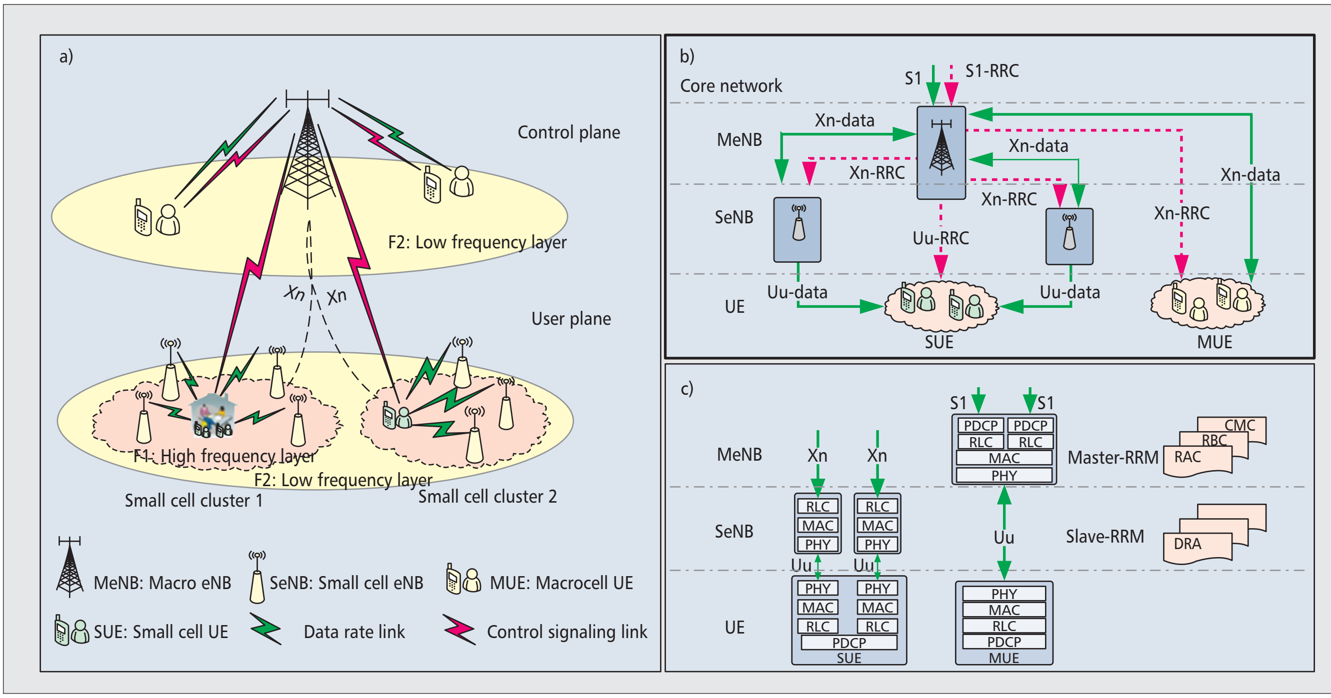


Figure 1. a) Data-only carrier (DoC) architecture and system model based on the split of the control and data planes: b) the interface in the DoC network; c) the protocol stack in the DoC network.

LTE/4G Background and Motivations

Although 3GPP has introduced heterogeneous networks with low-power nodes (LPNs), there are several crucial challenges to current 4G systems.

High energy consumption: As Cisco reported in [2], there will be over 10 billion mobile-connected devices by 2018, and much more energy will be consumed in the future. The improvement of energy efficiency under statistical quality of service (QoS) constraints is of great importance, as concluded in [3].

Huge overhead and interference of signals: The delay of a terminal to get access to the best candidate base station will be several hundred milliseconds in an LTE system [4]. This situation will become even worse and intolerable in a UDN network. Meanwhile, the combined overhead is up to 21 percent in LTE uplink and 28 percent in LTE downlink channels [5], which is quite impractical for small cells.

Frequent handover and signaling overhead: In the scenario of ultra-dense small cell deployment, the radius of small cell coverage is dozens of meters or even several meters [4]. In this case, handovers will be more frequent, and thus signaling overhead cost will be much heavier.

Control and Data Decoupling from a 5G Perspective

3GPP standardization has discussed and proposed solutions for 5G networks, among which small cell enhancement (SCE) and the new architecture design are the most promising technologies [6, 7].

Several new architectures have been proposed taking advantage of control and user plane separation [8–10]. In [8], the author gives the idea of cell zooming in a hyper cellular network (HCN) where the size of the small cell can be adjusted according to traffic load, user requirements, and channel conditions. Huawei proposes a two-layer network functionality separation scheme by taking UE state, and network functionality and signals into consideration [9]. In addition, [10] suggests a new clean slate system architecture to realize the signaling-data separation.

Although these architectures have achieved better performance compared to current networks, there are still open problems in the future ultra-dense scenario:

- Cell-level reference signaling interference still exists: The interference of signalings still exists, and the overhead occupies lots of the bandwidth and energy resources [8, 9].
- Intolerable delay in ultra-dense networks: The interference prevents UE's quick access and handover, which is intolerable for a future network such as vehicular networking.
- Limited opportunity to sleep: Most of the small cells in current networks waste great opportunities to go into sleep mode for periodically broadcasting public downlink signals and some information in public downlink channels.

Figure 1a shows one of the main scenarios, in which the LPNs in high frequency F1 (e.g., 3.5 GHz) are deployed in indoor or outdoor clusters, and a macro eNB (MeNB) is deployed in low frequency F2 (e.g., 2 GHz) as the anchor eNB. The wireless network is also split into the

	Network functionality	Current LTE	DoC system	Notes
Downlink signals	Small cell discovery	PSS/SSS	Removed	Network listening uplink SRS
	Time and frequency synchronization	CRS	Removed	Enhanced uplink SRS assisted downlink
	Candidate cell selection			Network listening uplink SRS
	PDCCH channel demodulation			Macro assisted PDCCH transmission signals
	Channel estimation			Enhanced uplink SRS assisted
	Demodulation reference signal	DRS	DRS	Channel associated signaling reserved
Downlink channels	System information bearer	PBCH	Removed	Macro assisted PBCH broadcast
	Frame control	PDCCH	ePDCCH	Located in PDSCH region
	Frame control	PHICH	PHICH	Indicate ACK/NACK information
	Frame control	PCFICH	Removed	Removed along with PDCCH
	Date information bearer	PDSCH	PDSCH	Downlink data bearer
Uplink signals	Uplink channel estimation	SRS	Enhanced	Enhanced in functionality and the orthogonal SRS number
	Channel demodulation	DM-RS	DM-RS	Channel associated signaling reserved
Uplink channels	Uplink control	PUCCH	Enhanced	Macro assisted with ideal backhaul
	Random access	PRACH	Removed	Macro assisted random access
	Data information bearer	PUSCH	PUSCH	Uplink data bearer

Table 1. Improvements of small cell in a data-only carrier (DoC) over current LTE systems.

control plane and user plane. Through the separation, the control signaling is handled by the MeNB layer with large coverage, while the data transmission is enhanced by the small cell eNB (SeNB).

SeNBs in the proposed system are designed toward pure "data-only" base stations with the assistance of an MeNB. It should be noted that data-only here does not mean no signals, but implies no more "fixed" overhead in the SeNB; the data-related control signals still remain. In this way, overhead of public system information and cell-level control signals can be reduced, and the interference from other SeNBs can be minimized for efficient small cell discovery and measurements. Furthermore, more resource blocks can be used to transmit useful data traffic, contributing to the increase of spectrum efficiency. Thus, we need to address two key issues:

- The methods to decouple the functionality of the wireless network between MeNB and SeNB
- The strategy to realize the decoupled function of SeNB by the pure data carrier with the assistance of an MeNB

In the next section, a detailed comparison between the proposed system and the current LTE system is discussed.

Macro-Assisted Data-Only Carrier Design

Under the C/U decoupling network architecture in 3GPP, each UE may have a single radio resource control (RRC) connection with an MeNB, and dual data connection of both an MeNB and an SeNB, as shown in Fig. 1. Fundamental procedures to realize the DoC system are given as follows:

Step 1: Decouple the architecture based on MeNB and SeNB collaboration.

Step 2: Analyze and split the functions of cell-level signaling and public channels in SeNB.

Step 3: Redesign the data-only carrier with MeNB assisted procedures.

The key features and characteristics of the proposed system are illustrated in Table 1, together with the differences between current LTE standards and macro-assisted DoC systems.

Functionality Separation between MeNB and SeNB

Following step 1, the collaboration between MeNB and SeNB is analyzed. The radio resource management (RRM) functionalities of MeNB and SeNB are different in DoC, systems as shown in Fig. 1b. The macrocell keeps a master RRM (M-RRM), and small cells are deployed

Network functionality	MeNB in DoC system		SeNB in DoC system	
Small cell discovery	E	Wake up small cell for listening	N	Listening SRS
Small cell selection	E	Based on SRS and UE requirement	×	Macro-assisted
Sync	✓	PSS/SSS	N	SRS
Information broadcast	E	Enhanced PBCH for SUE	×	Macro-assisted
Random access	E	Enhanced PRACH for SUE	×	Macro-assisted
Paging	E	Enhanced paging for SUE	×	Macro-assisted

✓: Preserved
×: Removed
E: Enhanced small cell signalings or channels
N: New design

Table 2. Functionality comparison of MeNB and SeNB in a DoC system.

with slave RRMs (S-RRMs) in Fig. 1c. The MeNB provides all necessary control information from the aspects of the control plane, such as the RRC S1-MME (mobility management entity) connection maintenance, security input, and non-access stratum (NAS) mobility information. The M-RRM functionality provides "slow" control for UE, such as radio admission control (RAC), radio bearer control (RBC), and connection mobility control (CMC), while the SeNB provides dynamic resource allocation (DRA) with power control.

For the user plane, the traffic to macrocell UE (MUE) beyond the coverage of SeNB is routed directly from the MeNB. However, the small cell UE (SUE), located in the ultra-dense region within the coverage of an SeNB, can get the traffic flow from both eNBs to achieve high data rates. In this way, all of the SeNB traffic flows are routed via the MeNB.

Downlink Signals for Data-Only Carrier

In this subsection, the comparison between LTE and our proposed scheme is shown in Table 2 as step 2. For better understanding and comparison, the names of signalings and channels of DoC are borrowed from LTE for explanation of similar functionality, but may not be exactly the same design as those of LTE.

PSS/SSS: Primary synchronization signals (PSS) and secondary synchronization signals (SSS), repeated every 5 ms, are used for UE to get the physical cell ID (PCI), time slot, and frame synchronization for this cell. These periodic system signalings can be totally removed by the network listening strategy. With an uplink sounding reference signal (SRS), transmitted along with normal data or RRC information from UE to MeNB in DoC systems, the UE can get uplink synchronization directly to a small cell.

CRS: As previously mentioned, CRS interference will become a serious problem for future 5G networks [4]. The function of time and frequency accurate synchronization and maintenance in CRS can be supported by uplink SRS signalings instead. In addition, the role of downlink channel quality measurement can be achieved by the calculation of uplink channel estimation, downlink transmission power, and other gains (e.g., BS and UE antenna gains) by utilizing the asymmetry between uplink and downlink channels, because of short distance (e.g., less than 100 m) between SUE and SeNB under typical dense urban scenarios.

DRS: DRS is the signal transmitted with data to demodulate useful data, which remains the same in future networks.

Downlink Channels for a Data-Only Carrier

Not only can the signals be removed for a carrier in an SeNB, but also the public channels with the assistance of an MeNB. In this subsection, several public downlink channels are analyzed and discussed.

Physical broadcast channel (PBCH): PBCH is used to broadcast the system information to all the terminals in the radio coverage area, and sends identification and access control parameters approximately every 40 ms [4]. Terminals accessing SeNB are within the coverage of MeNB, so the system information of SeNB can be transmitted by MeNB within PBCH of MeNB instead, and there is no interference to PBCH from different SeNBs.

Physical downlink control channel (PDCCH): PDCCH, located in the first n ($n <= 4$) orthogonal frequency-division multiplexing (OFDM) symbols, is utilized to transmit dynamic resource assignment information in LTE systems. Format and resource allocation related to DL-SCH and PCH, and hybrid automatic repeat request (ARQ) information related to DL-SCH are transmitted in this channel. In DoC systems, this information is called enhanced PDCCH (ePDCCH) and transmitted by PDSCH instead, together with the useful data.

Physical downlink shared channel (PDSCH): For the proposed system, PDSCH is almost the same as the one in a legacy eNodeB. This channel will allocate most of the resource blocks of the base station. As mentioned above, PDSCH and ePDCCH in DoC are transmitted together.

Other downlink channels: With the removal of PDCCH from DoC, PCFICH, the indicator of the number of symbols of PDCCH, can be removed as a consequence. In addition, downlink acknowledgment (ACK) and negative ACK (NACK) are also needed to represent the quality of the channel of UE to an SeNB and are only carried by the first symbol of each subframe.

Uplink Signals for a Data-Only Carrier

The two main signals in the uplink channel are the demodulation reference signal (DM-RS) and SRS, and these UE-specific signals should be kept in the DoC architecture. DM-RS is the reference signal used for uplink channel data demodulation transmitted with data traffic in PUSCH. SRS is also a UE-specific signal used

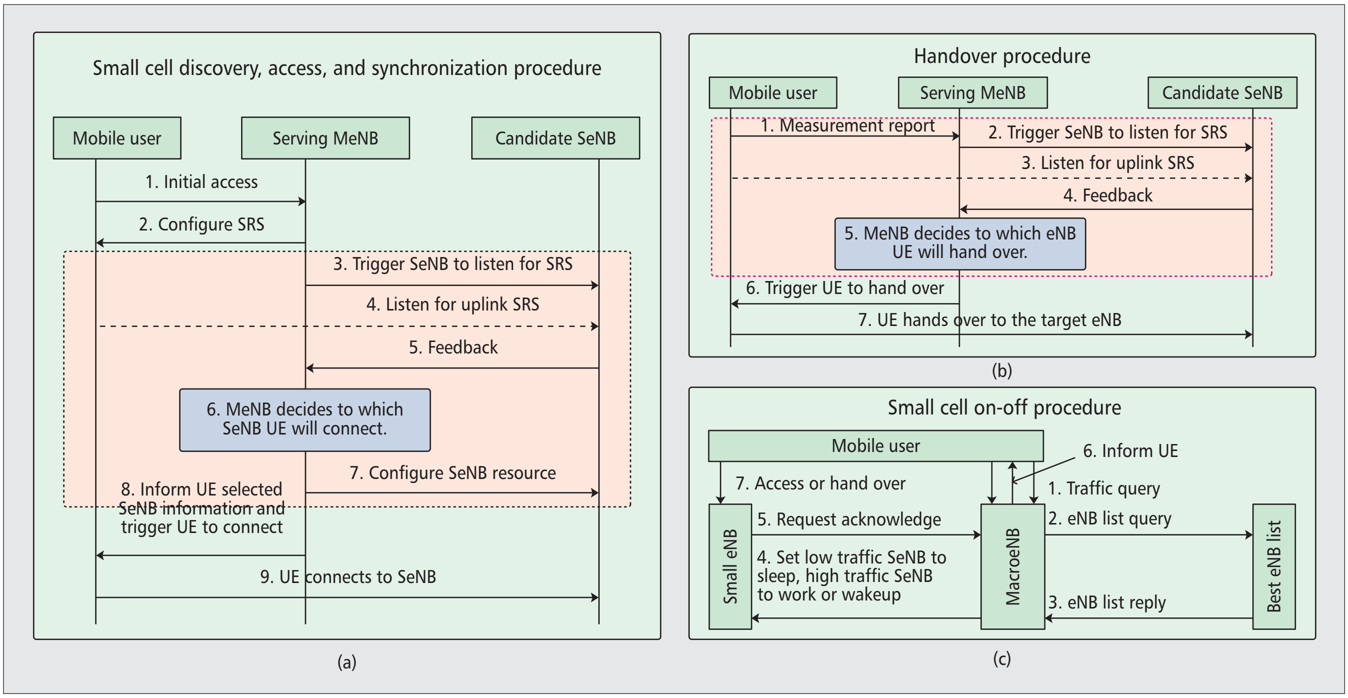


Figure 2. Procedures in DoC systems: a) small cell discovery, access, and synchronization; b) handover procedure; c) small cell on-off procedure. Note that the procedures in the red dotted region are the main differences between DoC systems and current LTE standards.

,for uplink channel quality measurements and channel estimation.

In a DoC system, the SRS will be enhanced through introducing additional functions of small cell discovery, uplink and downlink synchronization, and small cell selection, besides the function of uplink channel quality measurements and channel estimation. The number of SRSs is the key and big challenge for uplink SRS interference cancellation from UE to macrocell. Thus, the SRS in the future system should be enhanced by utilizing the multiplexing of SRS symbol sequence multiplexing in various domains to meet the demand in future networks.

Uplink Channels for a Data-Only Carrier

In DoC systems, ACK and NACK information on PUCCH in an SeNB may be removed and transmitted to a macrocell instead, if the backhaul is ideal (e.g., less than 2 μs in 3GPP standardization). On the contrary, these HARQ data can only be sent to an SeNB on PUCCH. As for PUSCH, it is also kept for future systems for uplink data transmission.

Security Issue and Application Scenario

Besides the discussion above, security and the application scenario are doubtless the significant issues to be considered. In a DoC system, the control and data transmission to SeNB are connected to the core network by the routing of the macro eNB instead of linking to the gateway directly. Thus, security on the Xn interface between the MeNB and the core network can be ensured.

The DoC network is aimed at the future 5G ultra-dense scenario, especially for centralized radio access network (C-RAN) architecture with baseband unit (BBU) pooling to support large-scale computing and storage. Furthermore, the procedure of pure-data transmission of SeNBs needs the assistance of an MeNB to maintain service to UE, which should be set as one mode of the configurations in small cells. In this way, the DoC small cells are somehow a "non-standalone and slim" design, and the extending DoC methodology in the standalone scenario is set as a future topic.

Key Procedures for a Macro-Assisted Data-Only Carrier System

In the DoC architecture, the procedures have been redesigned to realize the network's functionality as in step 3. We summarize the functions of MeNB and SeNB in a DoC system in Table 2.

Small Cell Discovery, Initial Access, and Synchronization

The initial access to small cells and synchronization are realized in the small cell discovery and access procedure. Different from the current LTE standard, where UE acquires the basic system information such as transmission bandwidth of a particular candidate cell by receiving the master system information (transmitted in a broadcast channel, specifically via a cell-specific reference signal). In a DoC system, the aforementioned procedures are only performed at the MeNB. For SeNB, the system information of the selected SeNB to which UE may connect is sent by the MeNB to UE. In this way, the PBCH of the MeNB should be enhanced, as shown in Table 2.

In the small cell discovery procedure, SRS is used for uplink listening. MeNB decides the best SeNB for UE, based on the measurement reports of SRS from the candidate SeNBs. This network-based method to find out the best serving SeNB is also used in the handover procedures.

After connecting to the candidate MeNB, the terminal will get time and frequency synchronization with the MeNB through the same procedure as in the current LTE system. If the UE could connect to an SeNB, it will initiate a non-contention-based random access supported by the MeNB. As the SeNB in a DoC system has no downlink control signals, the uplink listening is taken as the candidate solution. The proposed procedure is illustrated in Fig. 2a. After the MeNB has decided to which SeNB UE will connect, it will configure SeNB resources and inform UE of the system information of the selected SeNB and trigger the access procedure. Since the SeNB knows which resource block is allocated by the UE, UE will access the SeNB directly.

In the small cell discovery procedure, SRS is used for uplink listening. The MeNB decides the best SeNB for UE based on the measurement reports of SRS from the candidate SeNBs. This network-based method to find out the best serving SeNB is also used in the handover procedures. Features of SRS are summarized as follows:

- SeNB could listen for UE's SRS while UE is transmitting data with other eNBs (e.g., macrocell).
- SRS is configurable to various cases: traffic offloading, high data request, and so on.
- SRS is adequate for uplink synchronization. The SeNB cannot demodulate other signals, such as PUCCH, before synchronization with the UE.

Some prerequisites are needed to realize these procedures. First, the interface of Xn between SeNB and MeNB should be able to transfer terminal information and control signals. Second, a lower frequency listening module is required for the procedure of listening to SRS sent from UE to MeNB. Finally, an extra RRC signal is needed for transmitting system information of SeNB from MeNB to UE, and the SeNB is chosen by the MeNB based on the location information of the terminal.

Compared to the current LTE, in this procedure, only one-time initial access is performed to synchronize with both macro and small eNBs. In addition, SeNB needs no PRACH and other public broadcast channels.

Handover Procedure

The handover procedures in current LTE systems are not suitable for a DoC system since SeNBs no longer broadcast PSS/SSS/CRS or other downlink reference signals. The handover procedure of a DoC system are shown in Fig. 2b.

If the terminal is already connected to an SeNB in step 1, a macrocell will request a measurement report from the small cell, while setting aside the report from the terminal. Once the handover condition is fulfilled, the MeNB will trigger pre-handover procedures.

When the signal at the serving macro or small cell becomes lower than a threshold, the MeNB will trigger the SeNB adjacent to the UE to listen for uplink SRS. If the target cell in step 5 is the MeNB, the macrocell will inform UE to communicate with itself via RRC signals. If another small cell is chosen, access to and synchronization with that small cell will be triggered.

With the location information or measurement reports, the MeNB can send the information about terminal mobility to other macrocells. The handover procedure will be triggered, and small cells in another macrocell near the terminal will be notified to listen, and send the measurement reports back to the serving macro for decision. Afterward, UE needs to get initial access to the target macro eNB before handing over to the cells within its coverage.

Paging Procedure

The paging procedure in a DoC system is done by the MeNB, using the procedure in current LTE, and an SeNB will not send paging signals no matter to which kind of base station the UE is connected. When the paging procedure is finished, the small cell discovery procedure is performed to find out the best station for the UE under this circumstance.

Small Cell On-Off Procedure

The main characteristic of the small cell sleeping procedure in a DoC system is that the state of the small cell can be determined by the macrocell, as shown in Fig. 2c, which is different from the current LTE system. Since all of the small cells send the SRS listening report to the macrocell, the optimized centralized decision can be achieved easily compared to the current LTE system, which takes into account the long-term traffic distribution in the cell, the state and capacity of small cells, and the moving speed of terminals. Thus, the DoC system will be more highly efficient and optimal than the current LTE one, in which the small cells are individuals with standalone and a suboptimal distributed on-off strategy. In addition, the small cells have a high probability of staying in de-activated mode for much longer periods than in an LTE network because of the removal of periodic signals and channels, especially under low-traffic scenarios. Thus, the energy consumption of a DoC can be further reduced due to switching off the power amplifier in a long-term sleeping mode.

Performance Evaluation

The performance of the proposed data-only carrier and comparisons with the current LTE system is evaluated through OPNET-based system-level and Matlab-based link-level simulation. In the simulation platform, 19 macrocells (3 sectors/cell) with wrap-around and clusters of small cells are deployed according to [4]. Clusters are uniformly deployed within a macrocell. In each cluster, there are four small cells uniformly distributed across the cluster area. The inter-site distance (ISD) is set as 500 m, and the radius of clusters is 50 m. In the system, an MeNB is deployed in 2 GHz and SeNB is configured in 3.5 GHz with 10 MHz system bandwidth for each eNB. Two thirds of the UEs are uniformly distributed within the clusters, and one third of the UEs are uniformly-distributed across the macrocell area. 80 percent of the terminals are indoor with penetration loss (20 dB for MUE and 23 dB for SUE). The BS power consumption model is based on the EARTH Project. The BS maximum transmission power is

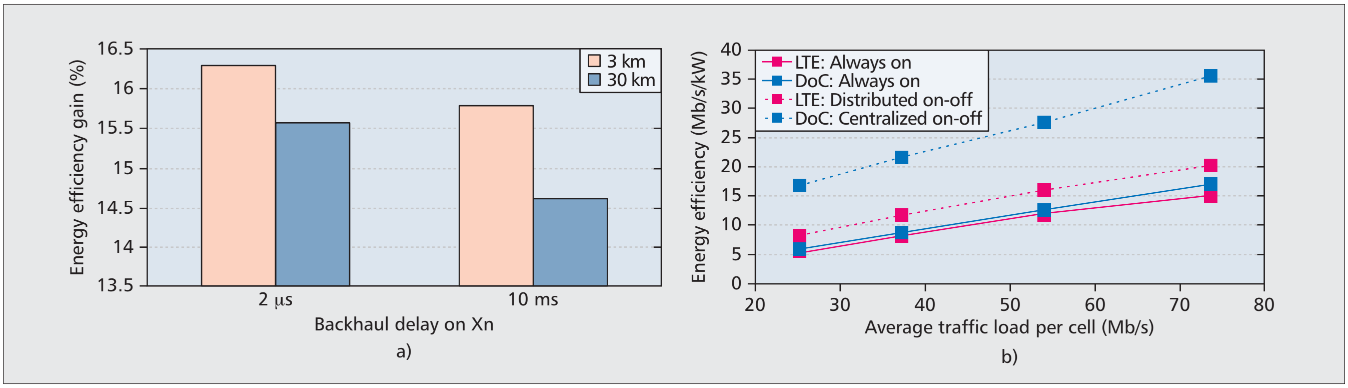


Figure 3. Energy efficiency comparison between the DoC and legacy LTE systems; a) energy efficiency gain under different UE speed and backhaul delay ; b) energy efficiency comparison with on-off strategy.

46 dBm for MeNB and 30 dBm for SeNB, respectively. The wireless channel models are based on 3GPP TR 36.814.

Energy Efficiency Comparison

According to [11], the delay of backhaul between MeNB and SeNB on Xn interface is set with the typical value of 2 μs and 10 ms for ideal and non-ideal backhaul, respectively. In Fig. 3a, the DoC system has an improvement of 14–16.5 percent energy efficiency over the current LTE systems. The benefits come from DoC design, interference cancelation, and faster small cell discovery and handover procedures. The pure-data small cell carrier contributes to the interference cancellation between small cells in a DoC system, so the terminals can hand over between cells quickly, and the modulation and coding scheme (MCS) selection is more timely and accurate than that in an LTE system. For the C-RAN architecture, ideal backhaul between SeNB and MeNB can be achieved easily, which makes the proposed system more practical and useful. In addition, as shown in Fig. 3a, the energy efficiency gain decreases with higher UE speed due to higher handover failure rate and the delay for channel estimation.

In Fig. 3b, the on-off strategies can also help the DoC network save more energy than LTE. As the LTE network with an energy saving mechanism is set as the baseline, nearly 75 percent of energy efficiency gain can be achieved under a DoC network with centralized on-off strategy. This is because the small cells are all standalone to maintain the service in LTE systems, which are hard to put to sleep for a long time. On the contrary, handling all the information of small cells, MeNB can perform a centralized sleep approach easily in the DoC network. Most of the small cells can be turned to deep sleep mode under the low-traffic scenario.

Throughput Performance

The simulation performance of system throughput in heterogeneous network deployment is shown in Fig. 4. It is shown that the DoC system has approximately 17 percent more throughput than the current LTE system, thanks to the reduction of signaling overhead and interference cancellation from co-channel public control signals. In this way, it could be inferred that this benefit could bring an upward trend with increased density of small cells, because a DoC system can help mitigate the interference between small cells in a UDN network.

Signaling Overhead

The typical value for signaling overhead cost of a current LTE system is roughly 25 percent for an eNB [9]. With the removal of cell-level channels and signals in a DoC system, the public overhead cost in SeNB could be reduced to approximately 0 percent. However, the overhead of MeNB will increase to 28.54 percent due to MeNB's assistance to SeNBs. In all, the DoC system will reduce system overhead considerably.

However, as a consequence, the interaction on Xn interface of the backhaul will be further burdened. The backhaul networks should also be optimized in the future as concluded in [12]. Under the typical scenario of ultra-dense network defined in [4], the measurement reports via backhual contribute to 40 * 60 = 2400 measurement messages every 5 ms, occupying the capacity of 11.52 Mb/s. Under a better situation, if all the small cells are only configured to listen to UEs within their own cluster, the measurement reports will occupy only 2.88 Mb/s on the backhaul. Thus, the capacity of backhaul is not the crucial problem for DoC network.

In summary, considerating the advantages of both system design and the energy-saving mechanism illustrated in Figs. 3 and Fig. 4, we can get a total of more than 92 percent energy efficiency gain in a DoC system compared to an LTE network applied with on-off strategy. As small cell deployment gets much denser, further advantages of a DoC system can be obtained.

Conclusions

In this article, a macro-assisted data-only carrier system is designed for small cell enhancement. With the collaboration between MeNB and SeNB, control channel overhead and cell-specific reference signals can be minimized to achieve a "pure" data carrier for small cells. Under this architecture, key procedures, including small cell identification and access, synchronization, handover, and so on, are illustrated. In this way, the proposed data-only carrier system are much "greener" and "softer," and help the wireless

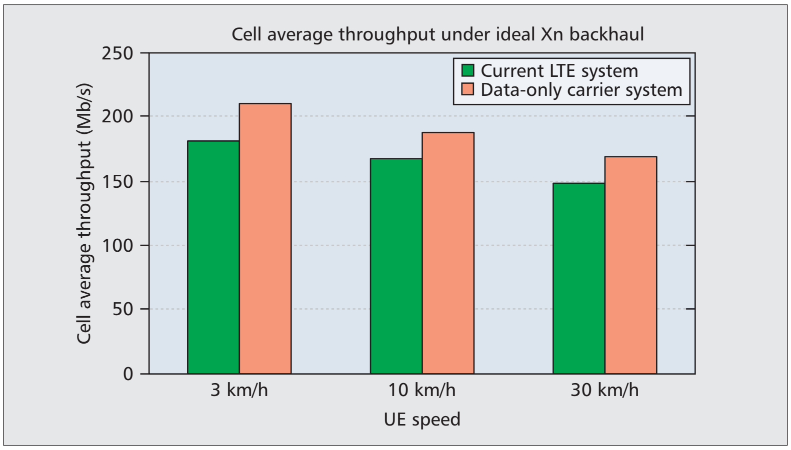


Figure 4. Throughput comparison between a current LTE system and a data-only carrier system.

network move toward a UE-centric "no more cell" architecture. Furthermore, the simulation results show that our proposed systems, achieving significant performance enhancement of signaling overhead, can lead to more than 17 percent throughput improvement and 90 percent energy efficiency gain over the current LTE HetNet with on-off strategy implemented. The proposed scheme can be recommended as a highly potential solution for future 5G mobile networks.

References

[1] C.-L. I et al., "Toward Green and Soft: A 5G Perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 66–73.
[2] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018.
[3] X. Ge et al., "Energy Efficiency Optimization for MIMO-OFDM Mobile Multimedia Communication Systems with QoS Constraints," *IEEE Trans. Vehic. Tech.*, vol. 63, no. 5, June 2014, pp. 2127–38.
[4] 3GPP TR 36.872 V12.1.0, Tech. Spec. Group Radio Access Network, "Small Cell Enhancements for E-UTRA and E-UTRAN — Physical Layer Aspects," Rel. 12.
[5] 3GPP TS 36.211 v11.2.0, Tech. Spec. Group Radio Access Network, "Evolved Universal Terrerstial Radio Access (E-UTRA); Further Advancements for E-UTRA Physical Layer Aspects," Rel. 9.
[6] T. Nakamura et al., "Trends in Small Cell Enhancements in LTE Advanced," *IEEE Commun. Mag.*, vol. 51, no. 2, Feb. 2013, pp. 98–105.
[7] C. Wang et al., " Cellular Architecture and Key Technologies for 5G Wireless Communication Networks," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 122–30.
[8] Z. Niu et al., "Cell Zooming for Cost-Efficient Green Cellular Networks," *IEEE Commun. Mag.*, vol. 48, no. 11, Nov. 2010, pp. 74–79.
[9] X. Xu *et al.*, "On Functionality Separation for Green Mobile Networks: Concept Study over LTE," *IEEE Commun. Mag.*, vol. 51, no. 5, May 2013, pp. 82–90.
[10] A. Mohamed *et al.*, "Physical Layer Frame in Signaling-Data Separation Architecture: Overhead and Performance Evaluation," *Proc. 20th Euro. Wireless Conf.*, May 2014, pp. 1–6.
[11] 3GPP TR 36.932 V12.1.0, Tech. Spec. Group Radio Access Network, "Scenarios and Requirements for Small Cell Enhancements for E-UTRA and E-UTRAN," Rel. 12.
[12] X. Ge *et al.*, "5G Wireless Backhaul Networks: Challenges and Research Advances," *IEEE Network*, vol. 28, no. 6, Nov. 2014, pp. 6–11.

Biographies

XING ZHANG [M'10, SM'14] is now a full professor with the School of Information and Communications Engineering, Beijing University of Posts and Telecommunications (BUPT), 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 60 papers in top journals and international conferences, and has filed more than 30 patents. He has served on the Editorial Boards of several international journals, including *KSII Transactions on Internet and Information Systems* and the *International Journal of Distributed Sensor Networks*, and as a TPC Co-Chair/TPC member for a number of major international conferences, including MobiQuitous 2012, IEEE ICC/GLOBECOM/WCNC, CROWNCOM, Chinacom, and others. He received the Best Paper Award at Chinacom 2014 and the 17th International Symposium on Wireless Personal Multimedia Communications, Sydney, Australia.

JIAXIN ZHANG (zhangjxbupt@gmail.com) received his B.E. degree in information engineering in 2012 from BUPT, and is currently pursuing his Ph.D. degree in BUPT, jointly with the University of Surrey, United Kingdom. His research interests lie mainly in heterogeneous hierarchical wireless networks, green communication, and 5G network architecture.

WENBO WANG received BS., M.S., and Ph.D. degrees from BUPT in 1986, 1989, and 1992, respectively. He is currently a professor and executive vice dean of the Graduate School, BUPT. He is also deputy director of the Key Laboratory of Universal Wireless Communication, Ministry of Education. He has published more than 200 journal and international conference papers, and six books. His current research interests include radio transmission technology, wireless network theory, and software radio technology.

YAN ZHANG received a Ph.D. degree from Nanyang Technological University, Singapore. In August 2006, he joined Simula Research Laboratory, Norway. He is currently head of the Department of Networks at Simula Research Laboratory and an adjunct associate professor with the Department of Informatics, University of Oslo, Norway. He has been a Regional Editor, Associate Editor, on the Editorial Board, or Guest Editor of a number of international journals. His recent research interests include wireless networks, cyber physical systems, and smart grid communications.

CHIH-LIN I (icl@chinamobile.com) received her Ph.D. degree in electrical engineering from Stanford University and has almost 30 years of experience in wireless communications. She has worked at various world-class companies and research institutes, including the Wireless Communication Fundamental Research Department of AT&T Bell Labs; the headquarters of AT&T, as director of Wireless Communications Infrastructure and Access Technology; ITRI of Taiwan, as director of Wireless Communication Technology; Hong Kong ASTRI, as vice president and founding group director of the Communications Technology Domain. She received the *IEEE Transactions on Communications* Stephen Rice Best Paper Award and is a winner of the CCCP National1000 Talent program. Currently, she is chief scientist of wireless technologies of China Mobile, in charge of advanced wireless communication R&D efforts of the China Mobile Research Institute (CMRI). She established the Green Communications Research Center of China Mobile, spearheading major initiatives including 5G key technologies R&D; high energy efficiency system architecture, technologies, and devices; green energy; and C-RAN and soft base stations. She was an elected Board Member of IEEE ComSoc, Chair of the ComSoc Meetings and Conferences Board, and Founding Chair of the IEEE WCNC Steering Committee. She is currently the Chair of FuTURE Forum 5G SIG, an Executive Board Member of GreenTouch, a Network Operator Council Member of ETSI NFV, and an adjunct professor of BUPT. Her research interests are green communications, C-RAN, network convergence, bandwidth refarming, EE-SE co-design, massive MIMO, and active antenna arrays.

ZHENGANG PAN (panzhengang@chinamobile.com) received his B.S.E. from Southeast University, Nanjing, and his Ph.D. degree in electrical and electronic engineering from Hong Kong University. After graduation, he worked for NTT DoCoMo Beijing Communication Labs and ASTRI in Hong Kong working on wireless communication (WiFi, WiMax, LTE), mobile digital TV (T-DMB, DVB-T/H, CMMB), and wireline broadband access (HomePlug, MoCA) for both system/algorithm design and terminal SoC chip implementation. He is currently a principal staff member of the Green Communication Research Center at the China Mobile Research Institute and is now leading a team working on the key technologies for the next generation 5G wireless communication systems. He has published more than 40 papers in top journals and international conferences, and filed 45 patents (20 granted). His research interests are time/frequency/sampling synchronization technology for OFDM/A-based systems, channel estimation, forward error correction coding, MIMO, space-time processing/coding, and cross-layer optimization.

GANG LI (ligangyf@chinamobile.com) received his B.A. degree in telecommunication engineering and M.E. degree in automation engineering from Sichuan University. After graduation, he worked for Lucent Technologies for four years as a team leader and software developer for the core network. He is now a senior researcher at the Green Communication Research Center of the China Mobile Research Institute and is working on key technologies for next generation 5G wireless communication systems. His research interests include radio access network architecture optimization, service-aware signaling/control redesign, and radio and core network convergence.

YAMI CHEN (chenyami@chinamobile.com) received her B.A. degree in telecommunication engineering from Wuhan University of Technology in 2009 and her M.E. degree in telecommunication and information system from BUPT in 2012. Since graduation, she is a researcher at the Green Communication Research Center of the China Mobile Research Institute and has been working on key technologies for next generation 5G wireless communication systems. Together with her colleagues, she has filed more than 30 patents. Her research interests include radio access network architecture design, user-centric signaling/control design, and end-to-end mobile network optimization.

Is Green Networking Beneficial in Terms of Device Lifetime?

Luca Chiaraviglio, Pawel Wiatr, Paolo Monti, Jiajia Chen, Josip Lorincz, Filip Idzikowski, Marco Listanti, and Lena Wosinska

Abstract

This article analyzes the impact that sleep mode (SM)-based green strategies have on the reliability performance of optical and cellular network elements. First, we consider a device in isolation (i.e., not plugged into a network in operation), showing how operational temperature and temperature variations, both introduced by SM, impact its lifetime. We then evaluate, from an operational cost perspective, the impact of these lifetime variations, showing that some devices are critical, that is, their achievable energy savings might not cover the potential additional reparation costs resulting from being put in SM too frequently. Moreover, we present a model for evaluating the impact of SM on the lifetime of a device plugged into an operational network. The analysis considers two case studies (one based on the optical backbone and one on cellular networks) showing that the lifetime of a device is influenced by both the hardware parameters, which depend on the specific design of the device, and the SM parameters, which instead depend on the energy-efficient algorithm used, the network topology, and the traffic variations over time. Our results show that (i) the changes in the operational temperature and the frequency of their variation are two crucial aspects to consider while designing a SM-based green strategy, and (ii) the impact of a certain SM-based strategy on the lifetime of network devices is not homogeneous (i.e., it can vary through the network).

Luca Chiaraviglio and Marco Listanti are with the University of Rome Sapienza.

Pawel Wiatr, Paolo Monti, Jiajia Chen, and Lena Wosinska are with KTH Royal Institute of Technology.

Josip Lorincz is with the University of Split.

Filip Idzikowski is with Poznan University of Technology.

The work of Filip Idzikowski was mainly performed while he was with the University of Rome Sapienza. A preliminary version of this article was published in the *Proceedings of the Fourth IEEE Online Conference on Green Communications* [1].

Introduction

In the past few years, the energy efficiency of communication networks has been the focus of extensive research work. Many green approaches have been proposed in the literature in order to reduce the energy consumption of both cellular and fixed networks, at all layers, and in all network segments (i.e., access, metro/aggregation, and core) [2, 3].

One of the most promising approaches to save energy is to put idle network devices in *sleep mode* (SM), a state in which a network element consumes less energy compared to fully operational mode. In the case of backbone optical networks, this means putting transponders, regenerators, reconfigurable optical add/drop multiplexers (ROADMs), and Erbium doped fiber amplifiers (EDFAs) into SM. In cellular networks SM can be used, for example, with base stations (BSs) and remote radio units (RRUs). However, the adoption of SM approaches can also trigger effects other than energy saving. In particular, SM cycles may vary the operating temperature of a device and, in turn, affect its lifetime [4].

More profoundly, temperature can impact the reliability performance of a device in different ways. For example, the lifetime of a device can be extended when its operating temperature is reduced [1]. Thus, SM-based green techniques could be beneficial. On the other hand, frequent temperature variations may accelerate the occurrence of failures [5] and, in turn, shorten the lifetime of devices. These aspects need to be carefully assessed, because any change in a device's reliability performance impacts the network operational cost in terms of extra failure reparation expenditures [4]. Therefore, it is important to make sure that the extra reparations caused by the decreased lifetime of some network devices does not exceed the potential savings from using an energy-efficient strategy.

This article presents a comprehensive study assessing the impact that SM-based green strategies have on the lifetime of network devices. The study considers the main network elements used in cellular and backbone optical networks. The objective of the study is twofold: identify the devices that may experience the highest impact on the operational cost increase due to a possible reliability performance degradation, and assess the reliability performance degradation measured in terms of *acceleration factor* (AF) of the network elements that are set to SM during network operation. This latter aspect is vital to assess the vulnerability of the network as a whole (i.e., defined in terms of how frequently network elements are likely to experience a failure).

Our results indicate that the AF of a device is a function of its hardware (HW) characteristics,

 0163-6804/15/$25.00 © 2015 IEEE

and the specific energy-efficient strategy, which inevitably sets the values of the device SM periods and frequencies. The results of the study also confirm that in order to benefit from an energy-efficient strategy, the SM switching frequency has to be kept to a minimum. Finally, when comparing cellular with optical devices, the latter seem to be more susceptible to reliability performance degradation from the operational cost perspective.

Lifetime Variations: Physical Phenomena

When a device goes into SM, its operating temperature may be reduced because a number of its internal components are switched to an off or stand-by mode. There are several models in the literature that can be used to characterize how much temperature variations impact the lifetime of a device. One of them is the Arrhenius law [6], which determines how much the failure rate increases/decreases if a device is operated at a temperature other than a reference one. According to the Arrhenius law, if the operating temperature of a device is reduced, its failure rate becomes smaller as well. This means that by considering only the effects of the Arrhenius law, an energy-efficient scheme would have a positive impact on the lifetime of a device, since the operating temperature of a device in SM is typically lower than one at full power.

However, there are also other physical phenomena that need to be considered, and may negatively impact the lifetime of a device. It is well known that temperature changes may affect the expansion of different materials within the same device differently if they have different coefficients of temperature expansion. In turn, a device may suffer strain and fatigue when temperature conditions change, in particular when this happens in a cyclic way. This phenomenon can be observed for many electronic devices, especially for solder junctions. The Coffin-Manson model [7, 8] describes the effects of material fatigue caused by cyclic thermal stress and is used to predict the number of temperature cycles a component can endure before failing. In particular, the more often a device experiences temperature variation, the shorter its lifetime might become. This effect occurs when the device passes from full power to SM and vice versa. In the remainder of this article, the lifetime variations of network devices are modeled using the Arrhenius law and the Coffin-Manson model.

Energy Efficiency vs. Lifetime Variations: An Operational Cost Perspective

This section presents an assessment, from the operational cost perspective, of the maximum tolerable lifetime decrease when setting the main active devices used in wavelength-division multiplexing (WDM) optical backbone networks and cellular networks to SM.

A lifetime degradation introduces an additional operational expenditure (OPEX) in terms of failure reparation cost. When green strategies are used in a network, this has to be taken into account in the overall OPEX calculation. It is then important to understand up to which point the savings coming from a green strategy can still compensate the extra costs related to the decreased lifetime of a device. This trade-off can be measured using the maximum allowable lifetime decrease (in percentage compared to the nominal conditions, i.e., when SM is not applied) so that the reparation costs will not exceed the saving obtained by lowering the energy consumption by a given threshold (i.e., 10 percent). The maximum allowable lifetime decrease can be expressed in the following way [5]:

$$MaxLD_{10\%} = \frac{\overbrace{10\% \cdot P_{eq} \cdot C_{kWh}}^{\text{monetary energy saving}}}{\underbrace{10\% \cdot P_{eq} \cdot C_{kWh}}_{\text{monetary energy saving}} + \underbrace{\frac{FR}{10^6} \cdot (MTTR \cdot Pers. \cdot C_m + C_{eq}}_{\text{reparation costs}}}, \quad (1)$$

where P_{eq}[W] represents the power consumption in active mode of the device under exam, C_{kWh}[US$/kWh] is the electricity cost, FR represents the device failure rate expressed in failure in time (FIT) [units] (i.e., the failure in time unit which corresponds to one failure per 10^9 hours of operation), $MTTR$ [h] is the mean time to repair the device, $Pers.$ [member] represents the number of reparation crew members necessary to repair the failed device, C_m [US$/h/member/failure] is the hourly rate of a reparation crew member, and C_{eq} [US$/failure] is the cost to buy a replacement unit of the device under reparation. For example, assuming that the monetary energy saving is US$400 and reparation costs are US$600, according to Eq. 1, the maximum allowable lifetime decrease will be 40 percent.

Table 1 presents the values of the maximum allowable lifetime decrease for a number of backbone WDM optical and cellular network devices, including linecards, transponders, ROADMs, EDFAs, BSs, and RRUs.[1] In particular, we consider two different scenarios: in the first one the device that fails has to be replaced, while in the second the device can be repaired without involving substitution. To compute the lifetime decrease, we adopt the parameters reported in the table, with the exception of the cost of energy C_{kWh} and the hourly rate of crew members C_m that are equal to US$0.16/kWh [12] and $US190/h/member/failure [4], respectively.

The lifetime decrease results presented in Table 1 show that in the scenario without replacement, the majority of network equipment (with the exception of EDFAs) is able to sustain a significant degradation of lifetime without a considerable impact on operational costs; that is, the maximum allowable lifetime decrease is greater than ~10 percent. In this case, the achievable energy savings from SM can easily compensate for the (possible) extra reparation costs. On the other hand, a different conclusion can be drawn in the case of EDFAs, where the

Our results indicate that the lifetime of a device is a function of its hardware characteristics, and of the specific energy efficient strategy that inevitably sets the values of the device sleep mode periods and frequencies.

[1] The input data come from [5, 9–11], and from the following sources: Cisco ONS 15454 Multiservice Transport Platform — MSTP (http://www.cisco.com/c/en/us/products/optical-networking/ons-15454-multiservice-transport-platform-mstp/index.html), CISCO Price List (http://www.kernelsoftware.com/products/catalog/cisco.html), EDFA-R Erbium Doped Fiber Amplifier with Redundant Power Supplies (http://www.pbnglobal.com/en/products/edfa-r), Huawei GBSS9.0 DBS3900 Product Description V2.1 (http://cosconor.fr/GSM/Divers/Equipment/Huawei/DBS3900 percent20product percent20description.pdf), Huawei eLTE3.1 DBS3900 LTE FDD Product Description (http://enterprise.huawei.com/ilink/cnenterprise/download/HW275863).

	Device type	*FR* (FIT)	*MTTR* (h)	*Pers.*	C_{eq} (kUS$)	P_{eq} (W)	Maximum allowable lifetime decrease (with replacement)	Maximum allowable lifetime decrease (without replacement
Optical	Multirate DWDM XPonder card	2900	2	1	10	50	2.59%	42.10%
	10 Gb/s full-band tunable multirate transponder card	4200	2	1	25	35	0.52%	26.00%
	40-channel single-module ROADM	3300	2	1	25	35	0.66%	30.90%
	100 Gb/s service line card	8600	2	1	190	130	0.18%	38.90%
	16-port wavelength mux/demux flex spectrum line card	6200	2	1	100	100	0.26%	40.40%
	Enhanced 96-channel EDFA	7100	6	2	15	40	0.52%	3.80%
	EDFA-R	10,000	6	2	10	18	0.23%	1.25%
Cellular	LTE RRU (micro/1 cell)	10,000	1.5	1	0.65	50	7.88%	21.92%
	LTE micro BS	6452	2	1	3.9	100	5.48%	39.49%
	LTE RRU (macro, 1 sector/cell)	10,000	4	2	2.6	120	4.45%	11.21%
	Main unit macro BS (3-sector LTE, 2 transceivers/sector)	6452	3	1	15.6	460	6.59%	66.68%
	Macro BS (3-sector UMTS, 2 transceivers/sector)	10,000	5	2	32.5	1700	7.33%	58.87%
	Macro BS (3-sector GSM, 2 transceivers/sector)	20,000	6	2	45.5	840	2.74%	37.09%

Table 1. Maximum allowable lifetime decrease analysis.

values of maximum allowable lifetime decrease are very small.

On the contrary, when replacement costs are taken into account, all network devices under exam (both optical and cellular) have a maximum allowable lifetime decrease below 10 percent, suggesting that the extra reparation cost plays a very crucial role from the operational cost perspective. It can also be noticed that the backbone optical devices show the worst lifetime decrease performance. Therefore, optical devices are more critical than their cellular counterpart and need to be considered more carefully.

Acceleration Factor: From the Single Device to the Network Perspective

The previous section highlighted how SM may impact the operational cost as a consequence of the potential variations of the component lifetime. The study considered a device in isolation. This section goes one step further by looking at the reliability performance of a device plugged into a network in operation when energy-efficient approaches are applied. The intuition is that when a device is plugged into a network, some of the parameters affecting the device lifetime (i.e., SM duration, frequency of sleep cycles) cannot be known in advance because they are dependent on the specific green strategy and the network status (e.g., connectivity, congestion, traffic conditions). In order to understand how green network operations impact the lifetime of all devices in a network, it is necessary to first model the lifetime variations, as presented next.

The considerations made previously highlighted a clear trade-off between SM duration (i.e., a positive effect modeled by the Arrhenius law [6]) and its frequency (i.e., a negative effect modeled by Coffin-Manson [7, 8]). More formally, the overall failure rate of an arbitrary device i in the network can be expressed in the following way:

$$\gamma_i = \underbrace{\left[(1-\tau_i^{sleep})\gamma_i^{on} + \tau_i^{sleep}\gamma_i^{sleep}\right]}_{\text{Impact of SM Duration}} + \underbrace{\frac{f_i^{tr}}{N_i^F}}_{\text{Impact of SM Frequency}}[1/\text{h}], \quad (2)$$

where t_i^{sleep} is the amount of time spent by the device in SM (normalized to the time the device is under observation), γ_i^{on} is the failure rate at full power, γ_i^{sleep} is the failure rate when the device is in SM, f_i^{tr} is the frequency of the SM cycles, and N_i^F is the number of cycles supported by the devices before a failure occurs. The first term in the equation is derived using [6]. It is the sum of the failure rates at full power and in SM, respectively, weighted by the amount of time the device is in SM. The second term is taken from [7, 8]. It represents the contribution to the failure rate that is a function of how frequently the device's operational state changes. The two terms are then added to compute the overall failure rate of the device, assuming that they are statistically independent and their effects are additive [1].

In order to model the lifetime variations of a device, the concept of *acceleration factor* (*AF*) is introduced. *AF* is a parameter measuring the mean lifetime decrease/increase w.r.t. the device operating in full power conditions (i.e., when SM is not applied). In particular, an *AF* larger than one means that SM has a negative impact on the device lifetime, while if *AF* is lower than one the device lifetime benefits from the introduction of SM. More formally, the *AF* of device i can be defined as

$$\begin{aligned} AF_i &= \frac{\gamma_i}{\gamma_i^{on}} \\ &= 1 - \underbrace{(1 - AF_i^{sleep})\tau_i^{sleep}}_{\text{Lifetime Increase}} + \underbrace{\chi_i f_i^{tr}}_{\text{Lifetime Decrease}}, \end{aligned} \tag{3}$$

where AF_i^{sleep} is the *AF* experienced by the device when it is always kept in SM (i.e., $AF_i^{sleep} = (\gamma_i^{sleep})/(\gamma_i^{on})$), which according to [6] is always lower than 1. Intuitively, AF_i^{sleep} is the minimum *AF* achievable by the device when SM is applied and the impact of temperature variation is not considered. This parameter depends on the technology adopted to implement SM of the device: the lower AF_i^{sleep}, the higher the gain in terms of reliability performance. χ_i is an HW parameter, defined as $\chi_i = 1/(\gamma_i^{on} N_i^F)$ [h/cycle], which acts as a weight for the frequency of the SM cycles.

The acceleration factor AF_i comprises two terms: $(1 - AF_i^{sleep})\ \tau_i^{sleep}$ tends to decrease the value of AF_i due to the temperature decrease during SM, while $\chi_i f_i^{tr}$ tends to increase the value of AF_i due to the temperature changes caused by the SM cycles. Moreover, AF_i is influenced by two types of parameters: technological (i.e., AF_i^{sleep} and χ_i), which are strictly related to the HW used to build the device, and SM-related (i.e., τ_i^{sleep} and f_i^{tr}), which instead depend on the green strategy used and the conditions of the network in operation.

Even though the proposed model is a first-order approximation, it is already useful to draw some interesting conclusions. The design of a device plays a crucial role; that is, devices that are designed to better sustain frequent temperature variations are likely to fail less frequently. In addition, the duration and frequency of SM strongly impact the lifetime. Intuitively, if SM happens frequently, a device experiences frequent power state transitions between full power and SM, and consequently its lifetime will be reduced.

Apart from the effects on a single device, an operator is mostly interested in evaluating the impact of SM on the reliability performance of all the devices deployed in the network. One possibility is to use a metric measuring (per each device type) the average value of *AF*, for example, something similar to the $\overline{AF}$ metric introduced in [1]. However, as the duration and frequency of SM periods are not the same for all devices, it might also be interesting to look into the best and worst case scenarios in terms of *AF* for a set of given devices. The next section presents a detailed case study of these aspects.

If sleep mode happens frequently, a device experiences frequent power state transitions between full power and sleep mode, and consequently its lifetime will be reduced.

Acceleration Factor Evaluation: A Network Level Case Study

This section evaluates the *AF* of a given set of devices when an SM-based green strategy is applied to a network. Two scenarios are considered:

- A backbone optical network using an energy-efficient routing strategy where energy is saved by setting EDFAs to SM
- A cellular scenario where BSs are set to SM when not needed to provide coverage and/or capacity

Details on each scenario are presented next.

Backbone Optical Network Scenario — We consider a green routing strategy called Weighted Power Aware Lightpath Routing (WPA-LR) [13] tested on the COST239 backbone optical network. WPA-LR is able to save energy by encouraging the routing of incoming lightpath requests over already used links, thus maximizing the number of unutilized fiber links, that is, maximizing the number of EDFAs that can be put in SM. A detailed description of the algorithm and the topology characteristics is available in [1].

Connection requests are bidirectional, and their source and destination pairs are uniformly chosen among the network nodes. They arrive according to a Poisson process, while their service time is exponentially distributed with an average holding time equal to six hours. In order to compute the value of *AF* for each EDFA deployed in the network, it is assumed that all of them have the same HW characteristics, while the frequency and duration of each sleep cycle for each EDFA are collected by simulating the WPA-LR algorithm over the COST239 network. Simulation results are averaged over a series of 10 experiments, with 10^5 connection requests in each experiment.

Cellular Scenario — We consider an energy-aware algorithm and a realistic cellular deployment scenario, both obtained from [14]. Due to the lack of space, we refer the reader to [14] for a comprehensive description. In brief, we consider a scenario with ~33 Universal Mobile Telecommunication System (UMTS) macro BSs and a service area (SA) of 9.2 × 9.2 km^2. Inside the SA, we assume more than 3000 user terminals (UTs) requesting voice and data services. Unless otherwise specified, we assume that the maximum data rate for each user is equal to 384

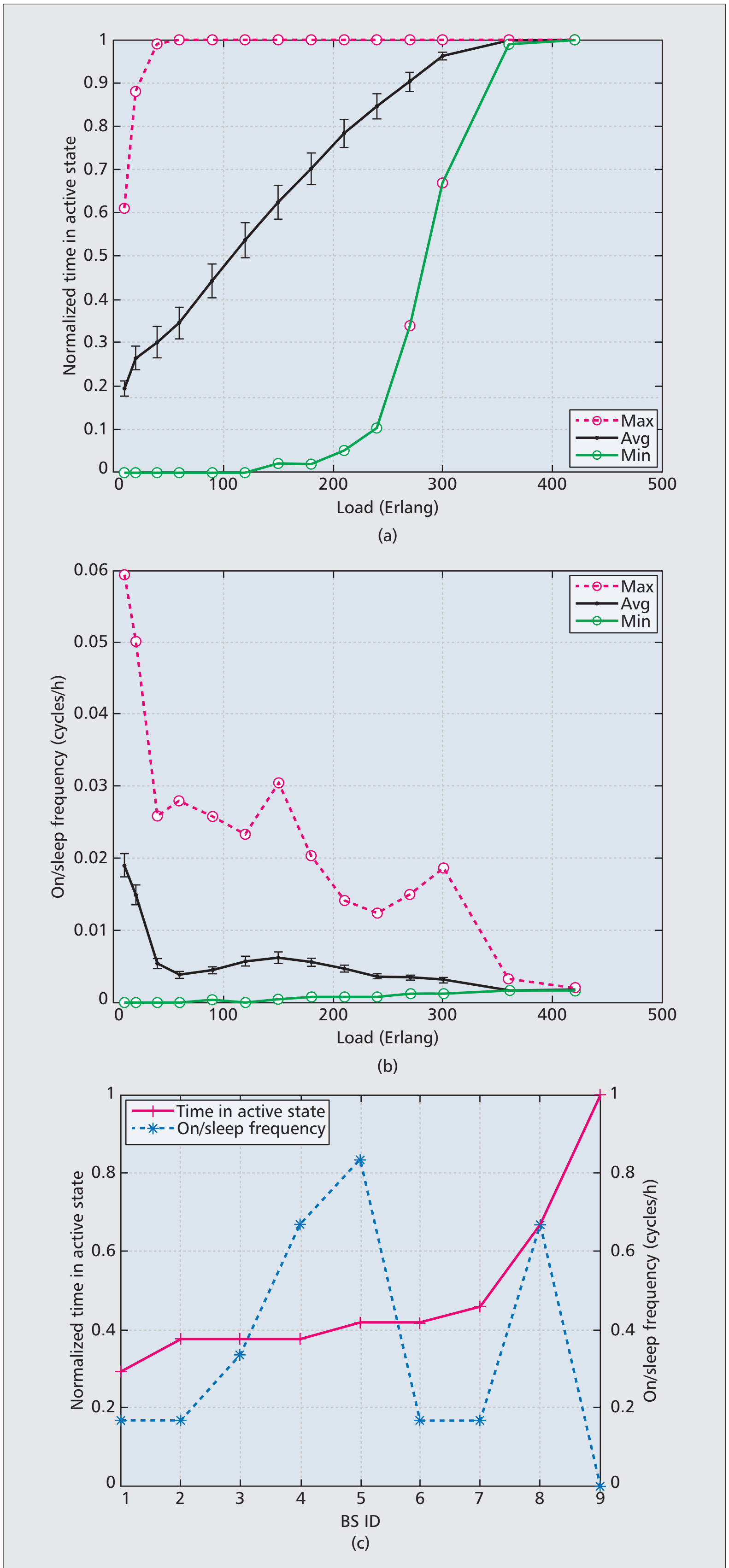


Figure 1. a) Maximum (Max), average (Avg), and minimum (Min) normalized time in active state and active/sleep switching frequency; b) for the optical scenario; c) (normalized) time in active state and active/sleep switching frequency for each base station experiencing SM cycles in the cellular scenarios.

kb/s. Moreover, we assume a day-night traffic variation with a deterministic profile over 24 h. For this scenario we solve the optimization problem aimed at minimizing the number of active BSs while guaranteeing the required coverage and capacity demand for all the UTs that are active in each time period. Similar to the optical case, we assume the same HW characteristics for all BSs, while we collect the frequency and duration of each SM cycle via simulation.

Sleep Mode Duration and Frequency

We first investigate the duration and frequency of the SM cycles, focusing on the optical scenario. In particular, we vary the traffic load between 10 and 420 Erlang. These values are chosen in order to investigate different conditions where the network blocking probability does not exceed 10 percent. Figures 1a and 1b report the normalized time in active state and the on/sleep frequency. In particular, given the duration and on/sleep frequency for each EDFA, we have computed the minimum, maximum, and average duration and the on/sleep frequency values. The bars in the figures indicate the confidence intervals for the average values, assuming 99 percent confidence level. The time in active state (reported in Fig. 1a) tends to increase with the load. This is due to the fact that EDFAs need on average to be powered on for a longer time in order to sustain the load increase. For load > 350 Erlang nearly all EDFA are on, since in this case all links are used. Focusing then on the maximum values, we can see that for load > 50 Erlang, there are already some EDFAs that are always powered on. However, focusing on the minimum values, we can see that until medium load (i.e., < 250 Erlang), there are EDFAs that are powered on for less than 10 percent of the time. This means that the duration of SM varies depending on the device, and consequently the effect of SM on the device lifetime is not the same for all the EDFAs.

Figure 1b reports the maximum, average, and minimum EDFAs on/sleep frequency values expressed in cycles per day. Interestingly, the average on/sleep frequency tends to be higher at lower load, due to the fact that WPA-LR only considers power minimization, while the on/sleep frequency is not taken into account. Consequently, solutions reducing power consumption might result in frequent on/sleep cycles. Focusing on the minimum values, we can see that there are EDFAs that seldom change their power state, resulting in frequency almost close to zero (as expected). These EDFAs are either in full power or in SM most of the time. Finally, we can see that the maximum on/sleep frequency is three times higher than the average for load of 10 Erlang. Thus, similar to the duration of SM, the on/sleep frequency also varies considerably among the EDFAs in the network.

Figure 1c reports the normalized time in active state and the on/sleep frequency for the cellular scenario. The figure presents results for each BS with at least one on/sleep cycle per day (we exclude BSs that are always at full power or in SM). Interestingly, here we can also see a variation in the time in active state and on/sleep frequency, suggesting that the impact of SM on the lifetime will not be the same for all BSs.

Acceleration Factor

Since AF is a function of the HW parameters (i.e., χ and AF^{sleep}), we vary them to compute the resulting value of AF for all the EDFAs in the optical scenario (Fig. 2) and for all the BSs in the cellular scenario (Fig. 3). In doing so we assume that all EDFAs or all the BSs, depending on the scenario considered, have the same HW parameters.

Figure 2b shows the average values of AF over all the EDFAs in the network, considering a load of 150 Erlang. The red dashed line highlights the level curve $AF = 1$. The region on the left of this line represents the zone where on average EDFAs in the network fail less often compared to the case in which the energy-aware algorithm is not used(i.e., $AF < 1$). On the contrary, the region on the right is the zone where $AF > 1$, that is, EDFAs on average are expected to fail more often than in the case in which WPA-LR is not used. From the figure, it can be noticed that the HW parameters play a crucial role in determining the impact of an energy-efficient strategy in terms of average lifetime decrease/increase. In particular, AF^{sleep} always has a positive effect on the lifetime; that is, in SM the operating temperature of an EDFA is lower. On the other hand, the frequency weight becomes the discriminating factor, meaning that EDFAs whose χ is very high (in this case higher than 2.7 h/cycle) will experience on average a reduced lifetime.

Besides the average, Fig. 2 reports the worst (a) and best (c) AF values. While computing them we did not consider the EDFAs that are either in SM or in full power all the time since they represent trivial cases. In the worst AF scenario, the curve $AF = 1$ is very close to the y-axis, meaning that SM will always result in a decrease of EDFA lifetime. This is due to the fact that the EDFA under consideration frequently changes its power state, with very short SM intervals. On the contrary, for the best AF scenario, AF is always lower than 1 (for the values of χ considered in the figure). For example, this is the case of an EDFA that is powered off for a long time (e.g., it is powered on only in the case of traffic spikes), resulting in an on/sleep frequency that is very low.

Figure 3b reports the average value of AF vs. the values of the HW parameters, focusing on the cellular scenario. The red dashed line again marks the curve $AF = 1$. Also, here we can see that there is clearly a trade-off between the zone in which using the SM option allows for a lifetime increase (area to the left) and the zone in which SM negatively impacts the lifetime (area to the right). Additionally, we report the worst (a) and best (c) AF scenarios. In the worst AF scenario, the BS AF is lower than 1 for $\chi \leq 2$ (h/cycle). This means that a BS exploiting an energy-efficient radio resource management that also accounts for the operating state transition frequency will be able to increase its lifetime (assuming a frequency weight $\chi \leq 2$ (h/cycle)). In the best scenario (Fig. 3c), AF is always lower than one regardless of χ. Similar to the optical scenario, this case represents those BSs that are normally in SM and are powered on during peak traffic periods.

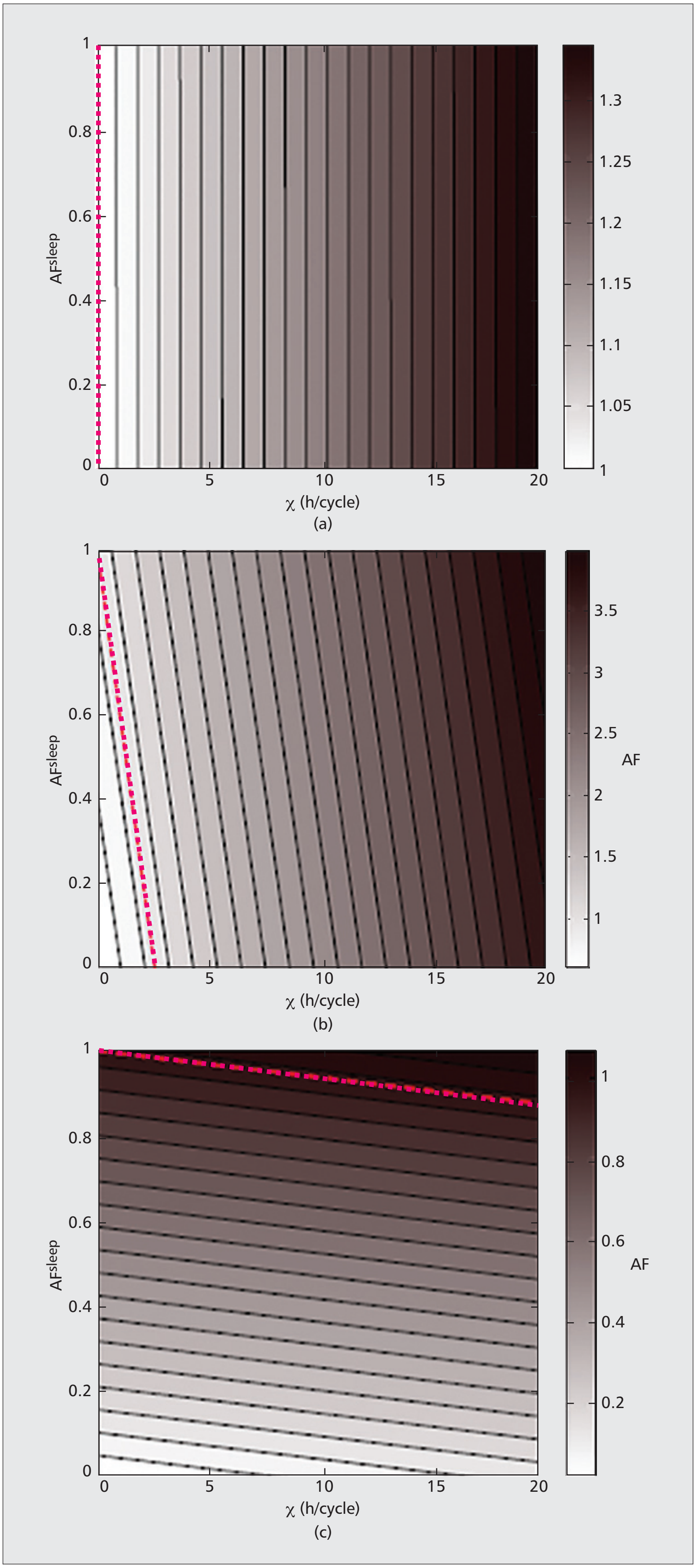


Figure 2. Acceleration factor in the optical scenario: a) worst AF value; b) average AF value; c) best AF value, as a function of χ and AF^{sleep} (load = 150 Erlang).

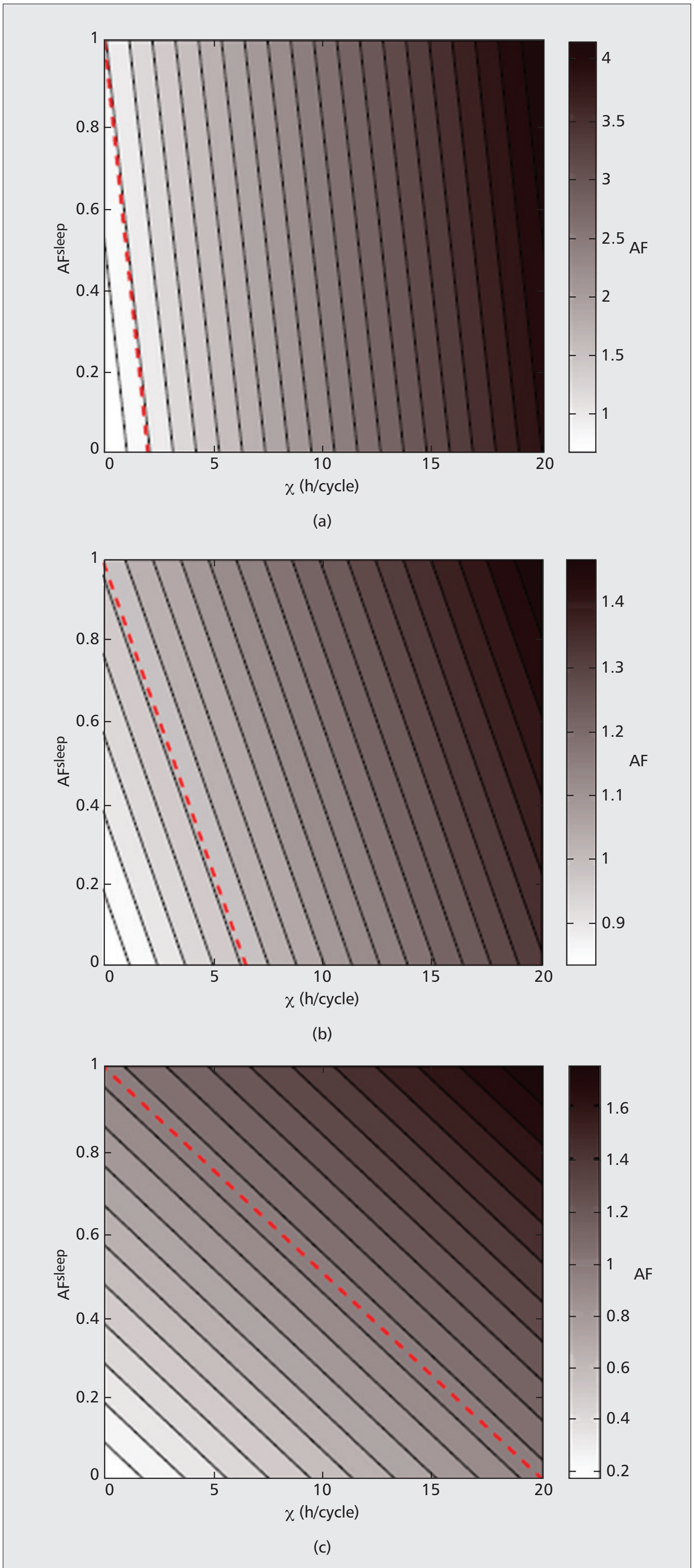


Figure 3. Acceleration factor in the cellular scenario: a) worst AF value; b) average AF value; c) best AF value, as a function of χ and AF^{sleep}.

Acceleration Factor Variations

It is important to consider not only the average, best, and worst AF values, but also to understand how AF varies over the set of devices in the network. For this purpose we introduce the concept of the efficient area, defined as the area in which $AF < 1$. In other words, the efficient area is the triangular region below the level curve $AF = 1$ in the bottom left region of all the plots in Figs. 2 and 3. If a device operates in that area, its lifetime will benefit as a result of the SM-based green strategy applied.

Figure 4 reports the cumulative distribution function (CDF) of the size of the efficient area, considering all the EDFAs and BSs deployed in the network for the optical and cellular scenarios. The efficient area value was computed by varying the values of the HW parameters.[2] It can be seen that the size of the efficient area strongly varies over the set of devices. Moreover, the larger variation in the optical scenario compared to the cellular one is due to the larger number of devices included in the scenario, which increases the chances to put the devices in the network in SM. Thus, the impact of SM on the lifetime of a specific device might be very different, with some devices that may fail more often and others that may increase their lifetime.

We have also investigated how much the load impacts the size of the efficient area. In particular, Fig. 5 reports average, maximum, and minimum sizes of efficient areas for increasing values of load. It can be observed that the size of the average efficient area does not significantly change for values of the load between 40 and 240 Erlang, suggesting that in this region WPA-LR always trades between on/sleep frequency and duration of SM. The size of the efficient area tends to decrease for higher values of load, since in this case the duration of SM is lower (i.e., more frequent SM cycles), and therefore the area in which $AF < 1$ tends to decrease. The figure also shows that there is always a non-negligible variation between the minimum and maximum values of the efficient area for each value of load (up to 5 orders of magnitude).

Finally, Table 2 reports load variations of the size of the efficient area in the cellular scenario. More specifically, the load is changed by varying the data traffic requested by each user considering 64, 128, and 384 kb/s, respectively. Similar to the optical scenario, the average size of the efficient area does not consistently vary with load. Moreover, the difference between maximum and minimum values is always about one order of magnitude for each load value, suggesting that the impact of SM on BS lifetime is not the same for all BSs in the network.

Conclusions and Future Work

We have investigated the impact of SM-based energy-efficient strategies on the lifetime of devices in cellular and backbone optical networks. In particular, we have first studied the problem from an operational cost perspective, showing that the energy savings introduced by SM in some cases can hardly cover the reparation costs when the devices are replaced as a

	64 kb/s	128 kb/s	384 kb/s
Min	0.91	0.78	0.99
Avg	4.37	6.95	5.28
Max	11.49	11.49	9.99

Table 2. Minimum, average, and maximum efficient area sizes (h/cycle) vs. load for the cellular scenario.

consequence of failures. Moreover, we have developed a simple model to evaluate the impact of SM-based green strategies on the device acceleration factor. In particular, we showed that SM lowers the operating temperature, thus potentially increasing the device lifetime, while frequent SM cycles have a negative effect on the lifetime. The model was used in two case studies, showing the dependence of the device lifetime variations on the specific HW and SM parameters. Additionally, we have shown that the AF varies considerably over the various devices in the network.

This work is a first step toward a more comprehensive approach in which device lifetime and energy awareness should be jointly considered. As future work, we plan to extend our analysis to consider the carbon footprint of network devices and the impact that SM might have on the switching elements in data center networks [15]. Moreover, we plan to collect measurements of the lifetime variations of real devices supporting SM. Finally, it would also be interesting to develop green techniques explicitly targeting the maximization of devices' lifetimes as well as a more detailed analysis of which components of each device are more susceptible to failures when SM is applied.

Acknowledgments

The research leading to these results was funded by the Sapienza Awards Project LIFETEL, the Polish National Science Center (decision DEC-2014/12/S/ST7/00415), and by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 318137 (ICT-DISCUS) and grant agreement no. 607078 (EPISECC).

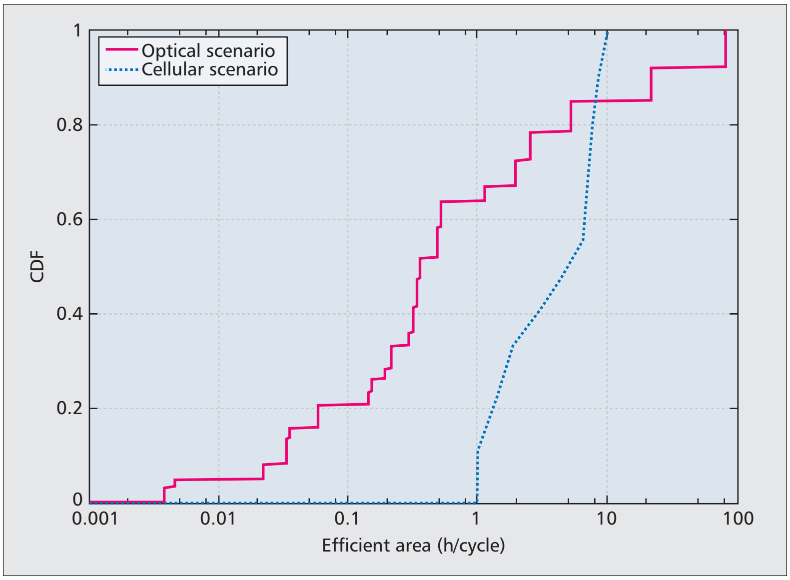


Figure 4. Cumulative distribution function of the size of the efficient area for the optical (load = 150 Erlang) and cellular scenarios.

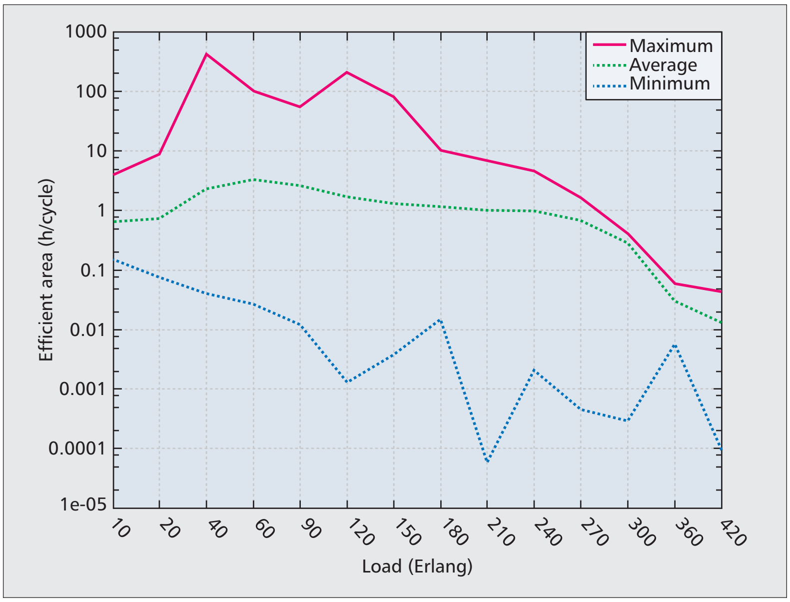


Figure 5. Maximum, average, and minimum values of the efficient area as a function of the load for the optical scenario.

References

[1] L. Chiaraviglio *et al.*, "Impact of Energy-Efficient Techniques on a Device Lifetime," *Proc. IEEE Online GreenCommun.*, online conference, Nov. 2014.

[2] W. Vereecken *et al.*, "Power Consumption in Telecommunication Networks: Overview and Reduction Strategies," *IEEE Commun. Mag.*, vol. 49, June 2011, pp. 62–69.

[3] W. Van Heddeghem *et al.*, "A Quantitative Survey of the Power Saving Potential in IP-Over-WDM Backbone Networks," *IEEE Commun. Surveys and Tutorials*, 2014.

[4] P. Wiatr *et al.*, "Energy Saving in Access Networks: Gain or Loss from the Cost Perspective?" *Proc. IEEE ICTON*, Cartagena, Spain, June 2013.

[5] P. Wiatr *et al.*, "Energy Efficiency Versus Reliability Performance in Optical Backbone Networks," *J. Optical Commun. and Networking*, vol. 7, Mar. 2015, pp. A482–91.

[6] S. Arrhenius, *Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren*, Wilhelm Engelmann, 1889.

[7] L. F. Coffin Jr. and U.S. Atomic Energy Commission and General Electric Company, *A Study of the Effects of Cyclic Thermal Stresses on a Ductile Metal*, Knolls Atomic Power Lab., 1953.

[8] S. S. Manson, "Behavior of Materials under Conditions of Thermal Stress," NACA Report 1170, 1954.

[9] K. Johansson, *Cost Effective Deployment Strategies for Heterogenous Wireless Networks*, Ph.D. thesis, KTH, Communication Systems, CoS, Nov. 2007.

[10] G. Auer *et al.*, "Framework for Energy Efficiency Analysis of Wireless Networks," *Proc. Wireless VITAE*, Chennai, India, Feb. 2011.

[11] J. Lorincz, T. Matijevic, and G. Petrovic, "On Interdependence among Transmit and Consumed Power of Macro Base Station Technologies," *Computer Commun.*, vol. 50, Sept. 2014, pp. 10–28.

[12] EC, "Energy Prices and Cost Reports," http://ec.europa.eu/energy/doc/2030/20140122swdprices.pdf, Mar. 2014.

[13] P. Wiatr, P. Monti, and L. Wosinska, "Power Savings vs. Network Performance in Dynamically Provisioned WDM Networks," *IEEE Commun. Mag.*, vol. 50, May 2012, pp. 48–55.

[2] The ranges of the parameters are sufficiently wide to properly compute all the areas.

[14] J. Lorincz, A. Capone, and D. Begusic, "Impact of Service Rates and Base Station Switching Granularity on Energy Consumption of Cellular Networks," *EURASIP J. Wireless Commun. and Networking*, vol. 2012, no. 342, Nov. 2012, pp. 1–24.
[15] M. Chen *et al.*, "Enabling Technologies for Future Data Center Networking: A Primer," *IEEE Network*, vol. 27, no. 4, 2013, pp. 8–15.

Biographies

LUCA CHIARAVIGLIO is an assistant professor in the Department of Information, Electronics and Telecommunications (DIET) of the University of Rome Sapienza, Italy. He holds a Master's degree in computer engineering and a Ph.D. in telecommunication and electronics engineering, both obtained at Politecnico di Torino, Italy. Between January 2011 and March 2013, he was first a post-doctoral researcher at the Telecommunication Network Group of Politecnico di Torino and then an ERCIM fellow at the Mascotte project of INRIA Sophia Antipolis, France. Moreover, in 2013 he was the Fabio Neri Fellow at the University of Rome Sapienza. Between August 2009 and April 2010, he was a visiting scholar at the WING group of Boston University, Massachusetts, under the supervision of Prof. Ibrahim Matta. He has coauthored more than 55 papers published in international journals and conferences, and he participates in the Program Committees of several conferences, including IEEE GLOBECOM, IEEE ICC, and IEEE Greencomm. He has also partecipated in different European projects (FP7 TREND, FP7 ECONET, FP7 BONE) and national projects (FIRB GreenNet). He is the coordinator of the national project LIFETEL (on increasing lifetime in telecommunication networks). His main research interests are in the field of sustainable networking, focusing on backbone, cellular, and data center networks.

PAWEL WIATR received his M.Sc. and Engineering degrees in telecommunications (2001) from the National Institute of Applied Sciences in Lyon, France (INSA de Lyon). Currently he is working toward his Ph.D. degree at KTH Royal Institute of Technology, Sweden. His research interests include optical networks, energy efficiency in networks, and network management. From 2005 to 2011 he worked at Comarch SA, Poland, as a technical sales manager with responsibility of proposing network management solutions (operation support systems) for network operators. Prior to that, he worked as a network administrator and research engineer at KTH in the domains of wireless networks and ubiquitous computing (interactive workspaces).

PAOLO MONTI [M] is an assistant professor at KTH Royal Institute of Technology. He received a Ph.D. degree (2005) from the University of Texas at Dallas, where he worked as a research associate until 2008. He serves on the Editorial Board of *Springer Photonic Network Communications* and has co-authored more than 100 technical papers, with three best paper awards. His main research interests are within the networking aspects of all-optical networks, while part of his work recently focused on the architectural and control challenges posed by next generation mobile communication (i.e., 5G) on optical transport networks.

JIAJIA CHEN [SM] received a B.S. degree (2004) from Zhejiang University, China, and a Ph.D. degree (2009) from KTH Royal Institute of Technology, Sweden. She is an associate professor in the Optical Networks Lab (ONLab) at KTH. She is a co-author of over 100 publications in international journals and conferences in the area of optical networking. Her main research interests are optical transport and interconnect technology supporting the future 5G and cloud environment. She has been involved in various European research projects like FP7 projects Integrated Project-Optical Access Seamless Evolution (IP-OASE) and Integrated Project-The Distributed Core for Unlimited Bandwidth Supply for All Users and Services (IP-DISCUS), and EIT-ICT projects (e.g., mobile backhaul, M2M, and Xhaul). Moreover, she is principal investigator/co-principal investigator of several national research projects funded by the Swedish Foundation of Strategic Research (SSF) and Swedish Research Council (VR). She has been involved in many professional activities including serving on the TPCs of many conferences (e.g., OFC, Photonics in Switching), as well as being a reviewer for scientific journals and project proposals. Currently, she is serving as an Associate Editor of the *Journal of the Franklin Institute*, Elsevier (impact factor: 2.260).

FILIP IDZIKOWSKI is an assistant professor at Poznan University of Technology, Poland. He received his Ph.D. (Dr.-Ing.) from the Technical University of Berlin, Germany, in 2014, and was with University of Rome Sapienza, for half of 2014. His main research interests include power consumption of multi-layer core networks, traffic modeling, protection, and routing.

JOSIP LORINCZ [M] received his M.S. and Ph.D. degrees in electrical engineering and computer science from the University of Split, FESB, Croatia, in 2002 and 2010, respectively. He was a visiting researcher at the Advanced Network Technologies Laboratory (ANTLab) of the Politecnico di Milano, Italy. He is co-chair of the Symposium on Green Networking and Computing, organized in the framework of the SoftCOM conference. He serves as a TPC member for many international scientific conferences, and a reviewer for scientific journals and project proposals. He was recognized as an outstanding young researcher by the Croatian Academy of Engineering in 2013. His current research interests include energy-efficient wireless and wired networks, optimization in telecommunications, and advanced design of heterogeneous communication networks. He is a member of ACM and the first president of the Croatian ACM Chapter.

MARCO LISTANTI is a full professor in telecommunication networks at the Department of Information of the University of Rome Sapienza. Since November 2013 he has been Dean of the Faculty of Information Engineering, Informatics and Statistics. In about 35 years of activity, his research interest has been mainly focused on the area of architectures and performance evaluation of telecommunication networks. He has participated in national and international standardization commissions and working groups in the field of telecommunication networks (CCITT, ETSI, ITU, IETF, etc.). He has coordinated the activity of many national and international study and research projects in the frame of the programs RACE, ACTS, ITS, COST, ESA, MIUR. He is an author of more than 300 papers, published in the most important scientific magazines and journals, and the proceedings of the main international conferences in the area of telecommunications.

LENA WOSINSKA [SM] received her Ph.D. degree in photonics and her Docent degree in optical networking from KTH Royal Institute of Technology, where she is currently a full professor of telecommunication in the School of Information and Communication Technology (ICT). She is founder and leader of the Optical Networks Lab (ONLab) at KTH/ICT. She has worked on several EU projects and coordinated a number of national and international research projects. Her research interests include fiber access networks, energy-efficient optical networks, photonics in switching, optical network control, reliability and survivability, and optical data center networks. She has been involved in many professional activities, including serving as an Associate Editor and Guest Editor of IEEE, OSA, Elsevier, and Springer journals, serving as General Chair and Co-Chair of several IEEE, OSA, and SPIE conferences and workshops, serving on TPC of many conferences, as well as being a reviewer for scientific journals and project proposals.

Call for Papers
IEEE Communications Magazine
Green Communications and Computing Networks Series

Background

Green Communications and Computing Networks is issued semi-annually as a recurring Series in *IEEE Communications Magazine*. The objective of this Series is to provide a premier forum across academia and industry to address all important issues relevant to green communications, computing, and systems. The Series will explore specific green themes in depth, highlighting recent research achievements in the field. Contributions provide insight into relevant theoretical and practical issues from different perspectives, address the environmental impact of the development of information and communication technologies (ICT) industries, discuss the importance and benefits of achieving green ICT, and introduce the efforts and challenges in green ICT. This Series welcomes submissions on various cross-disciplinary topics relevant to green ICT. Both original research and review papers are encouraged. Possible topics in this series include, but are not limited to:

- Green concepts, principles, mechanisms, design, algorithms, analyses, and research challenges
- Green characterization, metrics, performance, measurement, profiling, testbeds, and results
- Context-based green awareness
- Energy efficiency
- Resource efficiency
- Green wireless and/or wireline communications
- Use of cognitive principles to achieve green objectives
- Sustainability, environmental protections by and for ICT
- ICT for green objectives
- Non-energy-relevant green issues and/or approaches
- Power-efficient cooling and air conditioning
- Green software, hardware, device, and equipment
- Environmental monitoring
- Electromagnetic pollution mitigation
- Green data storage, data centers, contention distribution networks, and cloud computing
- Energy harvesting, storage, transfer, and recycling
- Relevant standardizations, policies, and regulations
- Green smart grids
- Green security strategies and designs
- Green engineering, agenda, supply chains, logistics, audit, and industrial processes
- Green building, factory, office, and campus designs
- Application layer issues
- Green scheduling and/or resource allocation
- Green services and operations
- Approaches and issues of social networks used to achieve green behaviors and objectives
- Economic and business impact and issues of green computing, communications, and systems
- Cost, OPEX, and CAPEX for green computing, communications, and systems
- Roadmap for sustainable ICT
- Interdisciplinary green technologies and issues
- Recycling and reuse
- Prospect and impact on carbon emissions and climate policy
- Social awareness of the importance of sustainable and green communications and computing

Submission Guidelines

Prospective authors are strongly encouraged to contact the Series Editor with a brief abstract of the article to be submitted before writing and submitting an article in order to ensure that the article will be appropriate for the Series. All manuscripts should conform to the standard format as indicated in the submission guidelines at

http://www.comsoc.org/commag/paper-submission-guidelines

Manuscripts must be submitted through the magazine's submissions website at

http://mc.manuscriptcentral.com/commag-ieee

You will need to register and then proceed to the Author Center. On the manuscript details page, please select "Green Communications and Computing Networks Series" from the drop-down menu.

Schedule for Submissions

Inaugural Issue: November 2014
Scheduled Publication Dates: Twice per year, May and November

Series Editors

Jinsong Wu, Alcatel-Lucent, China, wujs@ieee.org
John Thompson, University of Edinburgh, United Kingdom, john.thompson@ed.ac.uk
Honggang Zhang, UEB/Supelec, France; Zhejiang University, China, honggangzhang@zju.edu.cn
Daniel C. Kilper, University of Arizona, United States, dkilper@optics.arizona.edu

Energy-Efficient Infrastructure Sharing in Multi-Operator Mobile Networks

Angelos Antonopoulos, Elli Kartsakli, Alexandra Bousia, Luis Alonso, and Christos Verikoukis

Angelos Antonopoulos is with the Telecommunications Technological Centre of Catalonia (CTTC) and with the Technical University of Catalonia (UPC).

Christos Verikoukis is with the Telecommunications Technological Centre of Catalonia (CTTC)

Elli Kartsakli, Alexandra Bousia, and Luis Alonso are with Technical University of Catalonia (UPC).

[1] Pease note that the terms "MNO" and "operator" are used interchangeably in this article.

ABSTRACT

Network infrastructure sharing and base station switching off mechanisms have been recently introduced as promising solutions toward energy and cost reduction in cellular networks. Although these techniques are usually studied independently, their combination offers new alternatives to MNOs for serving their users and could potentially provide them with additional benefits. In this article we introduce the concept of intra-cell roaming-based infrastructure sharing, where the MNOs may switch off their BSs and roam their traffic to active BSs operated by other MNOs in the same cell. Motivated by the coexistence of multiple operators in the same area, we present possible network deployments and architectures in current and future cellular scenarios, discussing their particular characteristics. In addition, we propose an innovative distributed game theoretic BS switching off scheme, employing an integrated cost function that takes into account all the different cases for a given operator to serve its own traffic (i.e. through active BSs of neighboring cells or exploiting intra-cell roaming-based infrastructure sharing). Finally, we demonstrate some indicative simulation results in realistic scenarios to quantify the potential energy and financial benefits that our proposed scheme offers to the MNOs in multi-operator environments, providing them with the necessary incentives to participate in the infrastructure sharing.

INTRODUCTION

The deployment of 4G technologies, though in an early stage, is changing the landscape in cellular networking. More specifically, a nearly 11-fold increase is expected in global mobile data traffic in the next five years, reaching 15.9 exabytes per month by 2018 [1], thus compelling mobile network operators (MNOs[1]) to extend their network infrastructure in an effort to increase network capacity and meet these pressing traffic demands. This rapid expansion of mobile cellular networks has become a major issue both for society, due to the high energy consumption, and for operators, due to the increased financial cost. In particular, the use of information and communication technology (ICT) across a wide range of applications currently accounts for 5.7 percent of the world's electricity consumption and 1.8 percent of global carbon emissions, something that translates into electricity bills in the order of $10 billion for MNOs worldwide [2].

Network infrastructure sharing has been introduced as a promising viable solution for MNOs toward the reduction of capital (CAPEX) and operational (OPEX) expenditures associated with the deployment and the operation of the cellular networks, respectively. This new paradigm, promoted by legal regulations that obligate the operators to install their antennas on the same buildings [3], embraces a set of strategies that enable the MNOs to use their resources jointly in order to reach their common goal, which is to guarantee customer service while achieving energy and cost reduction. From a technical perspective, infrastructure sharing can be classified into three categories with different levels of cooperation and control [4]:

- Passive sharing, limited to the joint use of sites, masts, and building premises among MNOs.
- Active sharing, where the MNOs share the active network components such as antennas, switches, and backhaul equipment.
- Roaming-based sharing, where one MNO relies on the coverage of another MNO in a certain region on a permanent basis.

Despite the technical challenges that might arise in such scenarios, the potential benefits of sharing part of the cellular network have been recently estimated to be as much as €2 billion [5], further motivating the cooperation among different MNOs.

In the context of green cellular networking, the role of the macro base stations (BSs) and their contribution to the total energy consumption should be highlighted. More specifically, given the large number of deployed BSs (approximately five million worldwide) in conjunction with their relatively high operational power (~1.7 kW), it may be concluded that approximately 80 to 90 percent of total network energy is consumed to power radio sites. However, cellular networks are dimensioned according to peak-hour traffic demands, and consequently a portion of their resources remains unexploited during several hours per day, when the network traffic is low. These statistics have lately motivated the research community to shift toward the investigation of disruptive switching-off schemes in an effort to achieve drastic energy efficiency gains. The initial works on BS switching-off considered clusters of cells with a single operator, where part of the BS infrastructure can be temporarily switched off, while the remaining active BSs can extend their coverage range in order to serve the whole network area. However, by exploiting the coexistence of multiple MNOs in the same area, it is possible to conceive a new promising sharing technique based on MNO cooperation: intra-cell roaming-based infrastructure sharing.

Intra-cell roaming-based infrastructure sharing extends the traditional roaming-based sharing in two ways: the sharing takes place in the same region, where the MNOs offer their mobile services; and the sharing is dynamic (and not on a permanent basis), especially during low-traffic periods. In addition, the involvement of various stakeholders (e.g. MNOs, trusted third parties, regulatory authorities, etc.) in future cellular environments entails a diversity of network architectures. As a result, additional challenges arise with respect to the BS switching-off, since the conventional solutions are usually applied in plain hexagonal cellular scenarios, without considering the particularities resulting from the coexistence of multiple MNOs in the same geographic area.

In this article our goal is threefold. First, we present some existing and future multi-operator network architectures, identifying their characteristics and the possible roles of the relevant stakeholders. Next we discuss the state of the art with respect to the BS switching-off schemes in single-operator networks, and we draw attention to the new challenges that arise in multi-operator environments. Finally, we introduce a game theoretic framework that enables the operators to take individual switching-off decisions for their own BSs. Besides the expected energy efficiency benefits, the proposed scheme allows the MNOs to significantly reduce their financial costs independently of the strategies of the coexisting MNOs, providing them with the necessary incentives to participate in the game.

Multi-Operator Network Architectures

In this section we present some of the possible scenarios (Fig. 1) and we briefly discuss the challenges and the traits of each particular case.

Mobile Virtual Network Operators (MVNOs)

The spectrum and the network infrastructure in a given region are deployed and owned by a single operator (MNO_A in the example of Fig. 1a). All the other operators in the same area are virtual (MVNOs) and, since they do not own any spectrum or network infrastructure, they must lease resources from MNO_A in order to serve their clients. This model, also known as national roaming, has already been applied successfully in several countries thanks to its simplicity and its inherent advantages. In particular, the existence of MVNOs may be beneficial for the end-user, since it promotes market competition, while at the same time the basic MNO in the area may capitalize the deployed infrastructure. On the other hand, it is uncertain whether this model can be compliant with the foreseen traffic demands in cellular networks, which probably cannot be met using only the existing infrastructure and the limited amount of resources. Moreover, the high spectrum prices make the operators reluctant to share their resources (and the market) with their competitors [6].

> *The existence of MVNOs may be beneficial for the end-user, since it promotes the market competition, while, at the same time, the basic MNO in the area may capitalize the deployed infrastructure. On the other hand, it is uncertain whether this model can be compliant with the foreseen traffic demands in cellular networks.*

Trusted Third Party

The whole infrastructure in a certain area has been deployed and owned by an independent trusted third party, and the MNOs, who hold a spectrum licence, may enter into agreements for the employment of the access and core network (Fig. 1b). The main benefit of this model, which is gaining momentum in many countries (e.g. Spain [7]), is the significantly lower CAPEX for the MNOs, who are not concerned anymore with the maintenance of the hardware infrastructure, being also able to provide their services dynamically in a given geographic region. However, the lease of the network implies an increased OPEX, which in the long run may not be profitable for the MNOs. In addition, the expected participation of many MNOs in future cellular networks could potentially result in high leasing prices, raising additional barriers and challenges for the effective application of this model.

Unique Infrastructure Provider

There is one operator that has deployed the whole network infrastructure and leases part of it to other interested MNOs (Fig. 1c). This scenario is fueled by the fact that the same spots (e.g. rooftops) are usually appropriate for all MNOs, and consequently an operator that has built out its network is able to capitalize on the potential interest of other MNOs in the specific location. In addition, this architecture can be considered as a hybrid model that combines some basic characteristics of the two aforementioned schemes, having though two main differences: the interested MNOs are not virtual, since they have their own spectrum license, and the infrastructure owner is an operator and not an independent entity. In this case, the operator who provides the infrastructure is burdened with considerably high CAPEX and OPEX, which, however, can be depreciated through efficient network leasing. From the perspective of the other MNOs, there is a trade-off between the expected CAPEX (lower) and OPEX (higher), while it should also be taken into account that they rely on a competitive entity rather than a trusted third party. This architecture includes stakeholders with different levels of risk and profit, thus raising intriguing financial challenges, which require explicit models for the accurate analysis of each entity according to the different profiles.

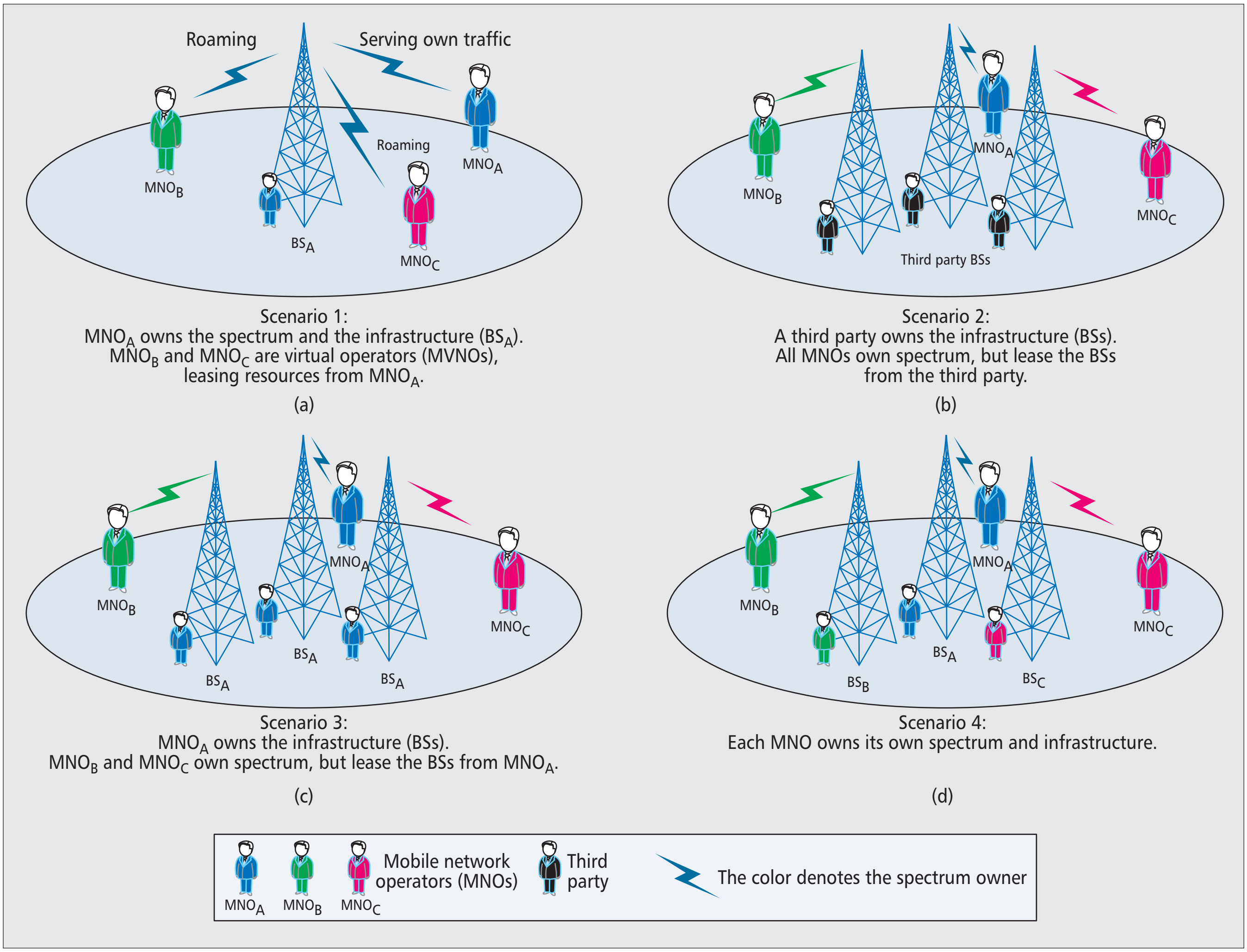


Figure 1. Multi-operator cellular network architectures.

Standalone

Various MNOs have deployed and run their own network in the same region (Fig. 1d). This model currently dominates in cellular networks, since it encompasses the lowest possible risk for the operators. Moreover, taking into account the rationality of the operators, along with the competitive nature of the telecommunications market, it is also very possible to appear in future architectures, especially in dense areas with high traffic. In this scenario, the MNOs have full control of their network, thus being able to estimate the expenses for both network deployment and operation.

Base Station Switching-Off and Infrastructure Sharing

In the previous section we shed some light on the different sharing solutions that can be encountered in multi-operator networks. Nonetheless, regardless of the particular sharing agreement, the energy issue has always been in the spotlight of cellular networking, especially as the design and engineering of traditional cellular networks typically take place considering the peak traffic demands. As a result, efficient joint resource allocation and network planning during low traffic periods (e.g. night zone) would allow the service of the traffic by a smaller number of active BS, thus enabling the deactivation of the remaining unused infrastructure. However, the decision of the particular BSs that should be switched off is not trivial, as it affects end user satisfaction, which is of top priority for telecommunication operators. In addition, this decision is even more complicated in multi-operator environments due to the conflicting interests of the operators and the diversity of the involved parameters. To that end, the design of BS sleeping mechanisms has recently become a very hot research topic, with the majority of the works focusing on the reduction of network energy consumption without compromising quality of service (QoS) for the mobile users. In this section we briefly discuss the state of the art in BS switching-off schemes in single- and multi-operator networks.

The core idea behind the works that deal with single-operator networks is that a given MNO is able to switch off part of its infrastructure and extend the transmission range (by increasing the transmit power) of the remaining active BSs in order to provide coverage to the whole network. Several research attempts study

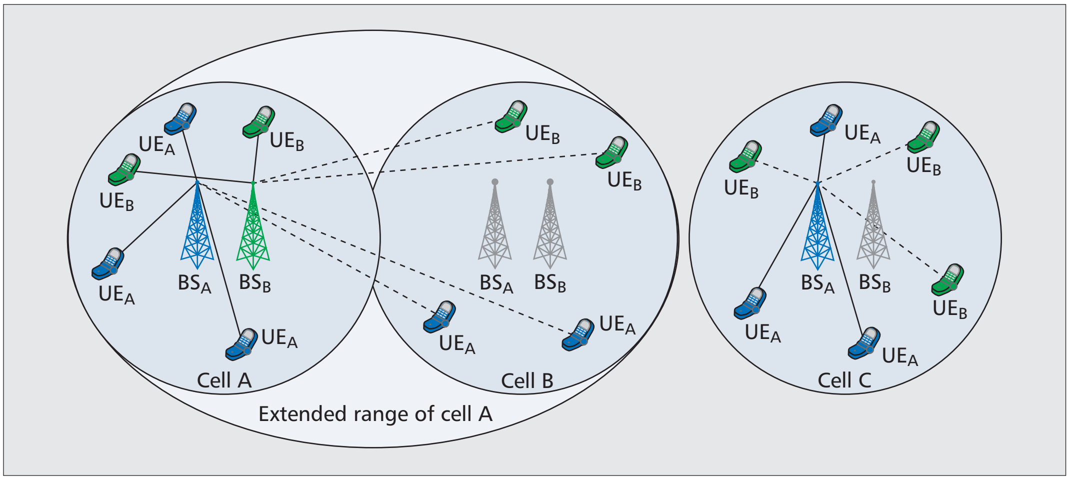


Figure 2. Possible cell operations in multi-operator environments.

We introduce a novel game theoretic switching off scheme for the BSs in a multi-operator environment, taking into account the conflicts and the interaction among the different MNOs, as well as the different available courses of action.

the potential impact on network performance of random [8] or specific [9, 10] patterns for switching off a subset of the BSs. These works focus rather on the spectral efficiency and the offered QoS (in terms of call-blocking probability) after the deactivation of the BSs, taking the reduced energy consumption for granted.

In addition, there are many works oriented to the switching-off decision itself. More specifically, different criteria (e.g. traffic load [11], user spatial distribution [12], or network-impact [13]) are considered in order to identify the optimal BS switching-off strategy, guaranteeing a certain level of user satisfaction in the network. In the same context, a very recent interesting approach is the application of reinforcement learning schemes that cope with the dynamic nature of the traffic load in current cellular networks [14] in order to overcome an important limitation of the existing works, which rely on past predefined traffic patterns based on the network history.

Despite their novelty and the promising energy savings, most works usually neglect the possible coexistence of the BSs, which offers to the MNOs the alternative of being served by other MNOs, co-located in the same area. Figure 2 illustrates three possible cases for the cell operation in multi-operator environments, where each operator controls and operates its own BS (i.e. the architecture presented in Fig. 1d). In the simple case where all the BSs are active (Cell A), each operator is responsible for serving the traffic of its users. In the case that all BSs have been switched off (Cell B), the active BSs of the neighboring cells (Cell A in the example) of each MNO can extend their range in order to prevent the generation of coverage holes. Finally, in the case that only a subset of the BSs has been switched off (Cell C), their respective traffic can be roamed to the active BSs of the same cell.

The aforementioned cases stress the need for new BS switching-off mechanisms that consider all the contingencies in future networks. In their pioneering work [15], Marsan and Meo have proposed two interesting BS sleeping schemes for multi-operator environments: roaming-to-one and roaming-balanced. In the former, only one BS remains active in every cell during low traffic conditions, serving the users of all operators, while with the roaming-balanced scheme, the MNOs in a particular cell switch off their BSs for different portions of time in order to balance the expected financial expenses associated with the roaming process. Although both schemes achieve noticeable energy savings, they neglect the possibility of switching off all BSs in a given cell and transferring the traffic to an active neighboring cell. In the next section, given the necessity for innovative solutions that take into consideration all the possible cases for cell operation along with the important parameters involved in future cellular networks (i.e. roaming and operational cost, energy consumption, QoS), we present a game theoretic switching-off algorithm that enables effective intra-cell roaming-based infrastructure sharing, providing significant gains to the participants operators.

Game Theoretic Base Station Sleeping Mechanism

In a multi-operator environment, the involvement of multiple stakeholders with opposing interests and different parameters that potentially affect the switching-off decision (e.g. energy savings, roaming cost, traffic load, etc.) make game theory a suitable tool to study this problem. To that end, we introduce a novel game theoretic switching-off scheme for the BSs in a multi-operator environment, taking into account the conflicts and the interaction among the different MNOs, as well as the different available courses of action.

We consider clusters of M cells (one central and $M - 1$ peripheral cells), with N operators offering their services in each cell through different co-located BSs. The N BSs in the central cell always remain active in order to offer an extra alternative to the operators. More specifically, in the case that all BSs in a peripheral cell have been deactivated and the intra-cell roaming-based infrastructure sharing is no longer possible, the MNOs are allowed to transfer the traffic load to the corresponding BS of the central cell. Therefore, the distributed network operation is further facilitated, enabling the modeling of the switching-off decision process in the peripheral

In the case that all BSs in a peripheral cell have been deactivated and the intra-cell roaming-based infrastructure sharing is no longer possible, the MNOs are allowed to transfer the traffic load to the corresponding BS of the central cell.

		Player B: N-1 MNOs (excluding MNO_i)	
	Actions	Remain ON	Switch OFF
Player A: MNO_i	Remain ON	$C_{const} + C_{tr}$	$C_{const} + (\bar{N}_{roam} + 1)C_{tr} - \bar{N}_{roam}C_{roam}$
	Switch OFF	C_{roam}	C_{inc}

Figure 3. Cost matrix of the proposed game.

cells as a static non-cooperative game. The specific game theoretic formulation provides two important benefits. First, it is a distributed solution, applied by every operator individually in each peripheral cell, something that constitutes an important and practical feature in multi-operator environments, as the operators are not obliged to follow a specific centralized policy. The second benefit is that the proposed scheme does not require the exchange of excessive control information among the operators, since the BSs of the same cell should make their decisions without any communication among them.

The proposed game consists of N players (i.e. the MNOs in each cell) that have two possible actions (i.e. either remain on or switch off) and a cost function C that corresponds to the actual cost paid by each operator in every peripheral cell, including the cost of operation, roaming, and power increase of the central BS. It is also worth highlighting the symmetry of the problem, as the participants (MNOs) have identical characteristics and the outcome of the game depends only on the selected strategies. Hence, by definition, the proposed game is symmetric, allowing us to:

- Characterize it as game of complete information, as the players have a common cost function.
- Formulate it with two macro-players (i.e. player A is a given MNO_i, while player B consists of the set of the remaining $N - 1$ MNOs).
- Study the problem from player A's point of view, generalizing the conclusions for every player.

Hence, considering the perspective of player A, four different cases can be distinguished:

Case 1: Player A is ON, all operators in Player B are ON: The total cost for the MNO under study includes the fixed operational cost for the BS (C_{const}) and the variable cost for serving its traffic (C_{tr}).

Case 2: Player A is ON and a subset of operators in Player B are OFF: In this case the switched-off BSs must roam their traffic to the active BSs. In general, player A may serve, on average, $\overline{N}_{roam}$ BSs. Thus, the total cost of player A must take into account the increased operational cost for serving the additional traffic $((\overline{N}_{roam} + 1) \cdot C_{tr})$, as well as the corresponding income from the roamed MNOs $(\overline{N}_{roam} \cdot C_{roam})$.

Case 3: Player A is OFF and at least one operator in Player B is ON: The operator under study does not have any operational cost, but it should pay the associated roaming cost (C_{roam}) to the active operator that serves its traffic.

Case 4: Player A is OFF, all operators in Player B are OFF: In this case the cost paid by each MNO corresponds to the extra energy consumption for the power increase of the central BS (C_{inc}) in order to cover the area of a switched-off BS.

The aforementioned cases can be summarized in the matrix representation of the game in Fig. 3. Having formulated the game in a strategic form, the next step is the identification of the stable states that define the actions of the players. These states, also known as Nash Equilibria in non-cooperative games, provide the players with a certain payoff (resp. cost) that cannot be increased (resp. decreased) if they decide to unilaterally alter their decision. However, in real systems, selecting a pure action is not always feasible or fair. In our problem, for instance, the selection of a pure strategy, where only a specific subset of the BSs remains active (i.e. Cases 2, 3), would require centralized control and coordination, while the deactivation of all BSs in all peripheral cells (Case 4) and the transfer of their traffic in the central cell might compromise the offered QoS in high traffic load scenarios.

Therefore, to overcome this limitation, we study the problem in the mixed strategies domain, where each MNO_i selects a specific action with a certain probability. To that end, we define s_i as the probability of player A (i.e. MNO_i) to switch off its BS, whereas the game's symmetry enables the grouping of the other $N - 1$ operators (excluding MNO_i), assuming a common switching-off probability s_j. Exploiting the strategic representation of the game, along with the switching-off probabilities, we may derive the expected cost function for player A as $\mathbb{E}[C] = f(s_i, s_j, N, Cost_k)$, where $Cost_k$, $k \in \{1, 2, 3, 4\}$ denotes the respective cost in each of the four different cases described above.

The probability that minimizes the cost function (i.e. the root of the equation $(\partial\mathbb{E}[C])/(\partial s_i) = 0$) corresponds to the equilibrium of the game. More specifically, following this strategy, the players minimize their cost in a distributed manner, thus having no incentives to change strategy. Regarding the details of the expected cost function, as we may also observe in Fig. 3, four different costs are involved:

- The fixed cost for the BS operation (C_{const}).
- The cost for serving the traffic (C_{tr}).
- The cost for increasing the power of the central BS in order to serve the traffic of neighboring cells (C_{inc}).
- The cost for roaming the traffic to another operator in the same cell (C_{roam}).

While the first three costs (i.e. C_{const}, C_{tr} and C_{inc}) are predictable and almost constant,[2] C_{roam} can significantly vary, depending on the intentions and the demands of the operators. To that

Roaming cost coefficient

Number of MNOs: N	α = 0.1	α = 0.3	α = 0.5	α = 0.7	α = 0.9
2	0.459	0.377	0.296	0.215	0.133
3	0.679	0.629	0.568	0.486	0.357
4	0.799	0.749	0.675	0.584	0.466
5	0.854	0.826	0.795	0.758	0.713
6	0.880	0.861	0.843	0.819	0.798

As *N* increases, MNOs have a stronger incentive to switch off their BS and roam their traffic to other operators

As α increases, the additional cost for traffic roaming discourages MNOs from switching off

Figure 4. NE switching off probabilities.

end, let us define $C_{roam} = a \cdot (C_{const} + C_{tr})$, with $a \in [0, 1]$ (as it does not make sense to have a roaming cost higher than the operational cost), and present some indicative Nash Equilibrium results in Fig. 4 in order to study the impact of N (number of MNOs) and a (roaming cost) on the switching-off probabilities.

Two main conclusions can be derived from Fig. 4. First, the NE switching-off probabilities increase with the number of MNOs in each cell. In particular, as the number of coexisting operators in the same cell grows, MNOs have a stronger incentive to switch off their BSs and roam their traffic to other operators. Regarding the second basic observation, as expected the switching-off probability decreases for higher roaming cost, which is a prohibitive factor for BS deactivation. However, it is worth noting that the NE probability reduction is more intense in networks with few MNOs due to the risk of switching off all BSs in the cell. On the other hand, in networks with many MNOs, where the aforementioned danger is not so evident, the switching-off probability for high a is still significant.

It is worth noting that the NE probability reduction is more intense in networks with few MNOs due to the risk of switching off all BSs in the cell. On the other hand, in networks with many MNOs, where the aforementioned danger is not so evident, the switching off probability for high a is still significant.

Quantifying the Gains

Simulation Scenario

We have developed a system-level C++ simulator for the network operation to assess the performance of the proposed infrastructure sharing scheme and quantify the expected gains in terms of energy efficiency and expenditures for the operators. The simulation scenario includes a typical urban 7-cell cluster (i.e., $M = 7$) of six peripheral cells and one central cell that always remains on. In each cell, N MNOs provide their services to a set of users,[3] while we focus on the night zone where all operators have low traffic. Regarding the technical parameters, we assume a service rate of 64 kb/s and 256 kb/s for voice and data sessions, respectively, while we consider two different transmission power levels for the BSs, equal to 40 dBm and 46 dBm for normal and extended operation, respectively. For the financial analysis, we assume $C_{const} \simeq 465$€, while the cost for the transmission and the power increase can take different values depending on the traffic level, that is, $C_{tr} \in [5, 592]$€ and $C_{inc} \in [6, 613]$€. In addition, we have adopted a price of 0.1€kWh for the electricity charge, and the roaming cost factor has been set equal to $a = 0.1$. In our experiments, we consider $N = \{4, 5, 6\}$ in order to study the impact of the number of MNOs in future networks. To evaluate the performance of our scheme, we compare the proposed game theoretic infrastructure sharing strategy (referred as GTIS in this section) with two state of the art approaches [15]:

- The roaming-balanced (R-bal) scheme, where the MNOs switch off their BSs for different portions of time in order to balance their roaming costs.
- The roaming-to-one scheme (R-to-1), where only the MNO with the highest traffic remains active and serves the total network traffic.

Performance Results

The normalized system throughput, defined as the percentage of the served users, is depicted in Fig. 5a. As can be seen, the proposed scheme guarantees the user satisfaction in the network in all cases, while the state of the art approaches experience losses in some scenarios. More specifically, as the number of MNOs increases in each cell, the network becomes overloaded and the deactivation of the BSs results in throughput degradation, as the infrastructure and the resources cannot be shared in an efficient way. As a result, in the case of $N = 6$, R-bal fails to satisfy approximately 2 percent of the network users, a percentage that can be prohibitive for wireless cellular networks. The losses are even greater in the R-to-1 scheme (~10 percent), and therefore the service of the traffic in future multi-operator environments by only one BS in each cell cannot be considered as a viable solution.

Figure 5b presents the network energy efficiency achieved by the three schemes under study, considering topologies with $N = 4$, 5, and 6 operators. The first important observation is that, unlike the normalized throughput trend, the energy efficiency increases with the number of operators. More specifically, the high number of BSs in each cell implies high traffic and large energy consumption. Consequently, switching-off schemes can contribute significantly in energy saving, and the energy efficiency increases as most of the sessions can still be served by the active BSs. Even in the case of the R-to-1 scheme in scenarios with $N = 6$ operators, the energy efficiency is extremely high, despite the throughput degradation, since the

[2] The traffic variations during night zone are negligible and, as a result, the C_{tr} can be considered constant.

[3] Each operator serves 100 users, where the portion of data and voice users is 70 percent and 30 percent, respectively.

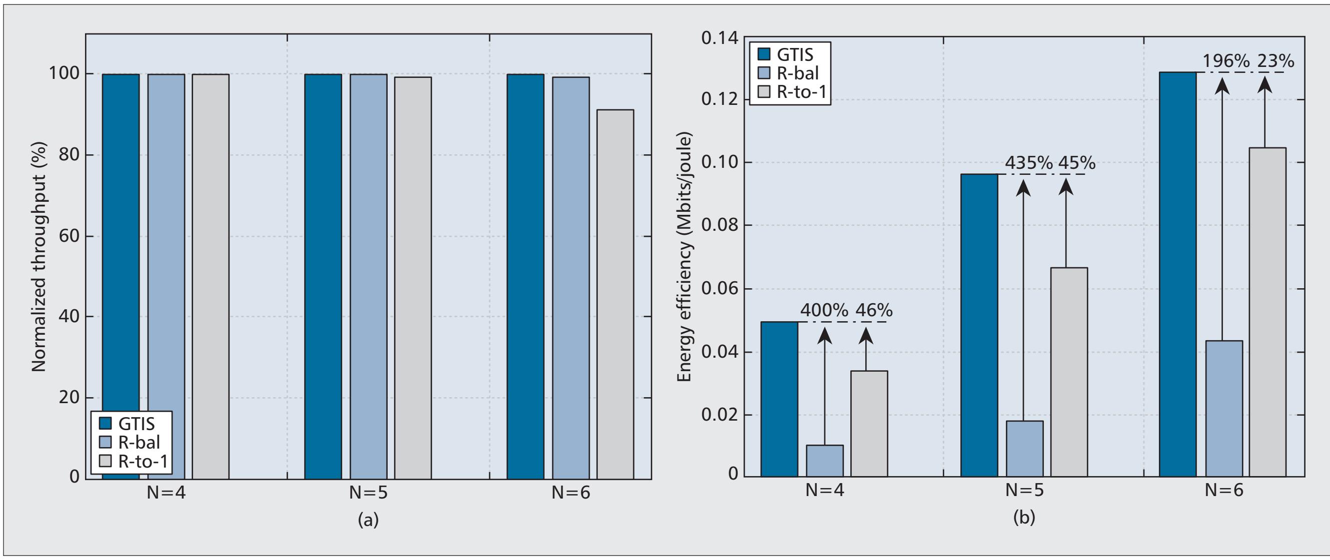


Figure 5. Network performance metrics vs. number of operators (N = 4, 5, 6): a) normalized throughput; b) energy efficiency.

deactivation of five BSs excessively diminishes the energy consumption in each cell. The second noticeable observation concerns the pronounced energy efficiency gains of the proposed GTIS scheme compared to the R-bal algorithm in all scenarios. Although the two schemes exhibit similar performance with regard to the QoS, the GTIS enables the deactivation of more BSs, thus achieving up to four times higher energy efficiency.

Regarding the financial analysis, we focus on the case of N = 4 operators, which is one of the dominant scenarios in European countries [16]. The individual revenue for each operator that participates in the infrastructure sharing is plotted in Fig. 6a, where we may observe that the R-to-1 scheme provides financial gains only to particular operators. More specifically, only the MNOs that switch off their BSs can benefit from this scheme, while significant expenses burden the operator that undertakes the service of the total cell traffic by maintaining its BS active. On the other hand, adopting the proposed GTIS approach, all operators in the network achieve higher gains compared to the R-bal scheme, as the infrastructure is shared in a more integrated way. Finally, the total annual cost for the network is shown in Fig. 6b. As can be seen, the difference in the network energy efficiency (Fig. 5b) is directly reflected in the financial cost. More specifically, the throughput guarantees provided by R-bal come at an increased energy and economic cost, while GTIS reduces this cost up to 77 percent by deactivating an additional number of BSs, which are not necessary during low traffic periods. In any case, these results (i.e. throughput, energy efficiency, and financial cost) should be jointly studied in order to decide the most appropriate switching-off strategy that maximizes the operators' gains without compromising end-user QoS.

Conclusion

In this article we discussed energy saving techniques in multi-operator cellular environments. Motivated by the coexistence of several MNOs in the same area, we introduced the concept of intra-cell roaming-based infrastructure sharing, which offers additional options to the operators for having their traffic served in regions where they decide to switch off their BSs. In addition, we proposed a novel game theoretic BS switching-off scheme that enables the MNOs to achieve significant energy and, consequently, cost savings, by deciding distributively the deactivation of their BSs in a given cell. The simulation analysis showed the potential gains of the proposed scheme in terms of energy and financial cost, providing the operators with the necessary incentives to participate in the network infrastructure sharing. Finally, it is worth mentioning that the foreseen massive small cell deployment in next generation networks is expected to have a key role in the design of effective infrastructure sharing and switching-off schemes, something that will be addressed in our future research.

Acknowledgment

This work has been supported by AGAUR (2014 SGR 1551) and the research projects GREENET (264759) and GEOCOM (TEC2011-27723-C02-01).

References

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013-2018," Feb. 2014.

[2] E. Oh *et al.*, "Toward Dynamic Energy-Efficient Operation of Cellular Network Infrastructure," *IEEE Commun. Mag.*, vol. 49, no. 6, June 2011, pp. 56–61.

[3] GSMA TM., "Base Station Planning Permission in Europe," Dec. 2013; available: http://www.gsma.com/publicpolicy/base-station-planning-permission-in-europe.

[4] A. Khan *et al.*, "Network Sharing in the Next Mobile Network: TCO Reduction, Management Flexibility, and Operational Independence," *IEEE Commun. Mag.*, vol. 49, no. 10, Oct. 2011, pp. 134–42.

[5] Applied Value Group, "Infrastructure Sharing Among MNOs," *Applied Value Telecom Series*, 2014.

[6] R. Cellan-Jones, "National Roaming — A Bad Call?" July 2014; available: http://www.bbc.com/news/technology-28149227.

[7] Abertis, "Abertis Reaches an Agreement with Telefónica and Yoigo on the Acquisition of a Minimum Package of 4.227 Mobile Telephone Towers," Press Release, Aug. 2013; available: http://www.abertis.com/dynadata/01082013_AdquisicionTorresMoviles_EN.pdf.

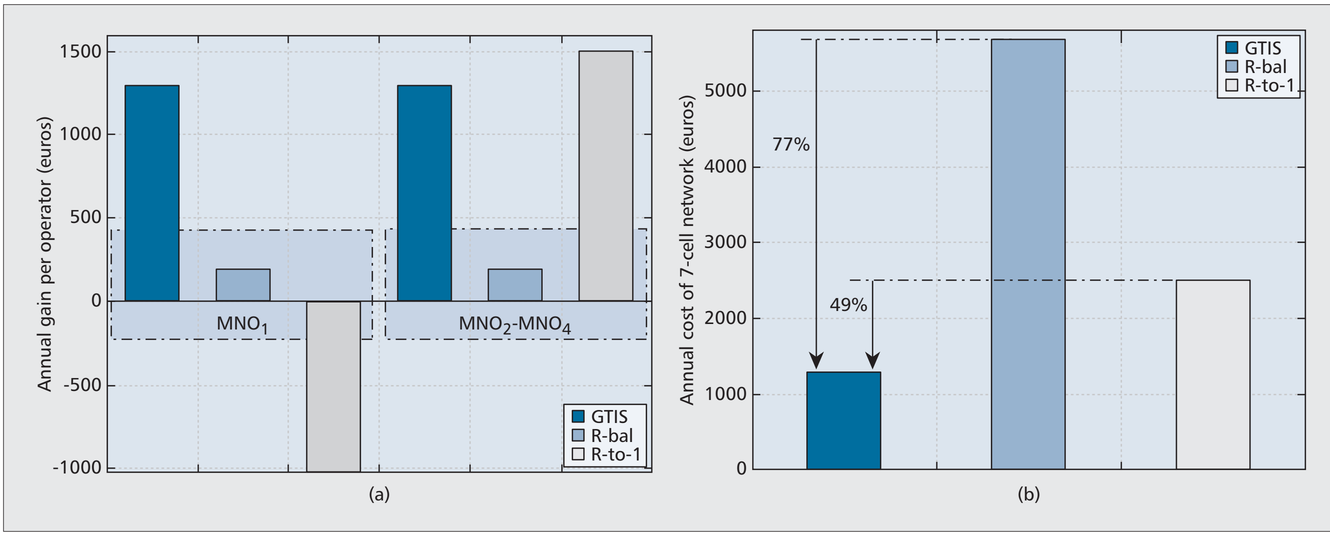


Figure 6. Annual financial data for networks with $N = 4$ operators: a) individual annual gains per operator; b) total annual cost per network.

[8] L. Chiaraviglio *et al.*, "Energy-Efficient Management of UMTS Access Networks," *Proc. 21st Int'l. Teletraffic Congress* (ITC), Sept. 2009, pp. 1–8.

[9] F. Han, Z. Safar, and K. J. R. Liu, "Energy-Efficient Base-Station Cooperative Operation with Guaranteed QoS," *IEEE Trans. Commun.*, vol. 61, no. 8, Aug. 2013, pp. 3505–17.

[10] H. Tabassum *et al.*, "Downlink Performance of Cellular Systems with Base Station Sleeping, User Association, and Scheduling," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, Oct. 2014, pp. 5752–67.

[11] J. Wu, S. Zhou, and Z. Niu, "Traffic-Aware Base Station Sleeping Control and Power Matching for Energy-Delay Trade-offs in Green Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, Aug. 2013, pp. 4196–4209.

[12] A. Bousia *et al.*, "Green" Distance-Aware Base Station Sleeping Algorithm in LTE-Advanced," *Proc. IEEE ICC*, June 2012, pp. 1347–1351.

[13] E. Oh, K. Son, and B. Krishnamachari, "Dynamic Base Station Switching-On/Off Strategies for Green Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, May 2013, pp. 2126–36.

[14] R. Li *et al.*, "TACT: A Transfer Actor-Critic Learning Framework for Energy Saving in Cellular Radio Access Networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, April 2014, pp. 2000–11.

[15] M. A. Marsan and M. Meo, "Energy Efficient Wireless Internet Access with Cooperative Cellular Networks," *Computer Networks*, vol. 55, no. 2, Feb. 2011, pp. 386–98.

[16] M. A. Marsan and M. Meo, "Network Sharing and its Energy Benefits: A Study of European Mobile Network Operators," *Proc. IEEE GLOBECOM*, Dec. 2013.

Biographies

Angelos Antonopoulos received his Ph.D. degree from the Signal Theory and Communications (TSC) Department of the Technical University of Catalonia (UPC) in 2012. He is currently working on various research topis, including energy efficient network planning, cooperative communications, e-Health applications, MAC protocols, and network coding. He has been granted three annual scholarships by the Greek State Scholarships Foundation (IKY) and the First Polytechnic Graduates Prize by the Technical Chamber of Greece (TEE-TGC). In December 2014 he received the best paper award at IEEE GLOBECOM and the best demo award in IEEE CAMAD, while in January 2015 he was nominated as an exemplary reviewer for *IEEE Communications Letters*.

Elli Kartsakli received her Ph.D. in wireless telecommunications from the Technical University of Catalonia (UPC) in February 2012. She holds a degree in electrical and computer engineering from the National Technical University of Athens, Greece (2003), and an M.Sc. in mobile and satellite communications from the University of Surrey, UK (2004). She is currently a post-doctoral researcher in UPC and has participated in several national and European projects (GREENET, WSN4QoL, etc.). Her primary research interests include wireless networking, channel access protocols, and energy efficient communication protocols.

Alexandra Bousia received her B.S. and M.S. degrees from the Department of Computer & Communication Engineering, University of Thessaly, in 2008 and 2009, respectively. She is currently working toward the Ph.D. degree in wireless communications at the Technical University of Catalonia (UPC). Since 2011 she has been involved in several national and European research funded projects. She received the best paper award at IEEE GLOBECOM'14. Her research interests include wireless networks, and energy efficient RRM algorithms.

Luis Alonso obtained a Ph.D. from UPC (Barcelona) in 2001, and was given a permanent tenured position at the same University becoming an associate professor in 2006. He was a co-founder of the Wireless Communications and Technologies Research Group (WiComTec), to which he currently belongs. His current research interests are in the fields of medium access protocols, radio resource management, cross-layer optimization, cooperative transmissions, cognitive radio, and QoS features for all kinds of wireless communications systems. He is an author of more than 50 research papers in international journals and magazines, one book, 12 book chapters, and more than 100 papers in international congresses and symposiums. He has received several best paper awards.

Christos Verikoukis received his Ph.D. from the Technical University of Catalonia in 2000. He is currently a senior researcher at CTTC and an adjunct professor at UB. His area of expertise is in the design of energy efficient layer 2 protocols and RRM algorithms for short-range wireless cooperative and network coded communications. He has published 68 journal papers and over 140 conference papers. He has also co-authored two books, 14 book chapters, and two patents. He has participated in more than 30 competitive projects while he served as the principal investigator in national projects in Greece and Spain. He has supervised 15 Ph.D. students and five post doc researchers since 2004. He is currently a vice-chair of the IEEE ComSoc Technical Committee on Communication Systems Integration and Modeling (CSIM). He received the best paper award at the Communication QoS, Reliability & Modeling Symposium (CQRM) at IEEE ICC'11, and the EURASIP 2013 Best Paper Award for the *Journal on Advances in Signal Processing*.

A Low-Cost Methodology for Profiling the Power Consumption of Network Equipment

Andrea Francini, Steven Fortune, Thierry Klein, and Marco Ricca

ABSTRACT

Rate adaptation technologies aim at establishing a linear relationship between power consumption and traffic load in packet networks. They rely on power profiles of network elements, which map system configurations and traffic loads onto power consumption levels, for the selection of network resources to place into low-power states and to identify new system designs with high power-saving yields. We introduce a methodology for profiling the power consumption of network systems that reconciles modeling accuracy with cost containment and rapidity in the preparation and execution of power measurements. We apply the methodology to network systems from multiple vendors and find it capable of delivering a clear message: the power savings enabled by protocol and system software upgrades that support demand-timescale rate adaptation are worthwhile, but also largely inferior to those attainable with a new generation of hardware platforms that pervasively deploy packet-timescale rate adaptation.

INTRODUCTION

In packet networks, the term rate adaptation designates a broad set of methods aimed at establishing a direct relationship between sustained workload and power consumption. In an ideal framework for energy efficiency, the network minimizes power consumption under full-load traffic conditions. Then, rate adaptation additionally ensures that the power-workload function is linear and that the network consumes no power when there are no packets to transport (in reality a small power offset for support of basic control functions is unavoidable). To support such behavior, rate adaptation schemes provide the network systems that they control with a discrete set of operating states, where each state maps a fixed traffic processing rate onto a respective power consumption level. A scheme is acceptable only if it does not introduce perceivable degradation of network performance.

The scope of the control exercised by a rate adaptation scheme can range from large subsets of network links and nodes to individual sections of a single traffic processing chip. We classify rate adaptation techniques based on the time consumed by their transitions between states. Demand-timescale rate adaptation (DTRA) techniques control the state of network links and nodes based on expected traffic demands between network endpoints [1–3]. DTRA state transitions involve network signaling and system-level power cycles, so their timescale ranges from seconds to minutes. Packet-timescale rate adaptation (PTRA) techniques adjust the clock frequency and supply voltage of data path hardware components to locally maintained workload indicators such as queue lengths and traffic arrival rates [4–6]. The timescale of PTRA state transitions ranges from microseconds to milliseconds depending on the underlying integrated circuit technology. Bit-timescale rate adaptation (BTRA) also applies to data path hardware components, but its transitions are much faster (down to nanoseconds) because they only involve control of the system clock (e.g. by gating of the clock signal [7]), at the expense of reduced power savings.

Profiling the power consumption of network equipment is instrumental in assessing the power-saving benefits of the different types of rate adaptation. With DTRA techniques, the power profiles quantify the benefits of enabling and disabling network ports, line cards, and systems in response to changes in traffic demands; in the case of PTRA and BTRA, the profiles identify the power-saving margins for the introduction of rate-adaptive hardware components. The power profile of a network element maps system and traffic configurations onto power consumption levels, typically by means of a simplified linear model [1–3]. Relevant system configuration variables include the number of cards plugged into the chassis (in slotted systems), the number of ports that exchange traffic over network links, and the transmission capacity of those ports. Traffic configuration variables

Andrea Francini, Steven Fortune, and Thierry Klein are with Alcatel-Lucent Bell Labs.

Marco Ricca is with Politecnico di Torino.

0163-6804/15/$25.00 © 2015 IEEE

include the traffic arrival rate and the statistical distribution of packet sizes and packet inter-arrival times at ports where traffic is present.

While power profiles are commonly available for computing systems and processors, studies that focus on networking systems and components have started appearing in the literature only recently [8–12] and suffer from a variety of limitations. The power profiles presented in those studies are not always complete in the identification of system configuration variables [8] or in the modeling of critical system components [9], or rely mostly on manufacturer power-rating data rather than experimental measurements [10], or involve expensive equipment that can be justified only in a specialized power-measurement lab [11, 12]. Profiling approaches that condense the energy-efficiency properties of a system into a small number of scalar indices [13, 14] cannot support the fine-tuned state-setting decisions that are at the core of all rate adaptation methods.

We describe for the first time a low-cost methodology for the fast compilation of dependable power profiles with a minimum of test equipment, which is ideal for research and development teams that have only occasional interest in power profiling. The methodology is based on a linear model for power consumption that crisply isolates the potential for power saving that is associated with rate adaptation techniques at all timescales. Application of the methodology to a variety of network systems shows that common commercial equipment still lacks support for the type of state transitions that can elevate the impact of DTRA techniques, and that important power savings can be obtained through the introduction of PTRA/BTRA capabilities in data path hardware components, especially those that reside in the chassis. We refer the reader to [15] for a more thorough discussion of the different types of rate adaptation and their afforded power-saving opportunities, together with the formal construction of the linear model for power profiling.

Linear Model for Power Profiles

In this section we summarize the linear model that we follow for power profiling. The model is adapted to the coarse resolution of a low-cost power meter that only reports the variations of electrical current absorbed by the system under test (SUT).

Definitions

Throughout the article, RJ-45 identifies an integrated twisted-pair Ethernet port, XFP identifies a 10GbE small form-factor pluggable module, and SFP identifies a 1GbE module or generically a 1GbE/10GbE module when the type distinction is not relevant.

A modular port of an SUT is *loaded* if it has an SFP module attached; otherwise it is an *empty* port. A loaded SFP port or an integrated RJ-45 port is *connected* if a network cable connects the port to a peering interface on the same system or on a traffic generator/sink, and *disconnected* otherwise. A network port is *enabled* if it is configured for operation at a set rate; otherwise it is *disabled*. A port can be switched between the enabled and disabled states when it is empty, when it is loaded but disconnected, or when it is connected. However, we are only interested in the distinction between the enabled and disabled states in the particular case where the port is connected, because this is the kind of state transition that DTRA techniques can control. A connected port is *busy* if it exchanges traffic with the attached network cable and *idle* otherwise. A busy port is *saturated* when it receives or transmits traffic at a rate equal to its nominal capacity.

Isolation of Traffic Contributions

In the data path of a network system we find hardware components whose power consumption depends mostly on the bit rate of the sustained traffic (e.g. switch fabric modules) and others for which it depends mostly on the packet rate (e.g. packet processors). A power meter that only captures fluctuations of the current absorbed by the system cannot identify the power consumed by each component, but can detect the effects of varying bit and packet rates on the overall power consumption. The bit rate β and the packet rate ρ should contribute with distinct terms to the power consumed by each controllable system component (chassis, line card, and port). However, since β and ρ are not independent of one another ($\rho = \beta/\bar{\sigma}$, where $\bar{\sigma}$ is the *average packet size*), we replace β with a *packet-size load* $q(\bar{\sigma})$ $(\sigma_{max} - \bar{\sigma})/(\sigma_{max} - \sigma_{min})$ such that the packet-rate contribution to the power consumption is null when the average packet size is maximum (i.e. the packet rate is minimum for the given bit rate), and maximum when the packet size is minimum (the packet rate is maximum for the given bit rate).

Linear Model

The ideal linear model for application in rate-adaptation contexts sums the power contributions of all system components whose state can be controlled by external action, whether by network signaling, by system management interface, or by physically plugging or unplugging hardware. These components include the chassis, the line cards (when present), and the network ports with respective pluggable modules. Every component should have one term that expresses the fixed cost of keeping it powered on and one that is sensitive to traffic. The contributions of bit rate and packet rate should be distinguished in parts of the system where the packet size is variable. The following subsections outline how we reconcile the above requirements with the coarseness of a low-cost power meter. A more formal description of our simplified linear model can be found in [15].

Chassis Power: The chassis power C should include a fixed term C_0 and a variable term that depends on the aggregate bit rate sustained by the switch fabric (data units are equally sized in the switch fabric, so the packet-size load is null). However, the measurement error of the power meter (between 0.05 W and 0.5 W in our experiments) systematically masks out the bit-rate sen-

The ideal linear model for application in rate-adaptation contexts sums the power contributions of all system components whose state can be controlled by external action, whether by network signaling, by system management interface, or by physically plugging or unplugging hardware.

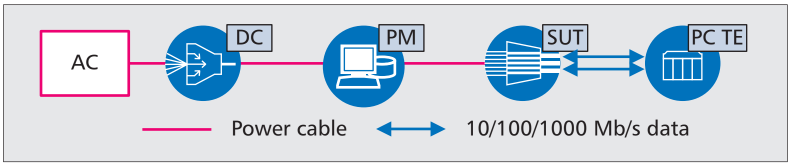


Figure 1. Experimental testbed for power measurements.

sitivity of the chassis. We conclude that it is acceptable, until PTRA capabilities become available in switch-fabric hardware, to reduce the chassis power to the fixed component alone: $C \triangleq C_0$.

Line-Card Power: The line card, when present, is the place where packets that are associated with multiple ports undergo the format conversion from network to switch fabric and vice versa. The line-card power L should include a fixed term L_0 and variable terms that depend on the line-card's input and output bit-rate and packet-size loads.

The switch-fabric adapter is one device in the line card where the bit-rate contribution dominates over its packet-rate counterpart because it handles fixed-sized data units formatted for the time-slotted operation of the switch fabric. Packet-rate dominance over bit rate is expected instead in the packet processor, where the amount of work applied to the packet header is independent of the packet size. However, the inline power meter compromises our ability to discern the traffic-sensitive power contributions of the line-card devices from those of the ports they serve. We decide to concentrate all traffic-sensitive terms at the port level, identifying the line-card power with its fixed term: $L \triangleq L_0$.

Port Power: Having simplified the chassis and line-card terms, the network port remains the only configurable component of the system where we can retain traffic-sensitive contributions to power consumption. The limited availability of traffic endpoints that is typical of a low-cost testbed induces measurement inaccuracies that hinder the distinction between input and output load parameters. The *bit-rate sensitivity* P_b and *packet-size sensitivity* P_r of our model only capture combined input-output port loads.

Preliminary measurements on idle systems [15] show that the fixed power contribution P_0 of a port must be split into two distinct terms: the *fixed hardware port power* $P_0^{(h)}$ and the *fixed software port power* $P_0^{(s)}$. The term $P_0^{(h)}$ captures the contribution of the port when it is loaded with an SFP, whether or not it is enabled for operation. Obviously it is null in the case of integrated RJ-45 ports. Its isolation offers the network operator the option to save power by unplugging the SFPs from modular ports that remain disabled for extended periods of time. It also offers system vendors an incentive to add to their designs provisions for controlling this power contribution (and the associated energy waste in the case of disabled ports) via software. $P_0^{(s)}$ is the contribution that a port adds when it is enabled for operation, before it starts handling traffic.

In DTRA techniques, a primary means for saving power is the switching of individual ports between the enabled and disabled states. Setting the operating rate of an enabled port to a maximum of 10 Mb/s, 100 Mb/s, or 1 Gb/s (and 10 Gb/s when available) is another dimension of dynamic configuration that DTRA techniques can exploit, because each rate generally presents a different value of $P_0^{(s)}$ [15].

Power Measurement Testbed

Figure 1 shows a schematic drawing of the laboratory testbed where we execute the power measurements. The power meter station (PM) consists of the power meter proper and the auxiliary data logging software that runs on a connected laptop. The power meter is an Extech Instruments 380801 true root mean square (RMS) single-phase power analyzer, placed between the power supply (whether AC or DC) and the SUT. The meter's resolution is 0.1 W for readings up to 200 W and 1 W for readings between 200 W and 2 kW. The data logging laptop acquires power samples at one second intervals over the serial port of the power meter. Some SUTs require a 48 V DC power supply, which we obtain from a Xantrex Technology XKW 1 kW module. We do not include the power consumption of the DC power supply module in our power measurements. For the other SUTs the power supply path bypasses the DC module.

As traffic endpoints (TEs) for the generation and termination of test traffic, we use two personal computers (PCs) with 1GbE network cards (PC TE in Fig. 1). We rely on the *iperf* utility for configuration and operation of the traffic sources and sinks used in traffic exchanges with RJ-45 SUT ports at rates up to 1 Gb/s. All traffic is constant-bit-rate UDP, with one source/sink per computer (in each experiment we need the traffic load to hold steady while we collect the power measurements; with TCP traffic the load would likely oscillate). We load the 10GbE ports (in SUTs that feature them) with two 10GBASE-LW/LR 10GbE XFP modules. For modular 1GbE ports we have twenty-four 1000BASE-T SFP modules that can be configured to work at 10/100/1000 Mb/s. We have matching ports on the traffic endpoints for the 1GbE ports but not for the 10GbE ports. (In [15] we used a Spirent SmartBit SMB-200 traffic generator to duplicate the *iperf* results over RJ-45 ports and profile one more type of optical SFP (BASE-SX). In this article we show that reliable power profiles can be obtained with much leaner tools for traffic generation.)

We obtain power profiles for the following five SUTs, manufactured by multiple vendors:

- ES1. Ethernet switch in fixed system configuration with an integrated control and switch module (no slots for plug-in cards), twenty-four 1GbE ports (SFP), two 10GbE ports (XFP), and AC power supply.
- ES2. Ethernet switch with twenty-four integrated 1GbE ports (RJ-45), four of which are dual-mode ports that also offer the alternative of loading an SFP module, two 10GbE ports (XFP), and AC power supply. The sum of the port capacities and functional capabilities are the same for ES1 and

ES2. One important difference is that the 1GbE ports are integrated in ES2 and modular in ES1.

- IR1. Edge/aggregation router in a fixed system configuration with integrated control and switch module, twenty 1GbE ports (SFP), six 10GbE ports (XFP), and AC power supply.
- IR2. Aggregation router in a fixed system configuration with integrated control and switch module, six 10/100 Mb/s Ethernet ports (RJ-45), two 1GbE ports (SFP), and DC power supply.
- IR3. Aggregation router in a modular system configuration with 8-slot chassis. In the configuration used in our experiments, the chassis is populated with one fan card, two control and switch module cards, and two eight-port Ethernet adapter cards (EACs). Each EAC includes six 10/100 Mb/s Ethernet ports (RJ-45) and two 1GbE ports (SFP). IR3 also works with a DC power supply.

We leave out vendor and model names for the five systems because the validation of our methodology for power profiling does not depend on the reproduction of our tests. Instead, the methodology can be applied to any switch or router, and then verified for accuracy with tests that involve more traffic endpoints and a better power meter, like those described in [11, 12].

Measurement Methodology

Our power measurements aim at obtaining accurate estimates for the six parameters of the simplified linear model: the fixed chassis power C_0; the fixed line-card power L_0; the fixed hardware port power $P_0^{(h)}$; the fixed software port power $P_0^{(s)}$; the port bit-rate sensitivity P_b; and the port packet-size sensitivity P_r. To optimize the accuracy of the estimates in spite of the coarse resolution of the power meter (0.1 W), in each experiment we must maximize the number of system components that operate in identical conditions, compatibly with the limitations of our testbed. The limitations are particularly relevant when the port parameter estimation requires the presence of traffic, because we only have two PCs for traffic generation.

In every experiment the system and traffic configuration remain unchanged during the entire measurement period. We collect power samples for 600 s at 1 s intervals after an initial warm-up period of at least 60 s, then we average the 600 power samples (60 s always proves more than sufficient to stabilize the power traces).

The specification of the system configurations for estimation of the fixed parameters is straightforward. For the chassis power C_0, we remove all line cards (if any), leave all modular ports empty, and disable all integrated ports. For the fixed line-card power L_0 in IR3, we load all the line cards available and leave all modular ports empty and all integrated ports disabled. For estimation of the fixed hardware port power $P_0^{(h)}$ (modular ports only), we load all the SFPs of the same type that we have available and leave all ports disabled. For the fixed software port power $P_0^{(s)}$, we enable all the ports of the same type that are available on the SUT.

The procedures for estimation of the port bit-rate sensitivity P_b and port packet-size sensitivity P_r are more elaborate. All measurements with RJ-45 and 1GbE SFP ports involve up to twenty-four ports; the measurements with 10GbE XFPs never involve more than two ports. In RJ-45 and 1GbE SFP measurements with packet traffic, the traffic enters and exits the SUT through the same type of port. In 10GbE XFP measurements, the traffic enters and exits the SUT through 1GbE ports that also contribute to the power measurements: we must subtract the estimated contribution of the ports facing the traffic generator and sink from the total system power before we can estimate the power consumed by the 10GbE XFP ports.

To ensure that every port in the system actually contributes to the measured power, we use loop-back cables to join all the pairs that we can form with ports that are not directly attached to a generator. Within each loop-back pair, one port transmits traffic to the loop-back cable and the other port receives it. We enable the Spanning Tree Protocol (STP) to prevent the volume of the injected broadcast traffic from exploding (with STP each receiving port internally forwards traffic to only one transmitting port). Since every port handles saturation-level traffic in one direction and no traffic in the other, relying on the assumption of linearity of the power consumption model, we use $\beta = 0.5$ to derive the value of P_b from the power meter readings.

For traffic measurements with the 10GbE XFPs we keep STP disabled: the replication of broadcast packets that occurs at each receiving port is necessary for expanding the traffic volume from the 1 Gb/s rate supplied by the traffic generator to the 10 Gb/s rate that each 10GbE port can sustain. Since every port that is enabled receives and transmits traffic at full capacity, we use $\beta = 1.0$ in the estimation of the bit-rate sensitivity.

To ensure that every port in the system actually contributes to the measured power, we use loop-back cables to join all the pairs that we can form with ports that are not directly attached to a generator. Within each loop-back pair, one port transmits traffic to the loop-back cable and the other port receives it.

Experimental Results

We present results from our experiments that gauge the compatibility of existing equipment with DTRA techniques and underscore the need for PTRA support in future system designs.

Power Profiles

For each SUT, Table 1 lists the sum of the port capacities (possibly larger than the actual switching capacity, not reported here because it is only partially characterized in the majority of our SUT data sheets) along with estimated values for five of the six parameters of our linear model, in the specific case where the SUT is loaded with 1GbE ports (RJ-45 in ES2, SFP in all other SUTs) enabled for operation at 1 Gb/s. The missing parameter is the (port) packet-size sensitivity, whose values are practically impossible to distinguish from zero in the system configurations used in the experiments of Table 1. We observe non-negligible values of the parameter only in the case of integrated RJ-45 ports in IR2 and IR3 probably because in these two routers, unlike IR1, packets do not bypass the header processing stages for more complex functions than the basic layer-2 switching that they exclu-

SUT (Sum of port capacities)	C_0[W] Chassis, idle	L_0[W] Line card, idle	$P_0^{(h)}$[W] Port, fixed hardware	$P_0^{(s)}$[W] Port, fixed software	P_b[W] Port, bit rate sensitivity
ES1 (44 Gb/s)	32.4	N/A	0.3	1.1	0.3
ES2 (44 Gb/s)	35.0	N/A	N/A	1.0	0.1
IR1 (80 Gb/s)	216	N/A	0.2	1.0	1.1
IR2 (2.6 Gb/s)	40.4	N/A	0.2	1.0	0.8
IR3 (15.6 Gb/s)	54.8	14.5	0.2	1.0	0.8

Table 1. Parameters of linear model (1GbE ports configured for 1 Gb/s operation).

SUT	$P_0^{(s)}$[W] Port, fixed software	P_b[W] Port, bit-rate sensitivity	P_r[W] Port, packet-size sensitivity
IR2	0.3	0.1	10
IR3	0.3	0.1	9

Table 2. Port parameters (RJ-45 ports in IR2 and IR3 configured for 100 Mb/s operation).

SUT	$P_0^{(h)}$[W] Port, fixed hardware	$P_0^{(s)}$[W] Port, fixed software	P_b[W] Port, bit-rate sensitivity
ES1	1.2	1.8	1.6
ES2	0.9	2.0	0.5
IR1	0.2	1.0	2.9

Table 3. Port parameters (10GbE XFP ports configured for 10 Gb/s operation).

sively require in our tests (see Table 2 for the values measured with RJ-45 ports configured at 100 Mb/s). Table 3 lists the parameters of the 10GbE XFP ports, which are only available in ES1, ES2, and IR1.

The results in Table 1 indicate that the fixed software port power is by far the dominant port term in the two Ethernet switches. The traffic-sensitive terms gain relevance in the IP routers, consistently with the increased variety and intensity of the packet processing functions in those systems. Table 3 shows similar trends for the 10GbE ports, although the traffic-sensitive terms are generally heavier than with 1GbE ports. Table 1 shows that the idle chassis power is much higher in IR1 than in all other SUTs, because IR1 is the only system in the set that combines a large sum of the port capacities (80 Gb/s) with the complex packet processing functions of a router in a non-modular architecture.

We define the margin for saving power with DTRA techniques as the entire portion of the total power consumption of a system that is associated with components that DTRA can control. This is a hard upper bound on the amount of power that DTRA can save. Network topology and traffic demands determine the tightness of the bound in practical applications.

To quantify the DTRA margin, we measure the relative weights of the line-card and port terms within the overall power consumption of each system. Figure 2 shows the breakdown of the total power consumption that our linear model yields for the five SUTs when all ports are enabled and saturated. In ES1, ES2, and IR1 we configure the 1GbE SFP ports as in Table I and the 10GbE XFP ports as in Table 3. In IR2 and IR3 the configurations are those of Table 1 (1GbE SFPs at 1 Gb/s) and Table 2 (RJ-45 at 100 Mb/s). We normalize the power levels in each column to the total power consumption of the respective system. We obtain the contributions of the port power terms by multiplying the maximum number of ports configurable for each type by the respective per-port values. With IR2 and IR3 the sum of the port capacities in this maximum configuration far exceeds the actual switching capacity of the system, with the effect of producing overestimated values for the traffic-sensitive power contributions. We would obtain more accurate estimates if we had a larger number of traffic generator ports to pair with the system ports in the power measurement experiments (up to 48 ports with IR3). Still, even if overestimated and maximized by the assumption of minimum-length Ethernet frames (64 B), the traffic-sensitive shares of the total power remain marginal in IR2 and IR3, causing no qualitative impact on the interpretation of the results.

In Fig. 2 the traffic-sensitive terms range between 5 percent and 21 percent across the five systems. If we also consider that the two highest values, in IR2 and IR3, are certainly overestimated, the maximum traffic power share is likely well below 15 percent. We can conclude that current designs are far from exhibiting the type of rate-proportional power consumption behavior that rate adaptation techniques aim at establishing at the system level. While the indication is disappointing in terms of overall energy efficiency, it signals that DTRA techniques have a clear window of opportunity in the short term for bringing along important power savings through relatively simple signaling extensions and software modifications applied to existing hardware platforms. If we compute the DTRA margin as the sum of the fixed-software and traffic-sensitive port power terms, that is without including the fixed-hardware port power and the line-card power, we see that it ranges between 46 percent and 49 percent in the two Ethernet switches and between clearly lower values (23 percent and 32 percent) in the three IP routers. The potential for DTRA savings increases substantially in modular systems where individual line cards can be switched on and off (upwards of 41 percent in IR3), and even more if DTRA can control the operating state of the entire chassis (although networks where an entire node can be powered off may not be easy to find [2]).

To ensure substantial power savings irrespective of the network topology, PTRA capabilities must be pervasively deployed in future generations of hardware platforms. Design challenges and performance properties are well understood for PTRA techniques in linear data-path devices with one input and one output [5]. The same is not true for devices with multiple inputs and outputs like the switch fabric in the system chassis. The challenge for those devices is to achieve direct proportionality between power consumption and aggregate switching throughput irrespective of the traffic load distribution across interfaces. Since the chassis contribution to total power usage is always large (between 26 percent and 76 percent in the SUTs of our testbed), future research efforts should target the identification of viable PTRA solutions for multi-interface devices.

Conclusions

The compilation of accurate power profiles is a useful tool in the process of planning for the design upgrades that can enable better energy efficiency in future generations of network systems. Building upon a linear model for power profiling that is well suited for supporting the operation of rate adaptation frameworks at multiple timescales, we defined a measurement methodology that is ideal for the occasional power profiler, as it yields reasonable estimates for the key parameters of the model even with rather minimal testing equipment (two PCs and an inline power meter). Power profiles obtained with the methodology show that the establishment of true proportionality between power consumption and traffic workload will mostly depend on the degree of deployment of packet-timescale rate adaptation in future hardware platforms. When systems with those capabilities become available, we expect our methodology to show that the fixed terms of the linear model yield prominence to the terms that represent sensitivity to traffic loads.

Acknowledgments

This work was completed with the support of the U.S. Department of Energy (DOE), award no. DE-EE0002887, and used resources of the Energy Sciences Network (ESnet) Advanced Network Initiative (ANI) Testbed, supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-05CH11231. Any opinions, findings, conclusions, and recommendations expressed herein are those of the authors and do not necessarily reflect the views of the DOE.

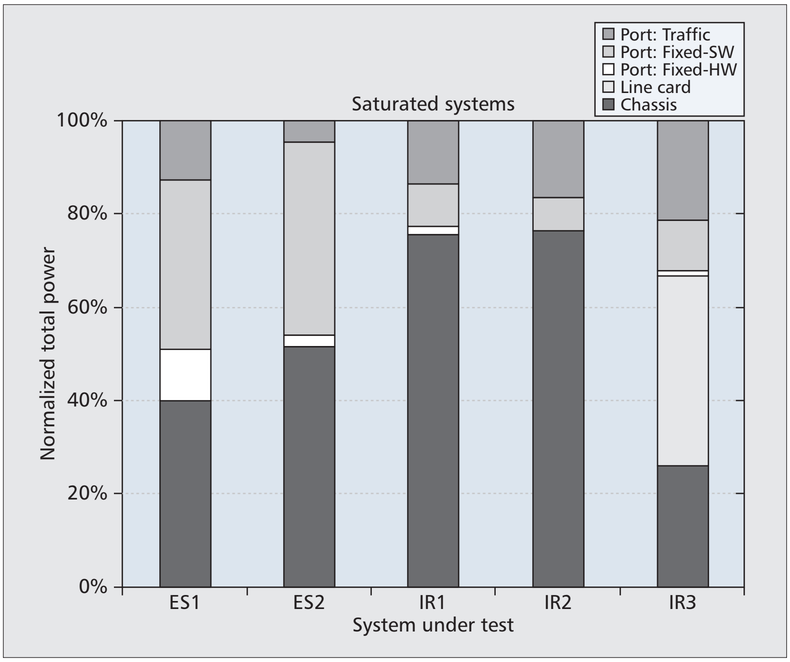


Figure 2. Estimated breakdown of system power when all ports in the system are saturated; 100 percent normalized power is 82 W (ES1), 68 W (ES2), 287 W (IR1), 59 W (IR2), and 234 W (IR3).

References

[1] S. Antonakopoulos, S. Fortune, and L. Zhang, "Power-Aware Routing with Rate-Adaptive Network Elements," *Proc. IEEE Global Commun. Conf. Workshops (GC Workshops '10), 3rd Int'l. Workshop on Green Commun. (GreenComm '10)*, Miami, FL, 2010, pp. 1428–32.

[2] A. P. Bianzino *et al.*, "Energy-Aware Routing: A Reality Check," *Proc. IEEE Global Commun. Conf. Workshops (GC Workshops '10), 3rd Int'l. Workshop on Green Commun. (GreenComm '10)*, Miami, FL, 2010, pp. 1422–27.

[3] P. Mahadevan *et al.*, "Energy Aware Network Operations," *Proc. 28th IEEE Int'l. Conf. Comput. Commun. Workshops (INFOCOM '09), 12th IEEE Global Internet Symp. (GI '09)*, Rio de Janeiro, Brasil, 2009.

[4] A. Francini, "Selection of a Rate Adaptation Scheme for Network Hardware," *Proc. 31st IEEE Internat. Conf. on Comput. Commun.: Mini-Conf. (INFOCOM '12)*, Orlando, FL, 2012, pp. 2831–35.

[5] A. Francini and D. Stiliadis, "Performance Bounds of Rate-Adaptation Schemes for Energy-Efficient Routers," *Proc. Int'l. Conf. High Perform. Switching and Routing (HPSR '10)*, Dallas, TX, 2010, pp. 175–82.

[6] S. Nedevschi *et al.*, "Reducing Network Energy Consumption via Sleeping and Rate-Adaptation," *Proc. 5th USENIX Symp. on Networked Syst. Design and Implementation (NSDI '08)*, San Francisco, CA, 2008, pp. 323–36.

[7] L. Benini, P. Siegel, and G. De Micheli, "Saving Power by Synthesizing Gated Clocks for Sequential Circuits," *IEEE Design and Test of Computers*, 1994, vol. 11, pp. 32–41.

[8] J. Chabarek *et al.*, "Power Awareness in Network Design and Routing," *Proc. 27th IEEE Conf. on Comput. Commun. (INFOCOM '08)*, Phoenix, AZ, 2008, pp. 1130–38.

[9] P. Mahadevan *et al.*, "A Power Benchmarking Framework for Network Devices," *Proc. 8th Internat. IFIP-TC 6 Networking Conf. (Networking '09)*, Aachen, Germany, 2009, LNCS vol. 5550, pp. 795–808.

[10] O. Tamm, C. Hermsmeyer, and A. M. Rush, "Eco-Sustainable System and Network Architectures for Future Transport Networks," *Bell Labs Tech. J.*, vol. 14, no. 4 (2010), 311–28.

[11] V. Sivaraman *et al.*, "Profiling Per-Packet and Per-Byte Energy Consumption in the NetFPGA Gigabit Router," *Proc. 30th IEEE Internat. Conf. on Comput. Commun. Workshops (INFOCOM '11), Workshop on Green Commun. and Networking (GCN '11)*, Shanghai, China, 2011, pp. 331–36.

[12] A. Vishwanath *et al*., "Modeling Energy Consumption in High-Capacity Routers and Switches," *IEEE JSAC*, vol. 32, no. 8, Aug. 2014, pp.1524–32.
[13] Alliance for Telecommunications Industry Solutions, "Energy Efficiency for Telecommunication Equipment: Methodology for Measurement and Reporting General Requirements," ATIS-0600015.2009, Feb. 2009.
[14] ECR Initiative, "Network and Telecom Equipment Energy and Performance Assessment: Metrics, Test Procedure and Measurement Methodology," Draft 3.0.1, Dec. 2010, <http://www.ecrinitiative.org/pdfs/ECR_3_0_1.pdf>.
[15] M. Ricca *et al*., "An Assessment of Power-Load Proportionality in Network Systems," Proc. 3rd IFIP Conf. on Sustainable Internet and ICT for Sustainability (SustainIT '13), Palermo, Italy, 2013.

Biographies

Andrea Francini (andrea.francini@alcatel-lucent.com) is a member of technical staff at Alcatel-Lucent Bell Laboratories. He joined Bell Labs in 1997, a few months before receiving his Ph.D. in electrical engineering and communications from the Politecnico di Torino, Italy. Andrea is a traffic management expert with hands-on experience designing quality-of-service architectures for packet switching chipsets, multi-service systems, and broadband access platforms. His current research focuses on rate adaptation for energy efficiency, congestion control for wireless networks, and TCP and network optimizations for adaptive-streaming video.

Steven Fortune (steven.fortune@alcatel-lucent.com) is a distinguished member of technical staff and formerly a technical manager at Alcatel-Lucent Bell Laboratories. He received his Ph.D. from Cornell University in 1979 and joined Bell Labs in 1983. Steven has published extensively in the area of computer science algorithms, principally computational geometry. He has developed a wide variety of practical algorithms with application to the fields of numerical analysis, radio propagation prediction in complex environments, and network design tools for DWDM optical networks.

Thierry Klein (thierry.klein@alcatel-lucent.com) is the network energy research program leader at Alcatel-Lucent Bell Laboratories. He received his Ph.D. in electrical engineering and computer science from MIT, and upon graduation joined Bell Labs in 2001. Thierry serves on the UNFCCC Momentum for Change Advisory Panel and is chairman of the Technical Committee of GreenTouch, a global consortium dedicated to improving network energy efficiency by a factor of 1000. He was voted "Technologist of the Year" at the 2010 Total Telecom World Vendor Awards.

Marco Ricca (marco.ricca@polito.it) is a research assistant with the Telecommunication Networks Group (TNG) in the Department of Electronics and Telecommunications (DET) of the Politecnico di Torino, Italy. From Politecnico he received the Italian Laurea degree in computer and communication networks engineering in 2008, and a Ph.D. in electrical engineering and communications in 2012. In 2011 Marco was a visiting researcher at Alcatel-Lucent Bell Laboratories in Murray Hill, New Jersey. His primary research interest is in the area of energy-aware control algorithms for packet networks.

Cost-Aware Green Cellular Networks with Energy and Communication Cooperation

Jie Xu, Lingjie Duan, and Rui Zhang

Jie Xu and Lingjie Duan are with Singapore University of Technology and Design.

Rui Zhang is with the ECE Department of the National University of Singapore.

This work was supported in part by the National University of Singapore under Grant R-263-000-A48-112 and the SUTD-MIT International Design Center Grant under Project IDSF1200106OH.

ABSTRACT

The energy cost of cellular networks is ever increasing to match the surge of wireless data traffic, and savings in this cost is important to reduce the operational expenditure of wireless operators in the future. Recent advancements in renewable energy integration and two-way energy flow in smart grid provide potential new solutions to save cost. However, they also impose challenges, especially on how to use the stochastically and spatially distributed renewable energy harvested at cellular BSs to reliably supply time- and space-varying wireless traffic over cellular networks. To overcome these challenges, in this article we present three approaches, *energy cooperation*, *communication cooperation*, and *joint energy and communication cooperation*, in which different BSs bidirectionally trade or share energy via the aggregator in smart grid, and/or share wireless resources and shift loads with each other to reduce the total energy cost.

INTRODUCTION

To meet the dramatically increasing mobile data traffic, recently cellular operators are deploying more and more base stations (BSs), and their daily energy cost amounts to a large portion of operational expenditure (OPEX) [1]. Many cellular operators want to reduce their energy costs by employing new cost-saving solutions [1–4], which in general manage either the energy supply or the communication demand of cellular networks.

On the supply side, one commonly adopted solution is to use energy harvesting devices (e.g., solar panels and wind turbines) at cellular BSs, which can harvest cheap and clean renewable energy to reduce or even substitute for the energy purchased from the grid [3]. However, since renewable energy is often randomly distributed in both time and space, different BSs are hard pressed to use only their individually harvested energy to power their operations. As a result, the power grid is still needed to provide reliable energy to BSs. Besides serving as a reliable energy supply, the power grid also provides new opportunities for BSs' cost saving with its ongoing paradigm shift from traditional to smart grid. Unlike traditional grid, which uses one-way energy flow to deliver power from central generators to electricity users, smart grid deploys smart meters at end users to enable both two-way information and energy flows between the grid and end users [5, 6]. The two-way energy flow in smart grid motivates a new idea of *energy cooperation* in cellular networks, as we discuss later in this article, which allows the BSs to trade and share their unevenly harvested renewable energy through the smart grid to support non-uniform wireless traffic in a cost-effective way.

On the demand side, various techniques have been proposed in cellular networks across different layers of communication protocols for reducing energy consumption [1]. Among them, communication cooperation (e.g., traffic loading [7], spectrum sharing [8] and coordinated multipoint [CoMP] [9]) is particularly appealing, allowing BSs to share wireless resources and shift traffic loads with each other for energy saving. However, the introduction of renewable energy at BSs imposes new challenges on the existing communication cooperation design: the conventional energy saving design may not be cost effective any longer. This is due to the fact that renewable energy, although unreliable in supplying, is in general much cheaper than the energy purchased from the grid, so BSs should maximally use it to save cost, whereas under the energy-saving design the harvested renewable energy at BSs may not be efficiently utilized when serving time- and space-varying wireless traffic. To overcome this problem, it is desirable to design new *cost-aware communication cooperation* approaches taking into account the cost differences between renewable and conventional energy.

In this article, we first overview the recent advances in energy cooperation and cost-aware communication cooperation. Figure 1 illustrates the general energy and communication cooperation concept for cellular networks at both the energy supply layer and the communication demand layer, respectively. Then we propose a new *joint energy and communication cooperation*

 0163-6804/15/$25.00 © 2015 IEEE

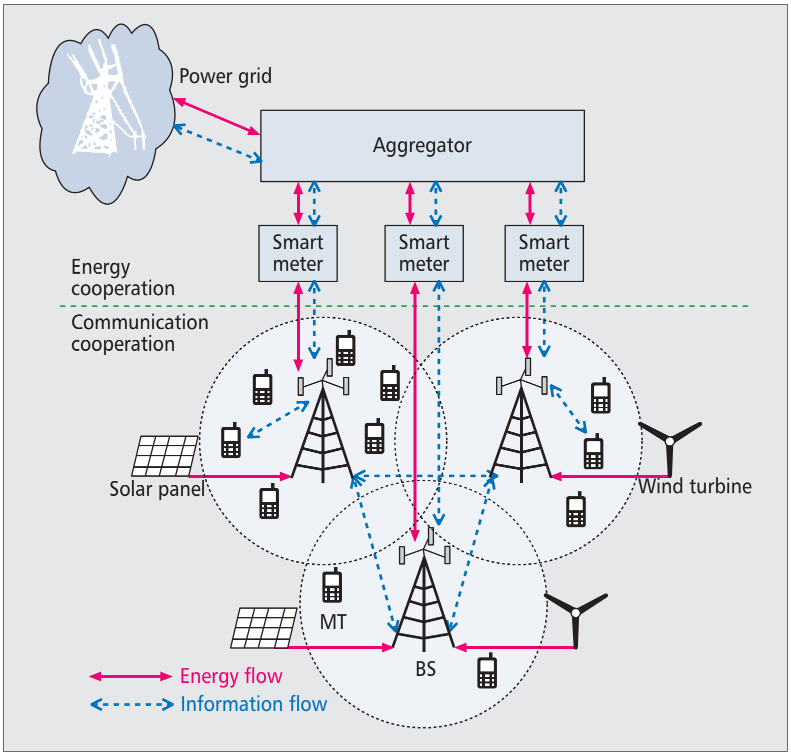


Figure 1. A general model of cellular networks with energy and communication cooperation among BSs.

to exploit both benefits. Specifically, we present the following three approaches.

Approach I — energy cooperation on the supply side: Cellular systems or BSs use the two-way energy flow in smart grid to trade or share renewable energy by taking the energy demands for communications as given.

Approach II — communication cooperation on the demand side: Cellular systems or BSs perform cost-aware communication cooperation to share wireless resources and reshape wireless load over space and time by taking the energy supply (renewable and/or conventional) as given.

Approach III — joint energy and communication cooperation on both sides: Cellular systems or BSs jointly cooperate on both the supply and demand sides to maximally reduce their total energy cost.

In the rest of this article, we first introduce the energy supply and demand models of cellular systems. Then we present the latest energy, communication, and joint cooperation approaches, respectively. Finally, we point out several future research directions and conclude this article. Due to the page limits, we cannot include everything here; more details can be found in http://arxiv.org/abs/1411.4139.

Energy Supply and Demand of Cellular Systems

In this section, we introduce the energy supply and demand models for cellular systems. For notational convenience, in this article we focus on one particular time slot and normalize the length of time slot into unity. Thus, we use the terms energy and power interchangeably throughout this article.

We consider a single cellular system with $N > 1$ BSs, in which each BS i is connected to the smart grid and also deployed with a renewable energy harvesting device with rate $E_i \geq 0$, $i = 1, \ldots, N$. The value of E_i at a given time instance depends on the type of renewable energy source (e.g., solar or wind), the harvesting capacity of the device (e.g., size of the solar panel), and the weather conditions at that location. As shown in the top sub-figure of Fig. 2, E_is are generally different among BSs at different locations.

On the demand side, the power consumption of each cellular BS i, denoted by $Q_i \geq 0$, is composed of two parts: the dynamic power consumption related to the transmission and reception of wireless signals for serving the mobile terminals (MTs), and the constant power consumption (e.g., at the circuits and air conditioners) for maintaining necessary operations. In reality, the value of Q_i varies according to the traffic load over the service coverage area of BS i. Due to MTs' mobility across cells and their time-varying service requests, the traffic loads (and thus the Q_is) are different among BSs and change over time, as shown in the middle sub-figure of Fig. 2.

By combining the supply and demand sides, we denote the *net load* at BS i as $\delta_i = Q_i - E_i$, where $\delta_i > 0$ shows the deficit status of renewable energy and $\delta_i < 0$ indicates the energy surplus status. Since Q_i's and E_i's are usually independent (Fig. 2), it is likely that some BSs are short of renewable energy to match demand (i.e., $\delta_i > 0$), while the other BSs have adequate renewable energy (i.e., $\delta_i < 0$). Such geographical diversity in net load requires some BSs to purchase energy from the grid (e.g., a δ_i amount of energy purchase for BS i with $\delta_i > 0$) but other BSs to waste extra renewable energy (i.e., a $|\delta_j|$ amount of energy waste for BS j with $\delta_j < 0$). Overall, the total amount of energy purchased from the grid by all N BSs is the total renewable energy deficit, denoted by $\Delta_+ \triangleq \sum_{i=1}^{N} [\delta_i]^+ \geq 0$ with $[x]^+ \triangleq \max(x, 0)$, while the total wasted renewable energy amount by them is the total renewable energy surplus, given by $\Delta_- \triangleq -\sum_{j=1}^{N} [\delta_j]^- \geq 0$ with $[x]^- \triangleq \min(x, 0)$. By denoting the price for BSs to purchase energy from the grid as $\pi > 0$, the total energy cost of the cellular system is $C_1 = \pi\Delta_+$, which is independent of Δ_-. This fact motivates us to use the wasted renewable energy surplus (Δ_-) to compensate for the deficit (Δ_+) for cost saving. To this end, we implement the energy and communication cooperation on the supply and demand sides, respectively, to reschedule and balance E_is and Q_is.

Energy Cooperation

Energy cooperation is a cost saving approach on the supply side, in which the cellular BSs are allowed to employ two-way energy trading or sharing to better utilize their otherwise wasted renewable energy surplus (Δ_-). Especially since it is too complex for the grid to directly control a large number of BSs, the energy trading and sharing in cellular networks should be enabled by using aggregators [10] (upper energy cooperation layer in Fig. 1). With aggregators, we can cluster BSs into a finite number of groups, and an aggregator can serve as an intermediary party to control each group of BSs for the grid, thus helping realize the two-way energy flow between the grid and BS groups.

Aggregator-Assisted Energy Trading

Aggregator-assisted energy trading is an energy cooperation scheme in which the aggregator performs two-way energy trading with the BSs by deciding buying and selling prices. In this scheme, the BSs adequate in renewable energy can sell their extra energy to the aggregator, from which the selling revenue can be gained to compensate the total energy cost; at the same time, the other BSs short of renewable energy can obtain such cheap energy from the aggregator at a lower price than the regular price π from the grid directly. As the coordinator in this trading market, the aggregator can also obtain some revenue by properly deciding the energy selling and buying prices.

Let $\pi_{buy} > 0$ and $\pi_{sell} > 0$ denote the unit prices for each BS to buy and sell energy from and to the aggregator, respectively. Here, $\pi_{sell} < \pi_{buy}$ holds to avoid the trivial case where a BS can benefit by reselling its bought energy from the aggregator; and $\pi_{buy} < \pi$ is also true, since otherwise all BSs short of energy will buy cheaper energy from the grid directly. With the two-way energy trading, the BSs adequate in renewable energy will sell their total Δ_- amount of energy surplus to the aggregator at the price π_{sell}, and accordingly an energy quota is set by the aggregator as Δ_-. The BSs short of renewable energy will first purchase a $\min(\Delta_+, \Delta_-)$ amount of cheap energy from aggregator at the price π_{buy} (with the quota limitation of Δ_-) to maximally use this resource, and (if not enough) will buy a $\Delta_+ - \min(\Delta_+, \Delta_-)$ amount from the grid at the price π. Depending on the relationship between Δ_+ and Δ_-, the total cost of all the N BSs, denoted by C_2, is equal to $\pi_{buy}\Delta_+ - \pi_{sell}\Delta_-$ in the case of $\Delta_+ \leq \Delta_-$, or $\pi_{buy}\Delta_- + \pi(\Delta_+ - \Delta_-) - \pi_{sell}\Delta_-$ otherwise. Note that C_2 can even be negative when Δ_- is sufficiently larger than Δ_+ such that $\pi_{buy}\Delta_+ < \pi_{sell}\Delta_-$. By comparing C_2 to the total energy cost C_1 without energy cooperation, it follows that $C_2 \leq C_1$, that is, the total energy cost is reduced.

Aggregator-Assisted Energy Sharing

Aggregator-assisted energy sharing is another energy cooperation scheme that allows BSs in a BS group to mutually negotiate and share renewable energy by simultaneously injecting and drawing energy to and from the aggregator, respectively. By matching the local renewable energy deficit (positive δ_is) and surplus (negative δ_is) between any two BSs, this scheme helps the group of BSs reduce their aggregate renewable energy deficit. The practical implementation of energy sharing requires this group of BSs to sign a *contract* with the aggregator at a contract fee that motivates the aggregator to support the energy sharing.

Specifically, suppose that BS i wants to transfer an $e_{ij} \geq 0$ amount of energy to BS j, $i \neq j$. This is accomplished at an appointed time by BS i injecting an e_{ij} amount of energy into the aggregator, and at the same time BS j drawing the same e_{ij} amount from the aggregator. Thanks to the mutual sharing of e_{ij}s among the N BSs, the total energy deficit Δ_+ and surplus Δ_- can be effectively matched. When $\Delta_+ \leq \Delta_-$, the N BSs can maintain their operation without purchasing any energy from the grid; otherwise, a total $\Delta_+ - \Delta_-$ amount of energy should be purchased from the grid at price π. By denoting the contract fee to the aggregator as $\overline{C}$, the total cost of all N BSs, given by C_3, is equal to $\overline{C}$ in the case of $\Delta_+ \leq \Delta_-$, or $\pi(\Delta_+ - \Delta_-) + \overline{C}$ otherwise. By comparing C_3 with the total

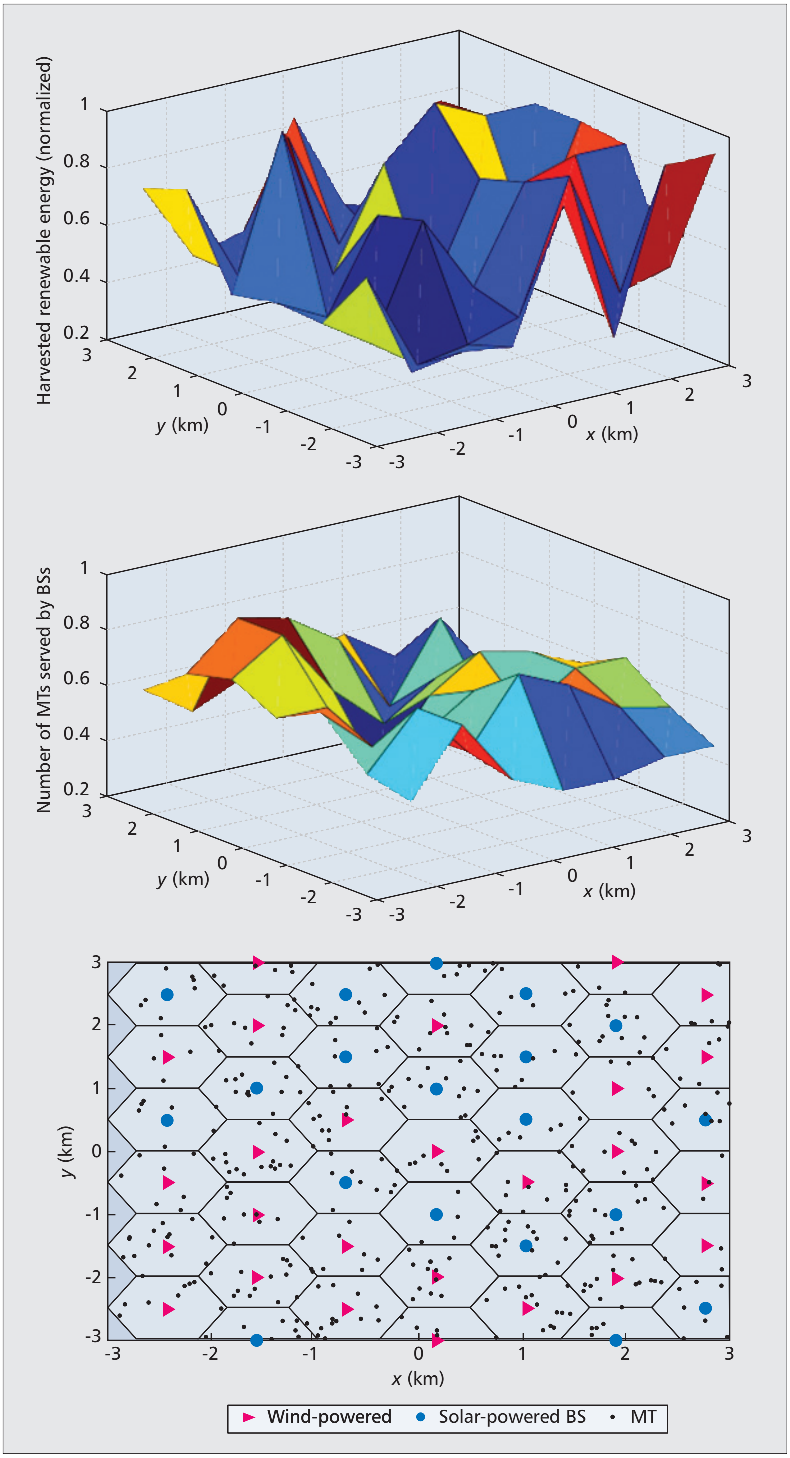


Figure 2. An example of a cellular system with BSs having spatially distributed traffic load and harvested energy at a given time instance. It is assumed that the (normalized) energy harvesting capacity of all solar-powered BSs is 1, and that of all wind-powered BSs is 0.5.

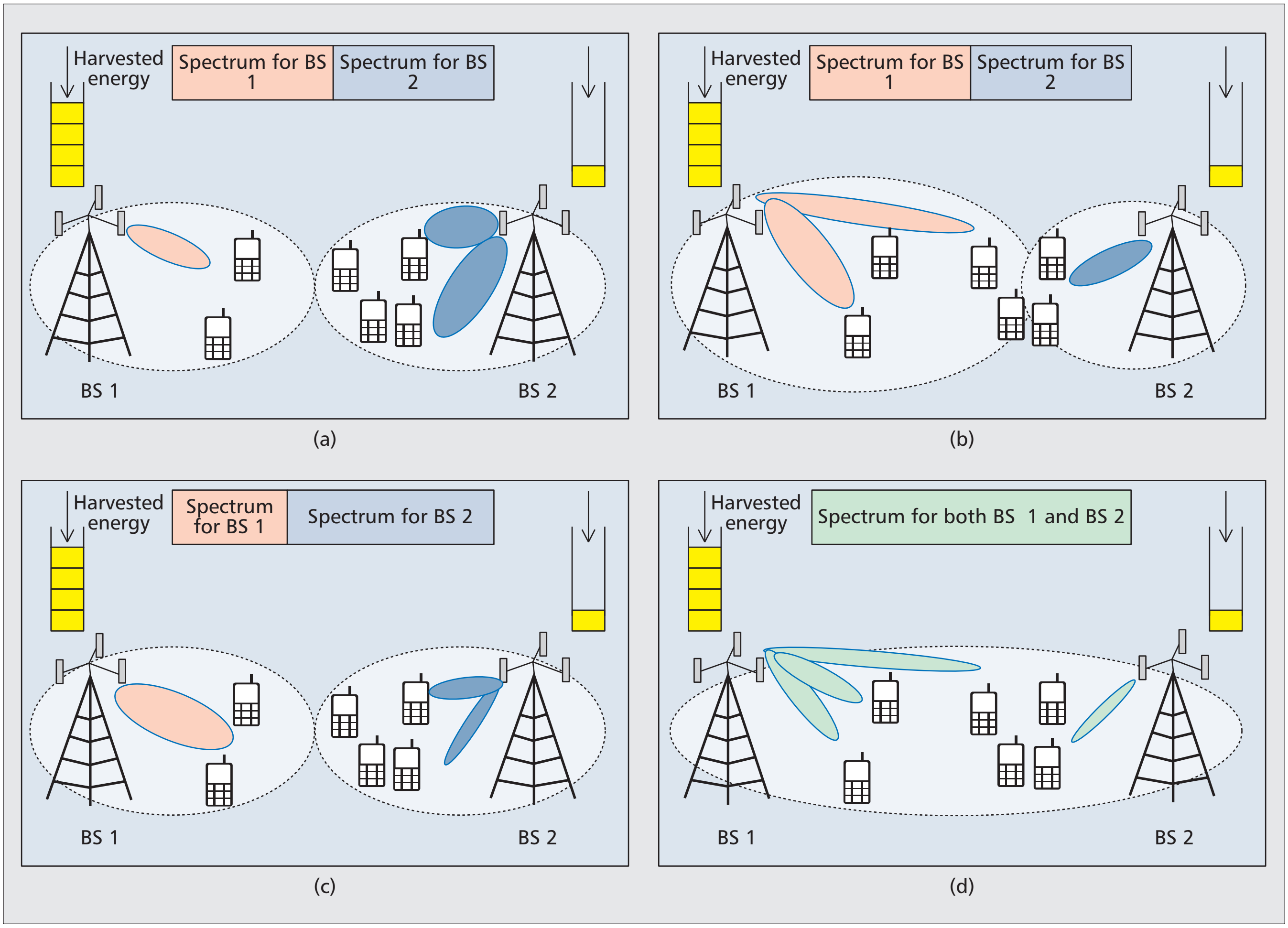


Figure 3. An example of different communication cooperation designs in a simple cellular network with two BSs: a) conventional design; b) cost-aware traffic offloading; c) cost-aware spectrum sharing; d) cost-aware CoMP.

energy cost C_1 without energy cooperation, it generally follows that $C_3 \leq C_1$, that is, the total energy cost is reduced, as long as $\overline{C}$ is sufficiently small.

Communication Cooperation

Communication cooperation refers to a cost saving approach on the demand side that exploits the broadcast nature of wireless channels and uses wireless resource sharing to reshape BSs' wireless load and energy consumption. Different from conventional communication cooperation (e.g., [7–9]) aiming to maximize data throughput or minimize energy consumption, the communication cooperation of our interest here seeks to minimize the total energy cost by optimally utilizing both cheap renewable energy and reliable on-grid energy. In so-called cost-aware communication cooperation, the rescheduling of BSs' traffic load and energy consumption should follow their given renewable energy supply such that the renewable energy can be maximally used to support the quality of service (QoS) requirements of the MTs, and the on-grid energy purchase is thus minimized.

In this section, we discuss three different cost-aware communication cooperation schemes: traffic offloading [7], spectrum sharing [8], and CoMP [9]. For the purpose of illustration, we consider a simple cellular system setup with two BSs as shown in Fig. 3a, in which BS 1 has sufficient harvested renewable energy and light traffic load (serving 2 MTs), with net load $\delta_1 < 0$; while BS 2 has insufficient renewable energy and heavy traffic load (4 MTs), leading to net load $\delta_2 > 0$.

Cost-Aware Traffic Offloading

Traffic offloading is traditionally designed to shift the traffic load (or served MTs) of heavily loaded BSs to lightly loaded ones for the purposes of avoiding the traffic congestion and improving the QoS of the MTs. Differently, the cost-aware traffic offloading here focuses on the new issue of energy cost reduction; that is, BSs short of renewable energy can offload their MTs to neighboring BSs with abundant renewable energy (even if they have more or similar traffic loads), thus reducing all energy drawn from the grid to save cost. As shown in the example of Fig. 3b, it is cost effective for BS 2 to offload 2 MTs (at its cell edge) to BS 1, such that the renewable energy at BS 1 is better utilized.

Cost-Aware Spectrum Sharing

Besides energy, spectrum is another scarce resource in cellular networks, and spectrum sharing has been considered as a solution to improve spectrum utilization efficiency [8]. Different from conventional spectrum sharing, cost-aware spectrum sharing is based on the fact that the energy and spectrum resources can partially substitute for each other to support wireless transmission, and sharing spectrum with a BS short of energy can better save the energy cost of that BS. As shown in the example of Fig. 3c, BS 1 shares part of its available spectrum with BS 2. Under the same QoS requirements of MTs, BS 2 can decrease its transmission power purchased from the grid, while BS 1 uses more renewable energy for its transmission. Hence, the total cost is reduced.

Cost-Aware Coordinated Multipoint

Traditionally, CoMP is considered as a technique to improve the spectral efficiency in cellular networks, by which BSs can implement coordinated baseband signal processing to cooperatively serve multiple MTs over the same time-frequency resources, transforming harmful intercell interference (ICI) into useful information signals [9]. Differently, cost-aware CoMP is motivated by the following observation: since different BSs can cooperatively send information signals to the MTs (in the downlink), their transmission power can be compensated by each other for satisfying the QoS requirements at MTs. Therefore, by adaptively adjusting the BSs' transmit signals, the cost-aware CoMP helps match the BSs' transmission power with their harvested renewable energy, thus minimizing the total energy drawn from the grid to save cost. For example, in Fig. 3d, BS 1, with adequate renewable energy, should use a high transmission power to provide strong wireless signals to the MTs, while BS 2, short of renewable energy, should transmit at a low power level in their CoMP transmission.

Joint Energy and Communication Cooperation

Joint energy and communication cooperation can maximally save cost by applying both energy cooperation on the supply side and communication cooperation on the demand side. To realize the joint operation, the BSs should share the energy information by using the two-way information flow supported by the smart grid (through smart meters), and also exchange the communication information through their backhaul connections (Fig. 1). In general, joint energy and communication cooperation is more complex than energy or communication cooperation only.

As there are two energy cooperation schemes (aggregator-assisted energy trading and energy sharing) and three communication cooperation schemes (traffic offloading, spectrum sharing, and CoMP), there are a total of six combinations for joint cooperation designs. In this section, we focus on three specific schemes: a joint energy and spectrum sharing design, and two joint energy cooperation and CoMP designs. The ideas can similarly be extended to the other three combinations.

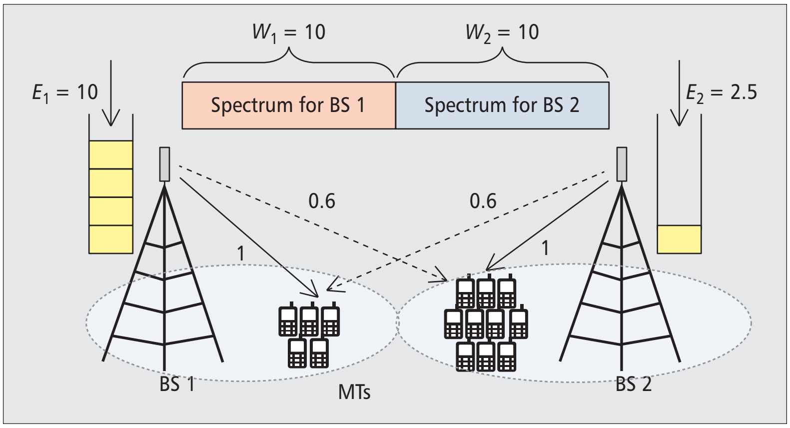


Figure 4. The case study model for comparing energy cooperation, communication cooperation, and joint cooperation.

Joint Energy and Spectrum Sharing

The joint energy and spectrum sharing [11] is a scheme that allows neighboring BSs to share energy and spectrum with each other through the aggregator-assisted energy sharing and the spectrum sharing, in the previous two sections. Accordingly, the BSs take advantage of resource complementarity for cost saving.

Building upon the spectrum sharing in Fig. 3c, [11] considered joint energy and spectrum sharing between two BSs to minimize their total energy cost, while ensuring the QoS requirements for all the MTs. It is shown that at the optimality, it is possible that one BS adequate in both energy and spectrum shares these two resources to the other (in unidirectional cooperation), or one BS exchanges its energy for spectrum with the other (in bidirectional cooperation).

Joint Energy Cooperation and CoMP

In this scheme, different BSs implement CoMP-based transmission/reception to serve one or more MTs over the same time frequency resources, and at the same time perform aggregator-assisted energy trading or aggregator-assisted energy sharing. Based on different types of energy cooperation schemes employed, [12] and [13] studied two different joint energy cooperation and CoMP designs for the downlink transmission.

In particular, when the aggregator-assisted energy trading is implemented among BSs, [12] proposed to jointly optimize the BSs' cooperative transmit beamforming in CoMP based communication and their two-way energy trading with the aggregator, so as to minimize their total energy cost. When the aggregator-assisted energy sharing is adopted, [13] considered to use it to enable a new purely renewable-powered cellular system, in which the BSs do not purchase any energy from the grid to minimize the cost, but use the harvested renewable energy together with the energy sharing to maintain their operations. The results in [12, 13] show that joint energy cooperation and CoMP optimization achieves a significant cost reduction compared to the design that separately optimizes CoMP-based communication and energy cooperation.

	BS 1's renewable energy supply	BS 2 's renewable energy supply	BS 1's energy consumption	BS 2's energy consumption	Total energy cost
Conventional design without energy or communication cooperation	10	2.5	4.14	18.28	15.78
Approach I: energy cooperation via aggregator-assisted energy trading	4.14	8.36	4.14	18.28	10.51
Approach I: energy cooperation via aggregator-assisted energy sharing	4.14	8.36	4.14	18.28	10.03
Approach II: communication cooperation via spectrum sharing	10	2.5	10.00	14.04	11.54
Approach II: communication cooperation via CoMP	10	2.5	10.00	3.75	1.25
Approach III: joint energy and spectrum sharing	5.00	7.50	5.00	15.00	7.60
Approach III: joint aggregator-assisted energy trading and CoMP	6.87	5.62	6.87	5.77	0.46
Approach III: joint aggregator-assisted energy sharing and CoMP	5.47	7.03	5.47	7.03	0.10

Table 1. Energy cost performance comparison.

A Case Study

Now we present a case study to compare the energy cooperation in the third section, the cost-aware spectrum sharing and CoMP in the previous section, and the three joint energy and communication cooperation schemes proposed in this section. Also, we consider the conventional design without energy or communication cooperation as the performance benchmark. For the purpose of illustration, as shown in Fig. 4, we consider the downlink of a cellular system with two single-antenna BSs (i.e., BS 1 and BS 2) each applying orthogonal frequency-division multiple access (OFDMA) to serve $K_1 = 5$ and $K_2 = 15$ single-antenna MTs (denoted by the MT sets $\mathcal{K}_1$ and $\mathcal{K}_2$), respectively. Each BS uses an orthogonal frequency band with the same bandwidth ($W_1 = W_2 = 10$). For simplicity, we randomly generate the channels based on the independent and identically distributed Rayleigh fading with the average channel powers from each BS to its own associated MTs (i.e., from BS 1 to any MT in $\mathcal{K}_1$ and from BS 2 to any MT in $\mathcal{K}_2$) being 1, and those from each BS to the other BS's associated MTs (i.e., from BS 1 to any MT in $\mathcal{K}_2$ and from BS 2 to any MT in $\mathcal{K}_1$) being 0.6. We set the noise power spectral density at each MT to be 1, and the QoS requirement of each MT to be a minimum data rate 1. On the demand side, we set the power consumptions Q_1 and Q_2 at the two BSs as their transmission power only; on the supply side, we set their harvested renewable energy as $E_1 = 10$ and $E_2 = 2.5$, respectively, and their energy buying price from the grid as $\pi = 1$. Additionally, for the aggregator-assisted energy trading, the BSs' energy buying and selling prices from and to the aggregator are $\pi_{buy} = 0.5$ and $\pi_{sell} = 0.4$, respectively; and for the aggregator-assisted energy sharing, the contract fee paid to the aggregator is $\overline{C} = 0.1$. Furthermore, in each scheme, the BSs employ equal bandwidth allocation among MTs, and there is only one MT served in each sub-band. Note that all units are normalized for simplicity here.

Based on the above setting, we summarize the results in Table 1, from which we make the following observations. For the two energy cooperation schemes, it is observed that the renewable energy supplies at BS 1 and BS 2 are rescheduled to better match the given energy consumption at the two BSs, thus resulting in reduced total energy cost compared to the conventional design. Regarding the two communication cooperation schemes, it is observed that the two BSs adapt their transmission power to match and better use the given cheap renewable energy supplies at the two BSs, thus leading to lower total energy costs than the conventional design. For joint energy and communication cooperation, it is observed that by exploiting both supply- and demand-side management, each joint scheme outperforms the corresponding schemes with only energy or communication cooperation. In particular, the joint energy cooperation and CoMP design is promising for maximum cost saving.

Extensions and Future Directions

Despite the aforementioned studies on energy and communication cooperation, a lot of interesting topics remain unaddressed. We list several of them as follows for future study.

Practically, energy harvesting rates in general change more slowly than wireless channel and traffic load variations, and as a consequence, the timescale of implementing energy cooperation is normally longer than that of communication cooperation. To cater for this issue, it is promising to consider the multi-timescale implementation of joint energy and communication cooperation, for example, by employing two-layer decision making with energy cooperation on a longer timescale and communications cooperation on a shorter timescale to balance the trade-off between the cost-saving performance and implementation complexity.

In practice, multiple self-interested systems (owned by different operators) can coexist or co-locate, and it is interesting to study their energy and/or communication cooperation. In this case, to establish joint energy and communication cooperation, different cellular systems may seek the advantage of resource complementarity. For example, in a preliminary study [11], it is shown that one system adequate in spectrum is willing to cooperate with another adequate in energy, since both systems can efficiently reduce their individual costs by exchanging spectrum and energy with each other. Overall, cooperation mechanism design is required to motivate or strengthen inter-system joint cooperation to a win-win situation for all systems involved.

Besides cellular networks, it is also appealing for heterogeneous communication networks (e.g., WiFi and small cells) to cooperate and reduce overall energy cost. However, these networks are different in service coverage, operated spectrum, and even energy harvesting availability (difficult indoors), and their joint energy and communication cooperation becomes more complicated than our design in the previous section.

Up to now, we have focused on the case without the use of energy storage at BSs due to the cost consideration. With the advancement of battery technologies, we envision that energy storage may be employed in future BSs, and it is promising to study energy and communication cooperation jointly with storage management (e.g., [14, 15] for initial studies on joint energy cooperation and storage management). Nevertheless, such joint time and space domain optimization problems are very challenging to solve, since any present decisions made by BSs would affect their storage status and traffic loads served in the future.

CONCLUSION

This article provides an overview on novel energy and communication cooperation approaches for energy cost saving in cellular networks powered by renewable energy sources and smart grid. These approaches use both the two-way energy flow in smart grid and the communication cooperation in cellular networks to reshape the non-uniform energy supplies and energy demands over the cellular networks for cost saving. It is our hope that these new approaches can bring new insights on energy demand management in smart grid by considering the unique properties of the cellular networks' communication demand, and also on the wireless resource allocation in cellular networks by taking into account the new characteristics of the emerging renewable and smart grid energy supply.

REFERENCES

[1] Z. Hasan, H. Boostanimehr, and V. Bhargava, "Green Cellular Networks: A Survey, Some Research Issues and Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 13, no. 4, 2011, pp. 524–40.

[2] J. Wu, S. Rangan, and H. Zhang, *Green Communications: Theoretical Fundamentals, Algorithms, and Applications*, CRC Press, Sep. 2012.

[3] T. Han and N. Ansari, "Powering Mobile Networks with Green Energy," *IEEE Wireless Commun.*, vol. 21, no. 1, Feb. 2014, pp. 90–96.

[4] S. Bu *et al.*, "When the Smart Grid Meets Energy-Efficient Communications: Green Wireless Cellular Networks Powered by the Smart Grid," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, Aug. 2012, pp. 3014–24.

[5] X. Fang *et al.*, "Smart Grid — The New and Improved Power Grid: A Survey," *IEEE Commun. Surveys & Tutorials*, vol. 14, no. 4, 2012, pp. 944–80.

[6] W. Saad *et al.*, "Game-Theoretic Methods for the Smart Grid: An Overview of Microgrid Systems, Demand-Side Management, and Smart Grid Communications," *IEEE Sig. Processing Mag.*, vol. 29, no. 5, Sep. 2012, pp. 86–105.

[7] Z. Niu *et al.*, "Cell Zooming for Cost-Efficient Green Cellular Networks," *IEEE Commun. Mag.*, vol. 48, no. 11, Nov. 2010, pp. 74–78.

[8] A. Goldsmith *et al.*, "Breaking Spectrum Gridlock with Cognitive Radios: An Information Theoretic Perspective," *Proc. IEEE*, vol. 97, no. 5, May 2009, pp. 894–914.

[9] D. Gesbert *et al.*, "Multi-Cell MIMO Cooperative Networks: A New Look at Interference," *IEEE JSAC*, vol. 28, no. 9, Dec. 2010, pp. 1380–1408.

[10] L. Gkatzikis, I. Koutsopoulos, and T. Salonidis, "The Role of Aggregators in Smart Grid Demand Response Markets," *IEEE JSAC*, vol. 31, no. 7, Jul. 2013, pp. 1247–57.

[11] Y. Guo *et al.*, "Joint Energy and Spectrum Cooperation For Cellular Communication Systems," *IEEE Trans. Commun.*, vol. 62, no. 10, Oct. 2014, pp. 3678–91.

[12] J. Xu and R. Zhang, "Cooperative Energy Trading in Comp Systems Powered by Smart Grids," *Proc. IEEE GLOBECOM*, Dec. 2014, pp. 2738–43.

[13] J. Xu and R. Zhang, "CoMP Meets Smart Grid: A New Communication and Energy Cooperation Paradigm," to appear, *IEEE Trans. Veh. Tech.*, http://arxiv.org/abs/1303.2870.

[14] Y. K. Chia, S. Sun, and R. Zhang, "Energy Cooperation in Cellular Networks with Renewable Powered Base Stations," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, Dec. 2014, pp. 6996–7010.

[15] J. Leithon, T. J. Lim, and S. Sun, "Energy Exchange Among Base Stations in A Cellular Network through the Smart Grid," *Proc. IEEE ICC*, June 2014, pp. 4036–41.

BIOGRAPHIES

JIE XU [S'12, M'13] received his B.E. and Ph.D. degrees from the University of Science and Technology of China in 2007 and 2012, respectively. He was a research fellow with the ECE Department, National University of Singapore. He is currently a postdoctoral research fellow with the Engineering Systems and Design Pillar, Singapore University of Technology and Design. His current research interests include smart grid, multiuser MIMO, energy efficiency, and energy harvesting in wireless communication.

LINGJIE DUAN [S'09, M'12] received his Ph.D. degree from the Chinese University of Hong Kong in 2012. He is an assistant professor in the Engineering Systems and Design Pillar, Singapore University of Technology and Design. His research interests include network economics, network optimization, and energy harvesting. He was the recipient of the Global Scholarship for Research Excellence at CUHK in 2011, and is a Finalist for the Hong Kong Young Scientist Award in 2014.

RUI ZHANG [S'00, M'07, SM'15] received his Ph.D. degree from Stanford University in 2007. He is now an assistant professor with the ECE Department of the National University of Singapore. His current research interests include multiuser MIMO, cognitive radio, energy-efficient and energy-harvesting-based wireless communications, and wireless information and power transfer. He was the recipient of the 6th IEEE ComSoc Asia-Pacific Best Young Researcher Award in 2011. He is now an Editor for *IEEE Transactions on Wireless Communications* and *IEEE Transactions on Signal Processing*.

These approaches use both the two-way energy flow in smart grid and the communication cooperation in cellular networks to reshape the non-uniform energy supplies and energy demands over the cellular networks for cost-saving.

Advertisers' Index

ADVERTISING SALES OFFICES

Closing date for space reservation: 15th of the month prior to date of issue

National Sales Office
James A. Vick
Sr. Director Advertising Business, IEEE Media
Email: jv.ieeemedia@ieee.org

Marion Delaney
Sales Director, IEEE Media
Email: md.ieeemedia@ieee.org

Mindy Belfer
Advertising Sales Coordinator
Email: m.belfer@ieee.org

Northern California
George Roman
Tel: (702) 515-7247
Fax: (702) 515-7248
Cell: (702) 280-1158
Email: George@George.RomanMedia.com

Southern California
Patrick Jagendorf
Tel: (562) 795-9134
Fax: (562) 598-8242
Email: pjagen@verizon.net

Mid-Atlantic
Dawn Becker
Tel: (732) 772-0160
Fax: (732) 772-0164
Email: db.ieeemedia@ieee.org

Northeast
Merrie Lynch
Tel: (617) 357-8190
Fax: (617) 357-8194
Email: Merrie.Lynch@celassociates2.com

Jody Estabrook
Tel: (77) 283-4528
Fax: (774) 283-4527
Email: je.ieeemedia@ieee.org

Southeast
Scott Rickles
Tel: (770) 664-4567
Fax: (770) 740-1399
Email: srickles@aol.com

Midwest/Central Canada
Dave Jones
Tel: (708) 442-5633
Fax: (708) 442-7620
Email: dj.ieeemedia@ieee.org

Midwest/Ontario, Canada
Will Hamilton
Tel: (269) 381-2156
Fax: (269) 381-2556
Email: wh.ieeemedia@ieee.org

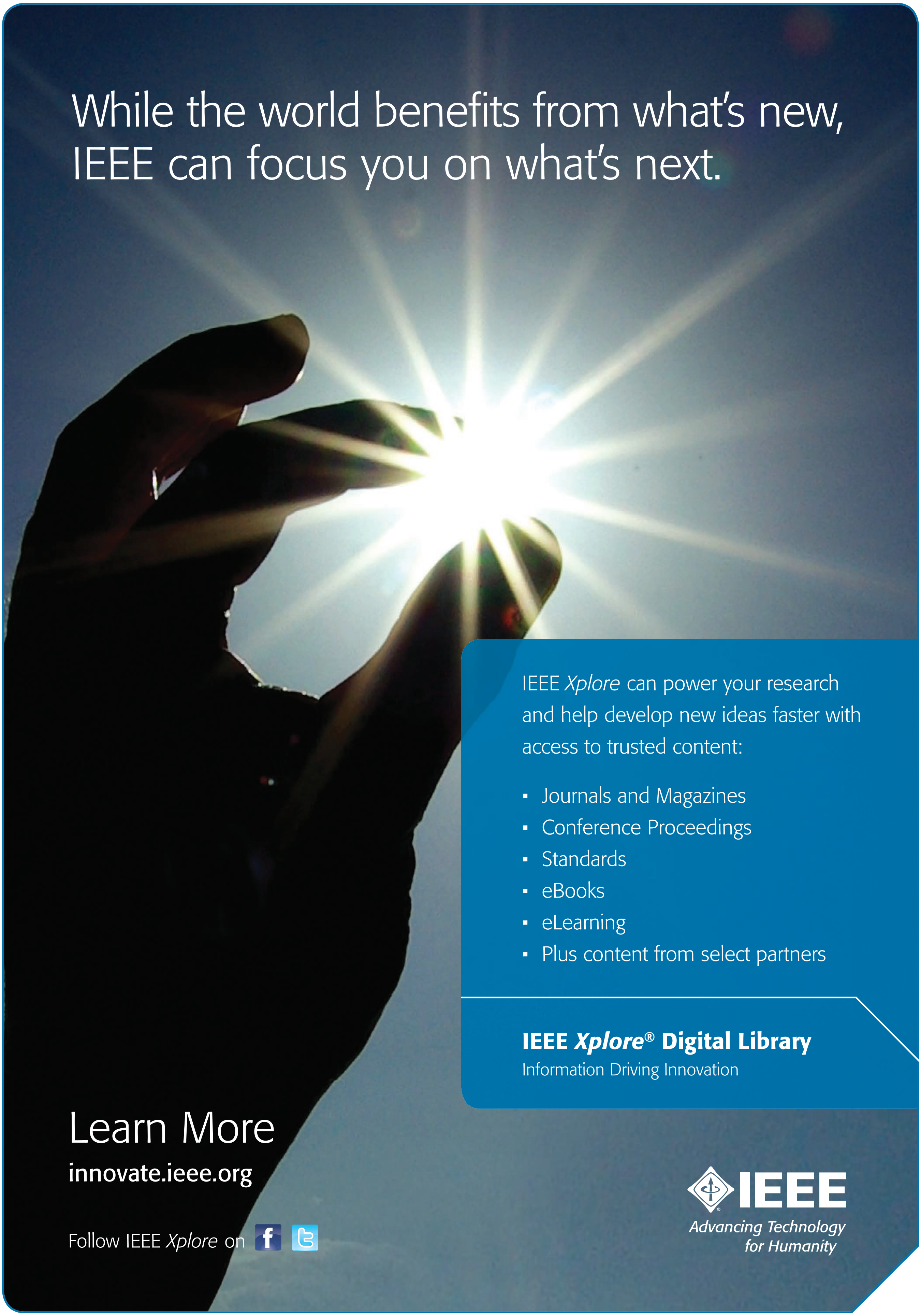
Texas
Ben Skidmore
Tel: (972) 587-9064
Fax: (972) 692-8138
Email: ben@partnerspr.com

Europe
Christian Hoelscher
Tel: +49 (0) 89 95002778
Fax: +49 (0) 89 95002779
Email: Christian.Hoelscher@husonmedia.com

Currently Scheduled Topics

Topic	Issue Date	Manuscript Due Date
Communications Education and Training: Ethics Training and Standards	November 2015	July 1, 2015
Towards Autonomous Driving: Advances in V2X Connectivity	December 2015	June 1, 2015

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


While the world benefits from what's new, IEEE can focus you on what's next.
IEEE Xplore can power your research and help develop new ideas faster with access to trusted content:
Journals and Magazines
Conference Proceedings
Standards
eBooks
eLearning
Plus content from select partners
IEEE Xplore® Digital Library
Information Driving Innovation
Learn More
innovate.ieee.org
Follow IEEE Xplore on
IEEE
Advancing Technology for Humanity

Fraunhofer
Heinrich Hertz Institute

VISIBLE LIGHT COMMUNICATION
UP TO 1 Gbit/s

Plug and play optical wireless links based on off-the-shelf LEDs featuring broadband data transmission up to 1 Gbit/s

Specifications
- Light source: any high-power LED
- Bidirectional data exchange
- Dynamic rate adaptation
- Peak data rate 1 Gbit/s LOS and NLOS scenarios supported
- Ultra-low latency (<2 ms)

Benefits
- Combination of lighting and data communication
- Well suited in environments where radio is not acceptable or feasible
- Available also with near infrared or ultraviolet LEDs
- Improved privacy
- CE certified

Applications
- **Wireless communication in industrial environments**
- Car-to-X-communication
- Optical Wi-Fi in RF-sensitive areas (hospitals)
- Backhauling for Wi-Fi and LTE
- Sensor networks
- Advertising, messaging
- Optical wireless hotspot

E-Mail: vlc@hhi.fraunhofer.de
Web: www.hhi.fraunhofer.de/vlc-prototype