June 2015, Vol. 53, No. 6

IEEE Communications MAGAZINE

www.comsoc.org

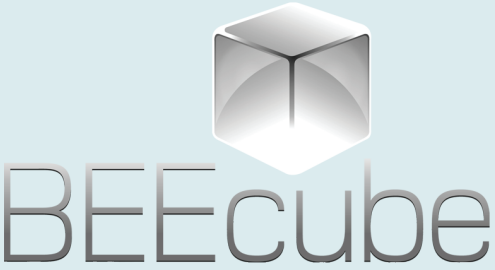
- Wireless Physical Layer Security
- Energy Harvesting Communications
- Automotive Networking and Applications
- Radio Communications

IEEE COMMUNICATIONS SOCIETY

A Publication of the IEEE Communications Society

IEEE Communications MAGAZINE

THANKS OUR CORPORATE SUPPORTERS



June 2015, Vol. 53, No. 6

IEEE Communications MAGAZINE

www.comsoc.org

- Wireless Physical Layer Security
- Energy Harvesting Communications
- Automotive Networking and Applications
- Radio Communications

IEEE COMMUNICATIONS SOCIETY

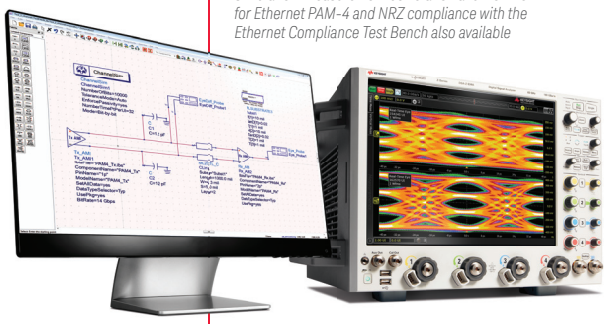
A Publication of the IEEE Communications Society

PAM-4 insights don't schedule meetings.

They come when they're good and ready.

Some call them Eureka moments. Others call them epiphanies. We call them insights, the precise moments when you know you've found great answers. As the networking industry considers transitioning to more complex signaling, we can help you achieve insights to meet the technical challenges of PAM-4 that lie ahead. From simulating new designs to characterizing inputs, outputs and connectors, we have the software, hardware and measurement expertise you need to succeed.

Keysight Advanced Design System bundle for signal integrity
Simulation-measurement correlation and workflow for Ethernet PAM-4 and NRZ compliance with the Ethernet Compliance Test Bench also available



Keysight Infiniium Z-Series oscilloscopes
Compliance solutions available for current and emerging PAM-4/Ethernet standards

HARDWARE + SOFTWARE + PEOPLE = PAM-4 INSIGHTS

PEOPLE

- *Member representatives in test working groups including IEEE, OIF-CEI, and Fibre Channel Industry Association*
- *Applications engineers in more than 100 countries around the world*
- *Nearly 1,000 patents granted or pending*

Download our app note *PAM-4 Design Challenges and the Implications on Test* at www.keysight.com/find/PAM-4-insight

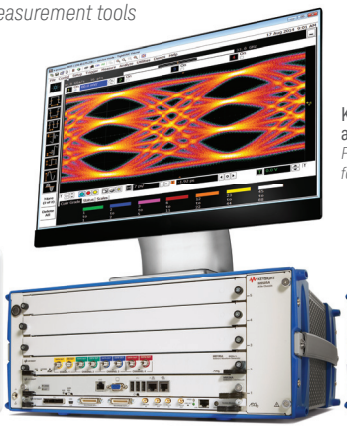
USA: 800 829 4444 CAN: 877 894 4414

HARDWARE + SOFTWARE

- *Instruments designed for testing PAM-4 from simulation to compliance*
- *Advanced Design System software for simulation-measurement correlation and workflow*
- *More than 4,000 electronic measurement tools*

Keysight 86100D Infiniium DCA-X wide-bandwidth oscilloscope
Compliance solutions for emerging optical and electrical PAM-4/Ethernet standards

Keysight N5245A PNA-X microwave network analyzer with N1930B physical-layer test system software
Gigabit Ethernet interconnect and channel test solutions



Keysight M8195A 65-GSa/s arbitrary waveform generator
Flexible PAM-4 pattern generation for 400G Ethernet and beyond

Keysight J-BERT M8020A high-performance BERT
The most integrated solution for 100G Ethernet input testing

KEYSIGHT TECHNOLOGIES

Unlocking Measurement Insights

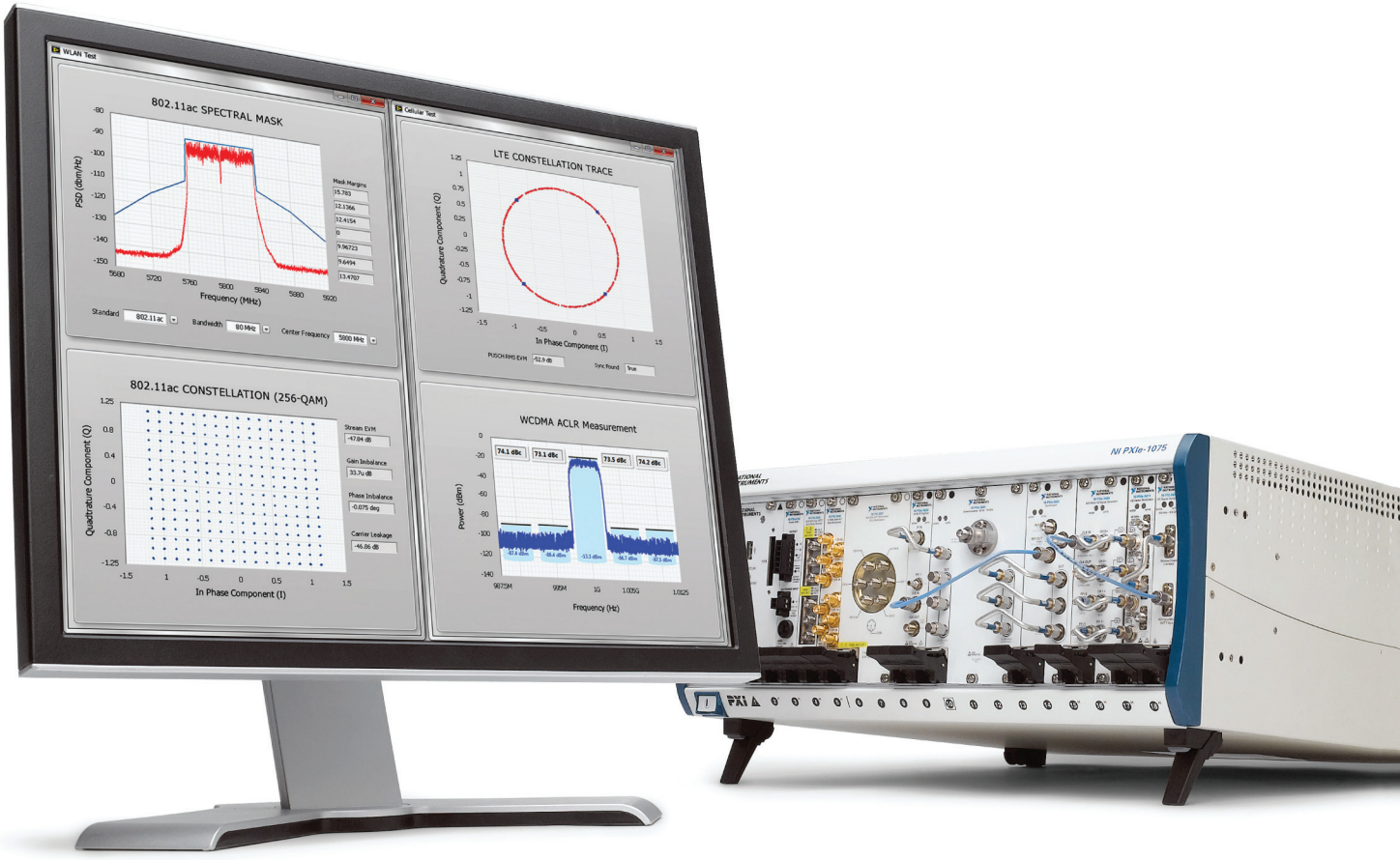
IEEE Communications MAGAZINE

JUNE 2015, Vol. 53, No. 6

www.comsoc.org/commag

Redefining RF and Microwave Instrumentation

with open software and modular hardware



Achieve speed, accuracy, and flexibility in your RF and microwave test applications by combining National Instruments open software and modular hardware. Unlike rigid traditional instruments that quickly become obsolete by advancing technology, the system design software of NI LabVIEW coupled with NI PXI hardware puts the latest advances in PC buses, processors, and FPGAs at your fingertips.

WIRELESS TECHNOLOGIES

National Instruments supports a broad range of wireless standards including:

802.11a/b/g/n/ac	LTE
CDMA2000/EV-DO	GSM/EDGE
WCDMA/HSPA/HSPA+	Bluetooth

>> Learn more at ni.com/redefine

800 813 5078

NATIONAL INSTRUMENTS™

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.

CURRENTLY SCHEDULED TOPIC

TOPIC	ISSUE DATE	MANUSCRIPT DUE DATE
COMMUNICATIONS EDUCATION AND TRAINING: ETHICS TRAINING AND STANDARDS	NOVEMBER 2015	JULY 1, 2015
WIRELESS COMMUNICATIONS, NETWORKING, AND POSITIONING WITH UNMANNED AERIAL VEHICLES	MAY 2016	NOVEMBER 1, 2015
BIO-INSPIRED CYBER SECURITY FOR COMMUNICATIONS AND NETWORKING	JUNE 2016	NOVEMBER 1, 2015
WIRELESS TECHNOLOGIES FOR DEVELOPMENT	JULY 2016	DECEMBER 1, 2015

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

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

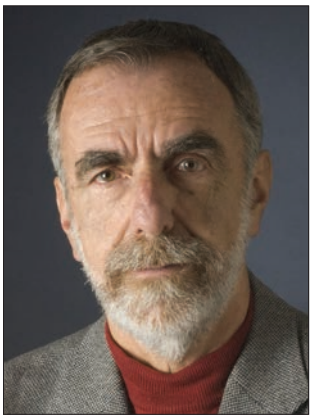
Proposals are also solicited for Feature Topic issues of the Communications Standards Supplement.

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.

THE GREEN ICT INITIATIVE: AN IEEE-WIDE FOCUS BUILDING UPON COMSOC'S LEADERSHIP



SERGIO BENEDETTO

BYEONG GI LEE

This month's President's Page is devoted to the IEEE Green Information and Communications Technology (ICT) initiative. By its very nature, Green ICT is a theme, not only of interest but also offering numerous opportunities, for virtually every IEEE Society and Council. Through this initiative, ComSoc seeks an IEEE-wide outreach to achieve even greater recognition for IEEE's mission of advancing technology for humanity.

The Green ICT initiative is jointly chaired by two ComSoc volunteers. Jaafar Elmirghani is Director of the Institute of Integrated Information Systems within the School of Electronic and Electrical Engineering, University of Leeds, United Kingdom. Charles Despins is President & CEO of Prompt, an ICT research consortium based in Montreal, Canada, as well as a faculty member at the Université du Québec. As shown by their bios at the end of this text, both of them bring a wealth of highly complementary industrial and academic experience to drive the initiative.

As a new endeavour formally launched at the beginning of this year, the Green ICT initiative is unveiling its web portal (greenict.ieee.org) this month at IEEE ICC 2015 in London. In view of Green ICT's breadth, the portal is meant to be a focal point for activities and the latest news on this theme throughout IEEE Societies, Councils, and other technology initiatives. Such activities can include publications, conferences and workshops, standards, or educational offerings.

The term "Green ICT" can often be perceived as controversial and conjuring images of a strict focus on environmental impacts to the detriment of other important issues for humanity. To the contrary, ICTs are well known enablers of productivity and economic development, as well as quality of life improvements and positive social impacts (when properly applied). The environmental benefits of ICT position it as one of the most important drivers of sustainable development in the 21st century. In a world where the economy and the environment are too often seen as incompatible, Information and Communications Technologies are in fact one of the few tools humanity can leverage to reconcile economic, social, and environmental benefits, i.e. the three pillars of sustainability.

WHAT IS GREEN ICT?

Green ICT generally refers to the design and application of information and communications technologies (ICT) in order to create environmental benefits. Moreover, as ICTs are finding applications in almost every sphere of human activity, the foreseen impact of "greening by ICT" is considered to be even ultimately greater than "greening ICT" itself. The environmental benefits associated with Green ICT can be measured through the following metrics:

Improved Energy Efficiency: This metric is well known to ICT engineers, and the ICT industry has been focusing on it, notably to reduce operational expenditures (OPEX) in the context of the huge traffic increases on communications networks, but also in various application areas of ICT (e.g. intelligent transportation systems, smart grids, etc.).

Reduction of Carbon (and Polluting Atmospheric) Emissions: ICT is currently estimated to offer the potential to eliminate throughout society seven times[1] the size of its own carbon footprint. As energy efficiency does not always lead directly to carbon emission reductions, strategies will be required to integrate renewable energy sources in the design of networks in order to make the latter and the myriad of applications supported by networks as low-carbon as possible. The environ- mental and social advantages resulting from low-carbon ICTs could further be monetized by the ICT industry if and when prices on carbon and polluting emissions proliferate throughout the world.

Life-Cycle Management: Targeting improvements in energy efficiency and reduction of carbon emissions is important not only in the operational phase of information and communications technologies, but also in their manufacturing and disposal phases. The rapid growth of the industry has notably led to an increase in e-waste.

Underlying the concept of Green ICT is the notion of convergence, with all the application sectors that can be greened by ICT, and also specifically between the ICT industry and the energy industry (which can be considered as one of these sectors). Just as our mastery of energy has transformed our way of life since the industrial revolution, our mastery of information is putting us in the midst of another similar revolution, transforming every facet of life, changing business models, and creating new opportunities that can significantly benefit humanity. ICT and energy are now fundamental enablers of our 21st century societies; a holistic approach to the design of ICT and energy infrastructures is a key element to fully leverage ICTs for sustainability.

This notion of convergence implies that Green ICT is in fact a very multi-disciplinary theme bringing together expertise not only from the various ICT sub-sectors (components and systems) but also from all those sectors that can be greened by ICT, e.g. healthcare, transport, buildings, etc. As an example, one can consider the numerous convergent Green ICT issues in the communications-energy-transport triumvirate, particularly when electric vehicles are brought into

[1] *http://gesi.org, Smarter 2020 report*

the mix along with Big Data, Internet of Things, and Cloud technologies.

THE GREEN ICT INITIATIVE

Although they may not always be labeled as such, the huge interest in various sub-themes of Green ICT is spurring a growing number of activities in IEEE (and in other global organizations), which speaks eloquently to the relevance of the theme. Nonetheless, this explosion of interest is a double-edged sword as it can also lead to the creation of technical interest "silos". While encouraging the development of various activities on the theme, the Green ICT initiative will therefore seek to develop holistic awareness on this topic, avoid duplication or overlap of activities, and provide a forum to build a cross-society and cross initiative IEEE consensus on outreach to other organizations. Through its interface with various IEEE societies and other technology initiatives (e.g. Cloud Computing, IoT, Smart Cities, etc.), the Green ICT initiative will foster the incorporation of green metrics and standards in design concepts for various technical domains. The initiative will also bring together expertise from different fields, in conferences and publications, with a view to foster holistic design and standardization approaches.

In view of the preceding, the Green ICT initiative mission statement has been defined as:

Build a holistic approach to sustainability through ICT by incorporating green metrics throughout IEEE technical domains.

The Green ICT initiative has in focus a set of priority areas that target participation from multiple Societies, Councils, and other IEEE technology initiatives.

Publications. The IEEE Green ICT initiative will target the creation of an IEEE Green ICT Transactions and Magazine, with broad IEEE Society and Council sponsorship, targeting holistic and multidisciplinary approaches to the incorporation of green metrics in all IEEE fields of interest. To incubate the new Transactions, IEEE ComSoc has started a new series on Green Communications and Networking in the *IEEE Journal on Selected Areas in Communications*. The first call for papers closed in March 2015 and was extremely popular and highly oversubscribed. The second call for papers has a closing date in July 2015. The Series has recruited an Editor-in-Chief and a number of experienced area editors and associate editors. The new IEEE Green ICT Magazine is expected to be a forum that brings together the different IEEE societies, industry, and academia.

Conferences and Events: The development of IEEE conferences, workshops, and symposia that take an integrated view of greening ICT and greening by ICT is one of the key objectives of the initiative. It will build collaboration with established industrial and policy bodies in this area, including GreenTouch and GeSi. Within IEEE ComSoc the first Green track at ICC/GLOBECOM was launched at GLOBECOM 2011 within the Selected Areas in Communications Symposium. This track has since grown significantly, and at GLOBECOM 2014 in Austin was elevated to a full Symposium. ComSoc started GreenComm, an entirely online event currently in its fifth edition as a pioneering development. ComSoc's Technical Committee on Green Communications and Computing (TCGCC) and Technical Committee on Transmission, Access and Optical Systems (TAOS) have also been very active on Green ICT. The Green ICT initiative, led by ComSoc, will build on this heritage and develop a pan-IEEE Green ICT conference that enables researchers and practitioners to address Green ICT within the full scope of IEEE.

Standards: The development of standards to properly assess the full environmental impacts (energy consumption, carbon and polluting emissions, e-waste, etc.) of ICTs throughout their life cycle presents a substantial challenge. As an example, in the realm of atmospheric emissions, common methodologies for carbon emissions, such as ISO 14064 used for thermal power plants, heavy manufacturing, etc., are difficult to apply in the ICT sector, notably as both steady network operators and spontaneous end users are involved. Building upon various sensing data both internal to ICT systems and external (e.g. electrical utilities), ICT carbon and polluting emission "foot-printing" typically incorporates[2] approaches at various levels of granularity and at various time intervals throughout the life cycle (manufacturing, operation, disposal). Real-time foot-printing can inform end-users of their environmental impact when using an ICT service or application; it can also help ICT network operators to optimize server management when these are located in areas with varying mixes of power generation sources.

Various international organizations are developing such standards, but these laudable activities remain fragmented. Beyond promoting the use of these methodologies in the R&D and design activities of the technical communities in various IEEE societies, a significant opportunity exists for IEEE as a neutral, fast-moving standards organization. Following a Green ICT standards SWOT analysis, an IEEE Green certification label could be developed.

Education: Green ICT training activities in the form of tutorials have been offered at major IEEE conferences, versions of which will be made available later to the wider IEEE membership through the initiative web portal. The initiative will also organize workshops targeting different application domains and industry sectors through white papers and interactive educational material. A number of tutorials have been delivered so far, for example at ICC 2013, SoftCOM 2013, ICC 2014, ICC 2015, and panels at CCEM 2014, COMPSAC 2014, ICCE 2014, ICC 2014, ICCE 2014, INTELECT 2015, CCNC2015, and ICC 2015. More are planned in the coming months in order to stimulate interest on Green ICT and to raise awareness of the breadth of the theme.

Outreach and Advocacy: A web portal has been designed to stimulate participation in the initiative from different IEEE technical societies. The portal will notably seek to foster inter-society exchanges and collaboration on the incorporation of green metrics in engineering design and research. The portal will also be used as an outreach tool to individuals and organizations outside IEEE's traditional communities. This will help grow IEEE's membership base. The portal is being launched at IEEE ICC 2015.

IEEE SOCIETIES AND STEERING COMMITTEE

Led by the IEEE Communications Society, 16 key IEEE stakeholders have expressed interest in this new initiative. These are: Communications Society, Aerospace and Electronic Systems Society, Computer Society, Consumer Electronics Society, Consumer Electronics & Product Safety Engineering Society, Council on Electronic Design Automation, Council on Superconductivity, Educational Activities, Member and Geographic Activities (MGA), Micro- wave Theory and Techniques Society, Oceanic Engineering Society, Photonics Society, Power and Energy Society, Standards Association, Technical Activities, and Vehicular Technology Society.

A steering committee has also been established, and in

[2] *Equation project (www.equationict.com), LCA team project report, private communication, December 2014.*

addition to the initiative's two co-Chairs, the steering group includes (in alphabetical order):

- Susan Brooks, IEEE ComSoc Executive Director, USA.
- Mohamed Chériet, Synchromedia Lab Director, Professor and Canada Research Chair, École de Technologie Supérieure (Université du Québec), Canada.
- Tarek El-Bawab, IEEE ComSoc Director of Conference Operations, Jackson State University, USA.
- Rob Fish, Vice-President, Standards Activities, ComSoc, and President, NETovations Group LLC, USA.
- Kathy Grise, Senior Program Director, Future Directions, IEEE Technical Activities, USA.
- Kerry Hinton, Principal Research Fellow, Centre for Energy-Efficient Telecommunications (CEET), University of Melbourne, Australia.
- Dan Kilper, Director of the Center for Integrated Access Networks, University of Arizona, USA.
- Thierry Klein, Chair of the GreenTouch Technical Committee and Network Energy Research Program Leader, Bell Labs Alcatel-Lucent, USA.
- Louise Krug, Senior Researcher, Carbon Reduction Strategy, British Telecom, United Kingdom.
- Fabrice Labeau, President, IEEE Vehicular Technology Society, and Associate Professor, McGill University, Canada.
- Magnus Olsson, Senior Researcher, Energy Performance, Ericsson Research, Sweden.
- Sarah Kate Wilson, Vice-President, Publications, IEEE ComSoc, and Associate Professor, Santa Clara University, USA.
- Ke Wu, President-Elect, IEEE Microwave Theory and Techniques Society, and Professor and Canada Research Chair, École Polytechnique de Montréal, Canada.

THE GREEN ICT INITIATIVE: A CALL TO ACTION THROUGHOUT IEEE

IEEE's mission statement defines its core purpose as fostering technological innovation and excellence for the benefit of humanity. When viewed through the lens of the triple bottom line (economic, environmental, social) of sustainability, the Green ICT initiative thus offers a compelling opportunity for IEEE to demonstrate the full impact of the technological innovation that it supports.

ComSoc is highly pleased to develop the Green ICT initiative and calls upon all its members to leverage the web portal to publicize all Green ICT activities (events, news, publications, technical committee developments, etc.). But as stated earlier, the success of the initiative will ultimately hinge on IEEE-wide participation. This call to action therefore goes out to all IEEE Societies, Councils and Initiatives. Green ICT … it's everybody's business!

If this broad participation in IEEE can be achieved, outreach and advocacy, beyond IEEE's technological communities, could then serve as a lever to generate even broader interest in IEEE's activities as well as to diversify its membership base.

THE GREEN ICT INITIATIVE CO-CHAIRS BIOGRAPHIES

PROF. JAAFAR ELMIRGHANI [M'91] is the Director of the Institute of Integrated Information Systems within the School of Electronic and Electrical Engineering, University of Leeds, UK. He joined Leeds in 2007. Prior to that (2000–2007) as the chair in optical communications at the University of Wales Swansea he founded, developed, and directed the Institute of Advanced Telecommunications and the Technium Digital (TD), a technology incubator/spin-off hub. He received the Ph.D. from the University of Huddersfield UK in 1994 and the D.Sc. in communication systems and networks from the University of Leeds, UK, in 2014. He is a Fellow of IET and a Fellow of the Institute of Physics. He has co-authored *Photonic Switching Technology: Systems and Networks* (Wiley), and has published over 400 papers. He has research interests in optical systems and networks. He was Chairman of the IEEE ComSoc Transmission Access and Optical Systems Technical Committee and was Chairman of the IEEE ComSoc Signal Processing and Communications Electronics Technical Committee, and an editor of *IEEE Communications Magazine*. He was the founding Chair of the Advanced Signal Processing for Communication Symposium, which started at IEEE GLOBECOM'99 and has continued since at every ICC and GLOBECOM. Prof. Elmirghani was also the founding Chair of the first IEEE ICC/GLOBECOM optical symposium at GLOBECOM'00, the Future Photonic Network Technologies, Architectures and Protocols Symposium. He was the founding Chair of the first Green Track at ICC/GLOBECOM at GLOBECOM 2011. He received the IEEE Communications Society Hal Sobol Award, the IEEE ComSoc Chapter Achievement award for excellence in chapter activities (both in 2005); the University of Wales Swansea Outstanding Research Achievement Award in 2006; the IEEE Communications Society Signal Processing and Communication Electronics outstanding service award in 2009; and a best paper award at IEEE ICC'2013 in Green Communications. He is currently an editor of *IEEE Communications Surveys and Tutorials* and *IEEE Journal on Selected Areas in Communications* series on Green Communications and Networking. He is Co-Chair of the GreenTouch Wired, Core and Access Networks Working Group and has been awarded in excess of £22 million in grants to date from EPSRC, the EU, and industry, and is an IEEE ComSoc Distinguished Lecturer for the term 2013–2016.

CHARLES DESPINS' [M'82] career has spanned 30 years in both the academic and industry segments of the information and communications technologies (ICT) field. He has held various posts in the private sector, namely at CAE Electronics, Microcell Telecommunications (Canadian cellular operator), and at Bell Nordiq Group (a network operator in rural and northern areas of Canada) as vice-president and chief technology officer. He has also worked as a consultant for wireless network deployments in India and China. Since January 2003 he has been President and CEO of Prompt inc., an ICT university-industry research and development consortium developing various collaborative, public-private partnership research activities tackling a broad range of ICT themes, including a recent four-year Green ICT project that brought together 40 organizations to develop more than 65 new Green ICT products, processes, and services. He is also a faculty member (on leave) at École de Technologie Supérieure (Université du Québec) in Montreal, with research interests in wireless communications, as well as a guest lecturer at the Desautels faculty of Management at McGill University in Montreal. He holds a bachelor's degree in electrical engineering from McGill University in Montreal, Canada, as well as M.Sc. and Ph.D. degrees, also in electrical engineering, from Carleton University in Ottawa, Canada. He is a Fellow (2005) of the Engineering Institute of Canada and a recipient (2006) of the Outstanding Engineer award from IEEE Canada. He was also the recipient of the 1993 Best-Paper-of-the-Year Award in *IEEE Transactions on Vehicular Technology*. He is currently a frequent advocate on issues regarding the opportunities ICT offer to achieve sustainability.

OMBUDSMAN

COMSOC BYLAWS ARTICLE 3.8.10

"The Ombudsman shall be the first point of contact for reporting a dispute or complaint related to Society activities and/or volunteers. The Ombudsman will investigate, provide direction to the appropriate IEEE resources if necessary, and/or otherwise help settle these disputes at an appropriate level within the Society…"

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

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

2015

OCTOBER

LANOMS 2015 — Latin American Network Operations and Management Symposium, 1– 3 Oct.
Joao Pessoa, Brazil
http://www.lanoms.org/2015/

IEEE CLOUDNET 2015 — 4th IEEE Int'l. Conference on Cloud Networking, 5–7 Oct.
Niagara Falls, Canada
http://www.ieee-cloudnet.org/

RNDM 2015 — 7th Int'l. Workshop on Reliable Networks Design and Modeling, 5–7 Oct.
Munich, Germany
http://www.rndm.pl/2015/

ATC 2015 — Int'l. Conference on Advanced Technologies for Communications, 14–16 Oct.
Ho Chi Minh, Vietnam
http://www.rev-conf.org/

APCC 2015 — 21st Asia-Pacific Conference on Communications, 14–16 Oct.
Kyoto, Japan
http://www.apcc2015.ieice.org/

IEEE HEALTHCOM 2015, 17th IEEE Int'l. Conference on e-Health Networking, Application & Services, 14–17 Oct.
Boston, MA
http://www.ieee-healthcom.org/index.html

WCSP 2015 — Int'l. Conference on Wireless Communications & Signal Processing, 15–17 Oct.
Nanjing, China
http://www.ic-wcsp.org/

MILCOM 2015 — Military Communications Conference, 26–28 Oct.
Tampa, FL
http://events.jspargo.com/milcom15/public/enter.aspx

IOT 2015 — 5th Int'l. Conference on the Internet of Things, 26–28 Oct.
Seoul, Korea
http://www.iot-conference.org/iot2015/

CNSM 2015 — 11th Int'l. Conference on Standards for Communications and Networking, 26–30 Oct.
Barcelona, Spain
http://www.cnsm-conf.org/2015/

IEEE CSCN 2015 — IEEE Conference on Standards for Communications and Networking, 28–30 Oct.
Tokyo, Japan
http://www.ieee-cscn.org/

NOVEMBER

IEEE/CIC ICCC 2015 — IEEE/CIC Int'l. Conference on Communications in China, 2–4 Nov.
Shenzhen, China
http://www.ieee-iccc.org/2015/

IEEE COMCAS 2015 — IEEE Int'l. Conference on Microwaves, Communications, Antennas and Electronic Systems, 2–4 Nov.
Tel Aviv, Israel
http://www.comcas.org/

(Continued on next page)

–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.

ComSoc 2015 Election: Take Time to Vote

Ballots were e-mailed and/or postal mailed 29 May 2015 to all Higher Grade* IEEE Communications Society Members and Affiliates (excluding Students) whose memberships were effective prior to 1 May 2015. You must have an e-ballot or paper ballot before you can vote.

Vote Now using the URL below. You will need your IEEE Account username/password to access the ballot. If you do not remember your password, you may retrieve it on the voter login page.

https://eballot4.votenet.com/IEEE

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

If you do not receive a ballot by 30 June, but you feel your membership was valid before 1 May 2015, you may e-mail ieee-comsocvote@ieee.org or call +1 (732) 562-3904 to check your member status. (Provide your member number, full name, and address.)

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

Voting for this election closes 24 July 2015 at 4:00 p.m. EDT! Please vote!

*Includes Graduate Student Members

Conference Calendar

IEEE SmartGridComm 2015 — 6th IEEE Int'l. Conference on Smart Grid Communications, 2–5 Nov.
Miami, FL
http://sgc2015.ieee-smartgridcomm.org/

IEEE LATINCOM 2015 — IEEE Latin American Conference on Communications, 4–6 Nov.
Arequipa, Peru
http://www.ieee-comsoc-latincom.org/2015/

IEEE OnlineGreenComm 2015 — IEEE Online Conference on Green Communications, 10–12 Nov.
Virtual
http://www.ieee-onlinegreencomm.org/2015/

IEEE NFV-SDN 2015 — IEEE Conference on Network Function Virtualization and Software Defined Networks, 18–21 Nov.
San Francisco, CA
http://www.ieee-nfvsdn.org/

DECEMBER

IEEE GLOBECOM 2015 — IEEE Global Communications Conference 2015, 6–10 Dec.
San Diego, CA
http://globecom2015.ieee-globecom.org/

ITU-K 2015 — ITU Kaleidoscope: Trust in the Information Society, 9–11 Dec.
Barcelona, Spain
http://www.itu.int/en/ITU-T/academia/kaleidoscope/2015/Pages/default.aspx

WF-IOT 2015 — IEEE World Forum on Internet of Things, 14–16 Dec.
Milan, Italy
http://www.ieee-wf-iot.org/

IEEE ANTS 2015 — IEEE Int'l. Conference on Advanced Networks and Telecommunications Systems, 15–18 Dec.
Kolkata, India
http://www.ieee-comsoc-ants.org/

IEEE VNC 2015 — IEEE Vehicular Networking Conference, 16–18 Dec.
Kyoto, Japan
http://www.iiitmk.ac.in/coconet2015/index.html

COCONET 2015 — Int'l. Conference on Computing and Network Communications, 16–19 Dec.
Trivandrum, India
http://www.iiitmk.ac.in/coconet2015/index.html

2016

JANUARY

IEEE CCNC 2016 — IEEE Consumer Communications and Networking Conference, 8–11 Jan.
Las Vegas, NV
http://ccnc2016.ieee-ccnc.org/

MARCH

OFC 2016 — Optical Fiber Conference, 20–24 Mar.
Anaheim, CA
http://www.ofcconference.org/en-us/home/

GCN GLOBAL COMMUNICATIONS NEWSLETTER

June 2015
ISSN 2374-1082

REGIONAL REPORT

IEEE ComSoc Iraq Chapter Activities Continue Despite the Severity of the Situation Inside Iraq

By Sattar Bader Sadkhan, Chair of IEEE Iraq Section

It is well known to the world how serious the situation is inside Iraq due to the terrorists activities of the ISIS-Daaish forces since June 2014. At that time very dangerous attacks were happened against one of the biggest cities in IRAQ, Ninawa, with a population of approximately 3,000,000. This occupation by the terrorism forces expanded to other cities (Tel-afar, Tikrit, Diyala, Kirkuk, Anbar, and Babylon). One of catastrophic results is the internal displacement of thousands of families from these cities to other cities. Approximately 3,000,000 people were internally displaced to many other cities such as Erbil, Duhook, and Sulimanyia, in the Kurdistan Region, as well as Baghdad, Kerbala, Al-Najef, Babylon, Waset, Al-Qadisyia, Missan, and ThiQar.

Our country has experienced much hardship in recent times, as can be seen in the effect on many daily activities of the people. Many members of the IEEE Iraq Section have faced extremely sad circumstances. They are professors in universities in these cities that were occupied by terrorism forces. However, the spiritual pain felt by the Iraq IEEE members has strengthened their desire to continue to volunteer to visit displaced families and offer them any possible support.

Prof. Dr. Sattar B. Sadkhan, the chair of the ComSoc Chapter, has made many visits during the past year to many campuses for displaced people, and six visits to the "Zuhair Al-Azdy Primary School" for displaced students established in Babylon City. One little girl's smile tells the story of the important work being done by Prof. Sadkhan and his fellow IEEE volunteers.

An Overview of IEEE Iraq ComSoc Chapyter Activities in Iraq

The IEEE Iraq ComSoc Chapter has been in operation since 2011, and in that time has taken major steps to expand activities, including conferences, workshops, specific lectures, and many social activities. The IEEE Iraq volunteers support the educational institutions by introducing the core values and objectives of the IEEE, as well as scientific activities that assist in promoting the communication engineering profession. Iraq has 18 cities and many rural regions that are not easy to access. As such, a major effort is required to distribute the activities throughout the country. The dedication of the IEEE Iraq ComSoc Chapter volunteers has made this possible.

Prof. Dr. Sattar B. Sadkhan, the chair of the ComSoc Chapter, has made many visits during the past year to many campuses for displaced people, and six visits to the "Zuhair Al-Azdy Primary School" for displaced students established in Babylon City. The volunteer team of the IEEE Iraq ComSoc Chapter spent the entire visit day listening to the hardships of the displaced students, educating them, and providing them with these words of inspiration: that this will all pass and that they have a bright future ahead.

(Continued on Newsletter page 4)

DISTINGUISHED LECTURE TOUR

Effectiveness of ComSoc DLT Program and Workshops in New Zealand

By Nurul I Sarkar, IEEE Joint NZ North, South and Central ComSoc Chair

The IEEE New Zealand (NZ) Communications Society (ComSoc) Chapter is a joint chapter of the IEEE NZ North, South, and Central Sections. This is a NZ-wide chapter that was formed in July 2014 as a result of the splitting of the NZ North joint chapter COM019/SP001/IT12. However, last year (2014) was a very productive for us from the perspective of professional development of the members of the society and the wider community.

Being a ComSoc chapter chair, Associate Professor Nurul Sarkar had nominated Professor Ekram Hossain (University of Manitoba, Canada) as IEEE ComSoc Distinguished Lecturer (DL). Professor Hossain gave three public lectures (covering the three major cities of NZ) in Christchurch, Wellington, and Auckland on November 10, 11, and 13, respectively. All three talks went very well for the professional development of the members of the society and the wider university community. A brief description of each of the talks is highlighted below.

Professor Hossain gave his first DL talk at the University of Canterbury, Christchurch on Monday, 10 November. This talk was organized by Professor Harsha Sirisena. Thanks go to Professor Krys Pawlikowski, who took over the role of organizing chair for the DLT tour in the absence of Professor Sirisena. Approximately 25 people attended the talk, and after the session there was a good discussion about the future generation cellular wireless networks, including 5G systems.

Next, Professor Hossain gave a talk in Wellington on 11 November which was organized by Dr Terence Betlehem. The talk went very well and was attended by more than 30 people, mostly from industry. There were a number of research students from Victoria University of Wellington attending the talk. After the talk the participants visited research laboratories, including the Sensing and Automation research labs.

DL-NSRG Workshop Attendees in Auckland

In Auckland, the DL talk was held at Auckland University of Technology (AUT), Auckland CBD. AUT's School of Computer and Mathematical Sciences (Network and Security Research Group) hosted this event in conjunction with a day-long Network and Security Research Group (NSRG) workshop on Thursday, 13 November, 2014. Associate Professor Jairo Gutierrez (head of Computer Sciences) gave an opening talk and outlined the program for the day. There were a series of presentations given by DLs, invited speakers, and research staff and students. In addition to Professor Hossain, we had two other high profile speakers: Professor Jaafar Elmirghani (University of Leeds, UK), and Professor Krys Pawlikowski (University of Canterbury, Christchurch). Professor Elmirghani's talk was hosted by the IEEE joint NZ North, South, and Central ComSoc Chapter and Network and Security Research Group (NSRG), AUT. The first DL talk presented by Professor Hossain, "Evolution Towards 5G Cellular Networks: Radio Resource and Interference Management Issues," was very interesting. Professor Hossain focused on the vision and requirements for 5G cellular networks, device-to-device communications, enabling technologies for 5G networks, and interference management challenges in 5G multi-tier networks. Some issues and open research areas were discussed.

Next, Professor Elmirghani gave an impressive talk on "Greening Core, Data Centre and Content Distribution Networks." More on Elmirghani's talk in other cities is discussed later in the report. Professor Pawlikowski then gave a useful talk on aspects of simulation credibility in telecommunication networks. The tutorial style presentations helped the audience understand the technical subjects very well. After each talk there was ample opportunity for question and answer and further discussion.

Among the other 10 presenters, Dr. William Liu gave a short presentation on "Energy Efficient and Resilient Network Design." The remaining nine Ph.D. students from NSRG gave mini presentations on the day. This event was financially supported in part by NSRG (AUT), IEEE ComSoc, and IEEE Region 10. Despite the busy time of year approximately 40 people (approximately 25 IEEE and 15 non-IEEE members) within and outside of AUT attended the event. People enjoyed talking and networking during coffee and lunch breaks. Organizing chair Associate Professor Nurul Sarkar received positive feedback from the participants, indicating that the event was successful.

Professor Elmirghani's DL Talks in Wellington and Dunedin

In addition to the talk at Auckland, Professor Elmirghani gave two other lectures, one in Wellington (11 November) and the other in Dunedin (14 November). The lecture in Wellington was organized by Dr. Mansoor Shafi, hosted by the IEEE NZ Central Section. Professor Elmirghani gave his final talk at Otago Polytechnic, Dunedin. This talk was organized by Dr. Tom Qi and hosted by the IEEE NZ South Section.

In both lectures, Professor Elmirghani highlighted the measures that can be used to reduce the power consumption of the Internet. Among the key measurers, Professor Elmirghani focused on the following: optimum use of time varying renewable energy in core networks; physical topology design consider-

(Continued on next page)

IEEE ComSoc DL-NSRG workshop attendees in Auckland. Left front row sitting: Professor Jairo Gutierrez, AUT, Professor Ekram Hossain, DL, (University of Manitoba), Professor Jaafar Elmirghani, DL, (University of Leeds, UK), Professor Krys Pawlikowski (University of Canterbury) and A/Professor Nurul Sarkar, AUT.

DLT Tour in New Zealand/*Continued from page 2*

ing operational and embodied energies; elastic optical networks using mixed line rates and optical OFDM; optimum resource allocation and green network design with data centers; dynamic energy-efficient content caching; energy-efficient peer-to-peer content distribution; energy-efficient distributed clouds; and energy-efficient network virtualization. The talks generated a lot of interest among the participants for further discussion and collaboration.

Professor Jayasumana's Invited DL Talk in Auckland

Professor Anura Jayasumana (Colorado State University, USA) visited both Christchurch and Auckland and gave invited DL talks. The effectiveness of Professor Jayasumana's talk in Auckland is briefly discussed below. The lecture was held at AUT on Thursday, 20 November, 2014. The talk, "Topology Coordinate System: A Novel Domain for Self-Organizing Large-Scale 2D and 3D Sensor Networks," generated much interest among the participants, and there was a good discussion and question and answer session after the talk. This event was jointly supported by AUT's School of Computer and Mathematical Sciences and the IEEE NZ North Section. Approximately 25 people attended the talk (mostly staff and students), and one IEEE member travelled from Whangarei (about 160 km North of Auckland) to attend the talk.

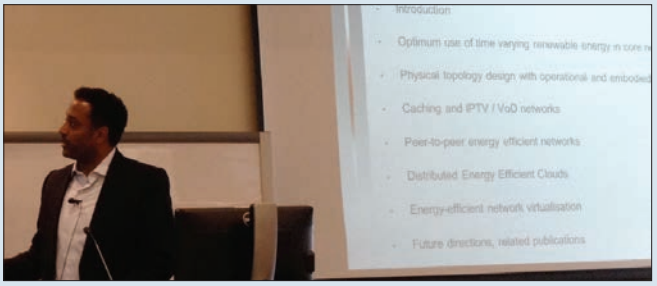
2014 IEEE NZ Wireless Workshop

The IEEE NZ ComSoc chapter organized a day-long IEEE NZ Wireless Workshop held on Friday, 5 September, 2014 at the University of Canterbury, Christchurch. This annual event brought together more than 70 engineers, researchers, industrialists, and policy makers working in the field of wireless communications and network technologies. There was a series of presentations by the participants from industry, wireless research centers, and academia, with ample opportunity for informal discussion and networking. The presentations covered various topics and provided a forum for experts in the wireless industry and academia to discuss innovative technologies and research currently being undertaken.

Conclusion

ComSoc DLT programs and workshops were very productive in 2014. We had three high profile IEEE DLs (Professor Ekram Hossain, Professor Jaafar Elmirghani, and Professor Anura Jayasumana) who travelled to NZ and gave several lectures in the major cities of NZ. In addition to the professional development of the members of the society and the wider community, we had ample opportunities for international links and collaboration. In summary, IEEE DL programs were very effective for the attendees for networking, forming academy-industry links, and sharing ideas. The DLT tour to NZ was supported by IEEE ComSoc, Region 10, IEEE NZ North, South, and Central Sections, and AUT University.

IEEE DL Prof Jayasumana's talk attendees in Auckland.



Invited DL Professor Jaafar Elmirghani's talk in Dunedin.

2014 IEEE NZ Wireless workshop participants.

ComSoc 2015 Election: Take Time to Vote

Ballots were e-mailed and/or postal mailed 29 May 2015 to all Higher Grade* IEEE Communications Society Members and Affiliates (excluding Students) whose memberships were effective prior to 1 May 2015. You must have an e-ballot or paper ballot before you can vote.

Vote Now using the URL below. You will need your IEEE Account username/password to access the ballot. If you do not remember your password, you may retrieve it on the voter login page.

https://eballot4.votenet.com/IEEE

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

If you do not receive a ballot by 30 June, but you feel your membership was valid before 1 May 2015, you may e-mail ieee-comsocvote@ieee.org or call +1 (732) 562-3904 to check your member status. (Provide your member number, full name, and address.)

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

Voting for this election closes 24 July 2015 at 4:00 p.m. EDT! Please vote!

*Includes Graduate Student Members

CONFERENCE REPORT

Tutorial Sessions on Quantum and Optical Wireless Communications and Exhibition at IEEE TENCON 2014, Bangkok, Thailand

By Kamol Kaemarungsi and Keattisak Sripimanwat, Thailand

During IEEE TENCON 2014 held October 22–25, 2014 in Bangkok, Thailand, the IEEE ComSoc Thailand Chapter assisted the local organizer by setting up three tutorial sessions for the conference, and also joined as a speaker in the IEEE leadership and membership development forum.

For more than three decades, TENCON of the IEEE Region 10 or Asia Pacific Region has been served as a networking and discussion forum for researchers and engineers in electrical and electronics engineering, computer science, and related fields. This year the IEEE ComSoc Thailand Chapter chair was invited to present our past and current activities in promoting the IEEE Communications Society and membership in Thailand in the IEEE leadership and membership development forum. We highlighted our continuous communications with our local audiences with monthly e-newsletters, special technical lectures, and free telecommunications related e-books.

The first tutorial session was entitled "Optical Wireless Communications: Challenges and Perspectives," presented by Dr. Anh T. Pham, who is a senior associate professor at the Computer Communications Laboratory, the School of Computer Science & Engineering at the University of Aizu, Japan.

The second tutorial session was entitled "The Language of Lighting Design," delivered by Dr. Chanyaporn Chuntamara, who is the head of the Lighting Research and Innovation Center (LRIC) at the King Mongkut's University of Technology Thonburi, Thailand.

The last tutorial session was entitled "Principle of Quantum Communication," presented by Dr. Suwit Kiravittaya, who is a lecturer from the Department of Electrical and Computer Engineering, Faculty of Engineering, Naresuan University, Thailand.

During the conference, our chapter also exhibited the Thai Telecommunications Knowledge Management (TTKM) project to the conference participants. This project is our continuous project that provides a free web portal, which archives telecommunications related resources for Thai people. The web portal contains links to the IEEE ComSoc Thailand Chapter website, the Thai Telecommunications Encyclopedia, and other related material.

During the event our chapter also presented a book of IEEE and Thai Telecommunications History & Milestones to Prof. Michel Howard, 2014 IEEE President-Elect. For more information please visit us at www.thaitelecomkm.org.

Presentation of the IEEE and Thai Telecommunications History & Milestones book to Prof. Michel Howard, 2014 IEEE President-Elect.

IRAQ COMSOC CHAPTER/*Continued from page 1*

We feel that these children are "Our little angels, and all of our wishes are directed at providing them with shelter and security, to give them happiness so that they can go back to their homes that they see only in their dreams".

The chair of IEEE Iraq ComSoc Chapter and many volunteers have visited many places in and around Babylon City offering any possible assistance. In all these areas, we have thousands of families who have been expelled from their homes in northern and western Iraq or left them in fear. Our regular visits provide a good picture about the status of these families. We have also made major efforts to visit the primary schools. Many of the IEEE Iraq ComSoc chapter members and their families have been collecting "financial support" for these students.

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

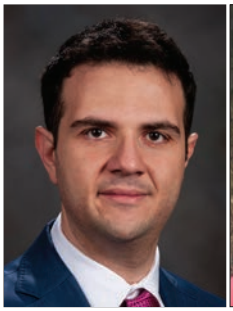
REGIONAL CORRESPONDENTS WHO CONTRIBUTED TO THIS ISSUE

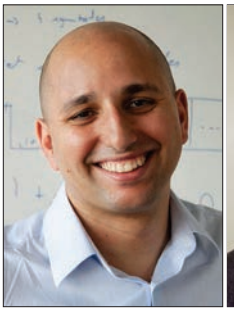
FAWZI BEHMANN (FAWZI.BEHMANN@GMAIL.COM)
EWELL TAN, SINGAPORE (EWELL.TAN@IEEE.ORG)

A publication of the IEEE Communications Society

www.comsoc.org/gcn
ISSN 2374-1082

WIRELESS PHYSICAL LAYER SECURITY: PART 1






Walid Saad *Xiangyun Zhou* *Mérouane Debbah* *H. Vincent Poor*

The ongoing paradigm shift from classical centralized wireless technologies toward distributed large-scale networks such as the Internet of Things has introduced new security challenges that cannot be fully handled via traditional cryptgraphic means. In such emerging wireless environments, devices have limited capabilities and are not controlled by a central control center; thus, the implementation of computationally expensive cryptographic techniques can be challenging. Motivated by these considerations, substantial recent research has been investigating the use of the physical layer as a means to develop low-complexity and effective wireless security mechanisms. Such techniques are grouped under the umbrella of *physical layer security*. These techniques range from information-theoretic security, which exploits channel advantages to thwart eavesdropping, to physical layer fingerprinting techniques that exploit physical layer features for device identification. In this context, providing state-of-the-art tutorials on the various approaches to physical layer security is of considerable interest. This Feature Topic gathers together such tutorial-style and overview articles that provide an in-depth overview of the broad spectrum of security opportunities brought forward by physical layer security.

This Feature Topic is composed of two parts; the second part is expected to appear in the December issue of this magazine. Part 1 begins with an opening editorial by Trappe that exposes the current and future potential of wireless physical layer security. Then, Kapetanovic *et al.* present a novel application of physical layer security: massive multiple-input multiple-output (MIMO) systems. In this article, the authors focus on the robustness of massive MIMO against eavesdropping while also outlining other important related challenges. The next article by Win *et al.* also focuses on secrecy with a particular emphasis on the role of interference. In particular, it discusses how one can engineer interference to ensure confidentiality. Next, the work by Zeng tackles the problem of using the physical layer for key generation. Apart from the passive eavesdropping attack commonly considered in the literature, the author also discusses three types of active attacks and proposes a new key generation scheme to defend against them. The next article by Kailkhura *et al.* describes the security of a distributed inference framework comprising a group of spatially distributed nodes that acquire observations about a phenomenon of interest and transmit computed summary statistics to a fusion center. The authors propose efficient schemes to mitigate the impact of eavesdropping on distributed inference, and survey the currently available approaches along with avenues for future research. This first issue concludes with an article by Yu *et al.* that exposes the importance of physical layer features as a means to fingerprint and authenticate wireless devices.

ACKNOWLEDGMENTS

The Guest Editors would like to thank the large number of people who significantly contributed to this Feature Topic, including the authors, reviewers, and *IEEE Communications Magazine* editorial staff.

BIOGRAPHIES

WALID SAAD [S'07, M'10] (walids@vt.edu) is an assistant professor with the Bradley Department of Electrical and Computer Engineering at Virginia Tech. His research interests include wireless and social networks, game theory, cybersecurity, smart grid, network science, cognitive radio, and self-organizing networks. He is the recipient of the NSF CAREER award in 2013, the AFOSR summer faculty fellowship in 2014, and the ONR Young Investigator Award in 2015, as well as several conference best paper awards.

XIANGYUN ZHOU (xiangyun.zhou@anu.edu.au) is a senior lecturer at the Australian National University (ANU). He received his Ph.D. degree from ANU in 2010. His research interests are in the fields of communication theory and wireless networks. He has a large number of publications in the area of physical layer security, including an edited book, *Physical Layer Security in Wireless Communications* (CRC Press). He serves as an Editor for *IEEE Transactions on Wireless Communications* and *IEEE Communications Letters*.

MÉROUANE DEBBAH [S'01, M'04, SM'08, F'15] (merouane.debbah@huawei.com) is vice-president of the Huawei France R&D center and director of the Mathematical and Algorithmic Sciences Lab. Since 2007, he is also a full professor at Supelec. His research interests lie in fundamental mathematics, algorithms, complex systems analysis and optimization, and information and communication sciences. He is a WWRF Fellow and a member of the academic senate of Paris-Saclay. He is the recipient of several awards such as the Qualcomm Innovation Prize Award.

H. VINCENT POOR [S'72, M'77, SM'82, F'87] (poor@princeton.edu) is with Princeton University, where his interests are in wireless networking and related fields. He is a member of the National Academy of Engineering and the National Academy of Sciences, and a foreign member of the Royal Society. He received the IEEE ComSoc Marconi and Armstrong Awards in 2007 and 2009, respectively, and more recently the 2014 URSI Booker Gold Medal and honorary doctorates from several universities.

The Challenges Facing Physical Layer Security

Wade Trappe

ABSTRACT

There has recently been significant interest in applying the principles of information-theoretical security and signal processing to secure physical layer systems. Although the community has made progress in understanding how the physical layer can support confidentiality and authentication, it is important to realize that there are many important issues that must be addressed if physical layer security is ever to be adopted by real and practical security systems. In this article, I briefly review several different flavors of physical layer security (at least for wireless systems), and then identify aspects (a.k.a. weaknesses) where the foundation for physical layer security needs to be strengthened. I then highlight that the opportunities for applying physical layer security to real systems will be quite rich if the community can overcome these challenges. In the course of the article, I note new directions for the community to investigate, with the objective of keeping physical layer security research targeted at having a practical impact on real systems.

INTRODUCTION

Physical layer security has become an emerging hot topic in wireless systems.[1] At the heart of this enthusiasm is the belief that the physical layer represents a previously untapped resource for enhancing wireless security. In particular, rather than rely solely upon generic higher-layer cryptographic mechanisms, as has been the norm, there is a belief that it is possible to design lower-layer services that support security objectives such as authentication and confidentiality. There are several good surveys and collections that explore the fundamentals of physical layer security [1–3], but briefly these services can be summarized as follows.

AUTHENTICATION/IDENTIFICATION SERVICES

Rather than employ a shared cryptographic authentication key between Alice and Bob, we instead exploit the uniqueness of the Alice-Bob channel relative to the Eve-Bob channel. The uniqueness of the channel between two locations provides a means of uniquely identifying wireless entities [4] or detecting an entity claiming multiple identities [5]. Devices may authenticate themselves based on their ability to produce an appropriate received signal at the recipient.

CONFIDENTIALITY

Confidentiality at the physical layer can generally be broken down into two different classes: dissemination methods that secretly convey information using the properties of the wireless medium, and extraction methods that seek to build secret information from the wireless channel's characteristics. Roughly speaking, for *dissemination* methods, researchers have shown that it is possible to secretly communicate if one can devise ways to ensure that the wireless channel between the correct transmitter and receiver is better than the channel to any illegitimate receiver. *Extraction* methods, which are philosophically similar to authentication methods, seek to use the unique space, time, and frequency characteristics of the wireless channel as the source of shared secret information (e.g., a key) between a transmitter and a receiver.

With all of the excitement surrounding physical layer security, I believe it is important to revisit the intent behind this research area and whether we, as a community, are on track to developing tools that will engender the trust needed for their adoption. Physical layer security is intended to secure real systems, and thus I think we should examine what we can do as a community to ensure that our methods will be warmly received by the broader security community. The purpose behind this critique is to present my opinions as a researcher and pragmatist who has conducted theoretical and systems research in physical layer security, and in doing so help steer the physical layer security community to high-impact topics that will ensure that our research will be integrated into real wireless systems.

BEING CRITICAL: WHAT ARE THE HURDLES?

When considering the hurdles facing physical layer security, it is natural to consider each type of physical layer security separately (authentication, secret dissemination, and key establishment). A quick analysis, however, will reveal that

The author is with Rutgers University.

[1] It must be recognized that there has been recent work in physical layer security for optical systems, but the focus of this discussion is on wireless systems.

0163-6804/15/$25.00 © 2015 IEEE

there will be significant overlap between the challenges that each of these different flavors of physical layer security will face. Therefore, the approach I have taken is to identify the fundamental assumptions on which these different flavors of physical layer security are built, and the potential hurdles that could prevent physical layer security from succeeding. Loosely, I would break these down into assumptions regarding the adversary and assumptions regarding the nature of the wireless channel, recognizing that weaknesses in our assumptions about the wireless channel would naturally be exploited by a clever adversary. Finally, there are practical matters that warrant investigation.

Hurdles from the Adversary Model

The adversary model used in physical layer security tends to be different than what is employed by the security and cryptography community. Consequently, bridging the gap between the different adversary models used by the different communities is perhaps the most important hurdle facing physical layer security. To many in the traditional security community, our models are perceived as weak, and below I elaborate on several aspects related to our adversary models that can be explored as a means to make our work more readily accepted by the broader security community.

The Adversary Is Passive: Many, although certainly not all, formulations assume that the adversary merely eavesdrops on communications. In classical cryptography parlance, such an adversary is a *ciphertext only* adversary. Many of the physical layer key establishment schemes assume that the adversary is merely monitoring the key establishment process and not actively injecting pilot symbols or imitating bit reconciliation messages. Furthermore, we rarely see active attacks like *replay attacks* employed against protocols involving physical layer security. Understanding the implications of an adversary being actively engaged in undermining the protocol will be paramount to getting physical layer security accepted. Modern cryptography recognizes that adversaries may cleverly set up *oracles* that support their cryptanalysis and attempt to undermine the system security. Chosen message attacks (e.g., chosen plaintext, chosen ciphertext, and adaptive versions thereof) are the well accepted starting point for analyzing security tools. As a community, we need to adopt a wider array of adversary modes in which the adversary is more active and, frankly, more clever.

The Adversary Does Not Have Many Observations: Many physical layer security approaches assume an Alice-Bob-Eve model where Eve makes a limited set of observations (e.g., at one extreme, Eve might be a single entity that exists at a single location). For secret dissemination methods, for example, the challenge is that if Eve is a distributed adversary and obtains multiple (say M) independent observations of Alice $\rightarrow$ Bob's single communication, the probability that one of these M channels is better than the Alice $\rightarrow$ Bob channels goes up. Furthermore, Eve could employ collaborative processing, and then the probability that Eve's (combined) effective channel is better than Bob's would go up quite rapidly — in fact, a collaborative adversary with merely a linear factor more observations than Bob can effectively make the secrecy rate 0. The community needs tools to counteract the challenge of adversarial resource advantages, and there has been promising work in leveraging feedback to overcome an adversary with multiple observations [6, 7]. Work needs to be done to ensure that such feedback mechanisms are themselves robust to adversarial manipulation and even impersonation attacks. I feel it is also worth mentioning two aspects of the Dolev-Yao model, which is frequently referred to in the security community: that there are adversaries anywhere they want to be in the network, and that these adversaries may eavesdrop, manipulate, inject, alter, duplicate, and reroute as befits their purpose. While the notion of an omnipresent adversary makes no sense in the context of physical layer security, we must nonetheless strive to strengthen our adversarial considerations in terms of where they are located in the system.

Beyond secrecy dissemination, one should consider how multiple adversaries would impact other forms of physical layer security. For methods that leverage the channel for authentication or key establishment, it must be recognized that the decorrelative properties of the channel are not absolute, and consequently, many observers located simultaneously near Alice and/or Bob may be able to collaboratively estimate the Alice-Bob channel. Understanding the implications of collusion or collaboration on physical layer security will be important to solidify the adversary model.

The Adversary Is Not Powerful: Often one hears a researcher say something along the lines of "physical layer security is based on information-theoretic security, and hence, no matter how much (computing) resources the adversary has, he will not be able to break our scheme." I would certainly agree that a considerable amount of research in physical layer security calls on information-theoretic tools. However, I would not agree that we are operating in the spirit behind information-theoretic security, which is a tool that has been successfully applied to many security problems outside of physical layer security, such as multicast security and key management. Information-theoretic secrecy uses impressive phrases like *perfect* secrecy, which suggest invincibility to resources employed by the adversary, and certainly in Shannon's original theory for secret communications the adversary is assumed to have unlimited computational resources. However, if I were an adversary challenging physical layer security, rather than put my money into buying computation, I would put my money elsewhere, such as buying better antennas. Modern cryptography and security is founded on the notion of an efficiency gap between legitimate users and illegitimate users, or, to put it another way, we expect the adversary to need greater than a polynomial amount of resources (computing, storage, money, or whatever) compared to legitimate parties. We often assume Eve has the same antennas and thus antenna gains as Alice and Bob, and this is worrisome. In fact, Eve's

The adversary model used in physical layer security tends to be different than what is employed by the security and cryptography community. Consequently, bridging the gap between the different adversary models used by the different communities is perhaps the most important hurdle facing physical layer security.

A slightly more sophisticated adversary might be able to employ ray-tracing methods to predict the channel, and an important question to explore is how much the secret key rate is affected by an adversary's ability to perform ray-tracing.

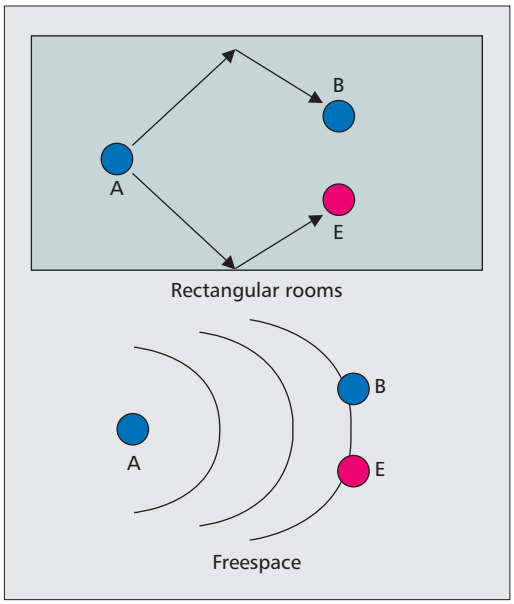


Figure 1. Symmetries inherent in the environmental geometry could pose potential weaknesses to physical layer authentication and key establishment. Two extreme examples include a rectangular room and freespace propagation.

gain is directly related to the aperture of her antenna, which is proportional to its physical aperture; hence, all Eve needs to overcome Alice-Bob is to devote more *area* to listening. Loosely speaking, this means that Eve's resources need only be quadratic in Alice-Bob's resources in order to have a competitive chance at undermining Alice-Bob's secret communication. But to put things in practical terms, a quick survey of antenna gains vs. cost finds gains ranging from 7 dBi to 20 dBi for anywhere between $25 to $200. I would thus suggest the following question: Can a non-nation-state Alice and Bob ever hope to overcome a nation-state Eve?

Hurdles from the Wireless Channel

The Channel Is Difficult for an Adversary to Predict: The security strength for both physical layer authentication and physical layer key establishment is intimately tied to the basic assumption that it is hard for an adversary to estimate or predict the channel. The literature is filled with references to the wide-sense stationary with uncorrelated scatterers (WSSUS) channel assumption and Jakes's well-known uniform scattering model to declare that a received signal rapidly decorrelates over a distance of roughly half a wavelength, and consequently that spatial separation of one to two wavelengths is sufficient for assuming independent fading paths. This decorrelative property is used as the basis to conclude that it is hard for an adversary to estimate the channel that Alice and Bob experience.[2] In actuality, there are several problems with this assumption that warrant discussion. First is quantifying when we are in a sufficiently rich scattering environment. Alice and Bob need to verifiably assess that they are in a richly scattering environment before commencing with physical layer key establishment (or authentication). It must be realized that simple environments are the bane of physical layer security. Two examples bring the severity of this problem to light: freespace and a simple rectangular room (Fig. 1). In freespace, if Bob and Eve are the same distance from Alice, they will experience the same propagation phenomena (notably, just path loss). While in a rectangular room, it is possible to construct benign scenarios where the Alice-Bob channel is the same as the Alice-Eve channel. The community is already examining questions related to the security of the propagation assumption, such as [9], and I expect that the community will continue to find cases where poor environmental scattering undermines the security of key extraction and physical layer authentication.

A slightly more sophisticated adversary might be able to employ ray-tracing methods to predict the channel, and an important question to explore is how much the secret key rate is affected by an adversary's ability to perform ray-tracing. Related to this is the interesting question of how complex the descriptive model for the environment must be in order to undermine physical layer security. I would note that ray-tracing has been used to validate physical layer authentication methods, and there is an amusing paradox here. The very tool used to validate physical layer authentication could also be employed by the adversary to undermine physical layer authentication: Eve could use the same ray-tracing to identify promising locations within a building to conduct spoofing attacks against Alice-Bob. Certainly, this is a matter of concern since many building blueprints are in the public domain, thereby facilitating ray-tracing analysis by an adversary. Developing tools that can estimate an environment's "propagation" complexity in real time will be an important practical tool for supporting physical layer security. Of course, this begs the fundamental question of what the right notion of environmental channel complexity is for physical layer security. It is quite unlikely that the notions of delay spread and the K-factor will carry the appropriate properties needed for physical layer security.

The Channel Needs to Be Dynamic: A similar concern arises when one considers the radio environment temporally; a completely static environment where the scatterers do not move also poses several concerns. Temporal decorrelation of the channel plays a significant role in the non-reciprocity of the channel observations between Alice and Bob, and thus has an important role in how a key establishment algorithm must process a sequence of bidirectional channel probes. In the first case, it is necessary to complete bidirectional probing (Alice $\rightarrow$ Bob and Bob $\rightarrow$ Alice) before the channel decorrelates in order to ensure that Alice and Bob are observing highly correlated observations of the same phenomena. In the second case, temporal coherence plays a role in subsequent Alice-to-Bob probing: either the algorithm must explicitly utilize the correlation to ensure that Alice and Bob arrive at the same decisions (e.g., the Radio Telepathy algorithm of [8]), or else the algorithm must explicitly ensure that subsequent rounds of channel probing have decorrelated to ensure independence in the secret bits that are

[2] I will be the first to admit that I have leveraged this argument myself, c.f. [8]!

established (e.g., as exists in the JRNSO quantization algorithm [10]). If the channel is completely static (coherence time is infinite), there is no "renewal" process, which leads to several problems in key establishment: Alice and Bob might not establish enough bits from a single channel usage; and a key formed at one instance might be the same as a key formed at a later time, and hence Eve can run her own measurements later to estimate Alice's and Bob's shared key. If the channel is somewhat dynamic, the key extraction algorithm must cope with correlated measurements. For example, in a level crossing algorithm, the correlation between subsequent Alice and Bob measurements is the basis for key extraction protocol, while in a quantization-based algorithm, subsequent measurements by Alice and Bob must be independent (e.g., either delayed measurements or by a whitening procedure).

The Channel Is Gaussian (or Symmetric): We often assume that our channels are Gaussian, but how Gaussian is "Gaussian" really when we consider fading? Without symmetry in the fading distribution, physical layer key establishment schemes may suffer from poor distillation if not properly considered. Gaussianity arises in fading from the sum of many independent non-resolvable multipaths. However, as bandwidth increases, resolvability sets in, and the standard application of the central limit theorem begins to fail. Hence, there is a potential problem as we take physical layer key establishment to the wideband regime. Even in non-wideband regimes, we need an assurance that the distribution of key bits coming out of physical layer key establishment is unbiased, which necessitates an assurance that the underlying channel is symmetric. This is a strong requirement, and merely saying that ideal fading is Gaussian and hence symmetric is not sufficient. Being able to reliably quantify and ensure that the channel Alice and Bob are experiencing is close to having a symmetric distribution and quantify the proper amount of privacy amplification is an important problem facing physical layer key establishment, which will necessitate connecting empirical channel estimation to statistical tests that can strongly verify the symmetric nature of the channel. Of course, there is the fundamental question: How close is close enough to symmetric?

The Adversary Cannot Control the Channel: The notion of an active or passive adversary typically has to do with whether the adversary is merely listening or actively injecting communications into the environment. In the context of physical layer methods, we should also consider that the adversary may attempt to manipulate the environment to its advantage. Although unlikely, a powerful attacker would be one that is able to manipulate the amplitude and phases of the signals being exchanged between Alice and Bob in a controlled manner by manipulating the environment. In particular, by manipulating the environment, it is possible to bias the resulting bits in the key establishment process. Such an attack was illustrated in [11] where naïve key establishment that merely uses received signal strength was shown to be able to be manipulated. An interesting question that remains is whether key establishment schemes that are based on complex channel characterizations can be similarly affected and controlled by an adversary.

Integrating Channel Knowledge into Practical Secrecy Dissemination: In secret dissemination, the approach taken to convey secret information depends on the amount of channel state information that is known to the various benign and adversarial participants. There is a considerable amount of work that characterizes secrecy under varying amounts of statistical information available regarding the Alice-Bob and Alice-Eve channel state [3]. A natural next step is to make the explicit connection between how Alice can assess what information she has and how she can adjust her secret communication methods appropriately in an online manner.

Hurdles from Practical Matters

Complementing the above hurdles is an array of practical matters that need to be addressed when moving to a system implementation. Although my intent is to highlight fundamental hurdles that we need to address, I would be remiss if I were not to mention my list of practical hurdles. There are several practical aspects of a transceiver's design that make utilizing channel reciprocity challenging when conducting physical layer key establishment. Beyond the issue of channel coherence time identified earlier, there are other matters that impact the validity of the reciprocity assumption, including properly quantizing channel estimates and non-isotropic noise conditions. Furthermore, there are matters related to calibration and associated amplifier discrepancies, transceiver burn-in, and frequency drift. We must have the appropriate tools in our toolbox to address these challenges, and although I am certain we may borrow methods from conventional communications engineering, I expect the security aspect of the problem will necessitate some new tricks when implementing in real systems. Turning to physical layer authentication, if Alice and Bob have lost their connection for a period of time, the channel no longer supports Bob's verification as the Alice $\rightarrow$Bob channel will have significantly changed. Thus, an important challenge is how to keep authentication going following communication outages. Lastly, in regard to secret dissemination, much of the literature focuses on results under assumptions of Gaussian signaling. This is unrealistic — no practical communication system employs Gaussian signaling, but instead actual transmissions involve discrete constellations, like quadrature amplitude modulation (QAM). For secret communication, discrete signaling behaves quite differently from Gaussian signaling. As an example, in a fast fading scenario, QAM can perform better than Gaussian schemes when Bob's channel is on average worse than Eve's channel as the discrete nature of the signaling effectively limits the information leakage when Eve's channel is better [12–14]. I think we should realize that if discrete signaling gives different conclusions than Gaussian signaling, and discrete constellations are practical, we ought to explore them more in our community's work. Furthermore, I believe that

The notion of an active or passive adversary typically has to do with whether the adversary is merely listening or actively injecting communications into the environment. In the context of physical layer methods, we should also consider that the adversary may attempt to manipulate the environment to its advantage.

Investigating approaches such as ciphers or encoding for physical layer confidentiality that are efficient and have little to no message expansion is a promising direction for investigation that would greatly benefit IoT devices. As a researcher and participant in physical layer security, I am excited to see how these challenges will be addressed in the years ahead.

not only will we learn valuable lessons when going to practical implementations, but we might also find characteristics beneficial in overcoming some of the other hurdles outlined earlier.

Discussion: There Is Light at the End of the Tunnel

I do not want the reader to think that there is only bad news. I do not believe this to be the case. The above discussion has identified hurdles we need to address, and for which I believe the community has more than ample ingenuity to overcome. By comparison, a survey of classical security research will reveal a long list of encryption algorithms and security protocols that have undergone refinement over the years, and this is as natural to security engineering as it is to communications engineering or any other form of research. Furthermore, as I have noted earlier, we must bridge the language gap between our community and the classical security community in order to have the security community adopt our methods. Toward this end, we can revisit many of our approaches, and explore them in the context of semantic security and indistinguishability — this is a relatively low-hanging fruit as some of the formulations we have devised share many of the same properties as desired by the modern cryptography community. As an example, there is a remarkable similarity between probabilistic encryption [15] and codes developed for secret dissemination.

I would also like to offer to the community my belief that physical layer security is the best approach to securing many emerging wireless systems that are not well suited for conventional cryptographic approaches. In particular, we have all heard the rumblings of the Internet of Things (IoT) as a new area of research. I would point out that a significant portion of the IoT will be low-end, low-energy, and lightweight computing devices that will come with real restrictions on how one designs their functions. For these low-end devices, most of the available energy and computation must be devoted to executing core application functionality, and there may be little left over for supporting security. This is where I feel that the physical layer security community can have a notable impact as an ideal approach to securing the low end of the IoT is to dual-use existing radio communication functions for security. For example, physical layer signal processing can be applied at a gateway receiver to authenticate whether a transmission came from the expected IoT transmitter in the expected location. Similarly, one might explore using the physical layer for confidentiality, and along these lines I would note that physical layer secrecy has its own costs; perhaps the most obvious is the message expansion needed for confidentiality. Similar to the message expansion problems associated with probabilistic encryption, a source coding approach to physical layer secrecy typically involves a coded ciphertext that is larger than the plaintext. Since the message is larger, there will be more strain on the energy needed by the system. Investigating approaches such as ciphers or encoding for physical layer confidentiality that are efficient and have little to no message expansion is a promising direction for investigation that would greatly benefit IoT devices. As a researcher and participant in physical layer security, I am excited to see how these challenges will be addressed in the years ahead.

References

[1] R. Liu and W. Trappe, *Securing Wireless Communications at the Physical Layer*, Springer, 2010.
[2] M. Bloch and J. Barros, *Physical-Layer Security: From Information Theory to Security Engineering*, Cambridge, 2011.
[3] Y. Liang, H. V. Poor, and S. Shamai, *Information Theoretic Security*, NOW, 2009.
[4] L. Xiao, L. Greenstein, N. Mandayam, and W. Trappe, "Using the Physical Layer for Wireless Authentication in Time-Variant Channels," *IEEE Trans. Wireless Commun.*, vol. 7, 2008, pp. 2571–79.
[5] L. Xiao *et al.*, "Channel-Based Detection of Sybil Attacks in Wireless Networks," *IEEE Trans. Info. Forensics Security*, vol. 4, 2009, pp. 492–503.
[6] X. He and A. Yener, "Providing Secrecy When the Eavesdropper Channel is Arbitrarily Varying: A Case for Multiple Antennas," *Proc. 2010 48th Annual Allerton Conf. Commun., Control, and Computing*, Sept. 2010, pp. 1228–35.
[7] X. He and A. Yener, "On the Role of Feedback in Two-Way Secure Communication," *Proc. 2008 42nd Asilomar Conf. Signals, Sys. and Computers*, Oct. 2008, pp. 1093–97.
[8] S. Mathur *et al.*, "Radio-Telepathy: Extracting a Cryptographic Key from an Un-Authenticated Wireless Channel," *Proc. 14th ACM Annual Int'l. Conf. Mobile Computing and Networking*, 2008.
[9] X. He *et al.*, "Is Link Signature Dependable for Wireless Security?" *Proc. IEEE INFOCOM*, 2013, April 2013, pp. 200–04.
[10] C. Ye *et al.*, "Information-Theoretically Secret Key Generation for Fading Wireless Channels," *IEEE Trans. Info. Forensics Security*, vol. 5, 2010, pp. 240–54.
[11] S. Jana *et al.*, "On the Effectiveness of Secret Key Extraction from Wireless Signal Strength in Real Environments," *Proc. 15th ACM Annual Int'l. Conf. Mobile Computing and Networking*, 2009, pp. 321–32.
[12] S. Bashar, D. Zhi, and Chengshan Xiao, "On the Secrecy Rate of Multi-Antenna Wiretap Channel Under Finite-Alphabet Input," *IEEE Commun. Lett.*, vol. 15, no. 5, May 2011, pp. 527–29.
[13] S. Basharand, D. Zhi, and Chengshan Xiao, "On Secrecy Rate Analysis of MIMO Wiretap Channels Driven by Finite-Alphabet Input," *IEEE Trans. Commun.*, vol. 60, no. 12, Dec. 2012, pp. 3816–25.
[14] Z. Li, R. Yates, and W. Trappe, "Achieving Secret Communication for Fast Rayleigh Fading Channels," *IEEE Trans. Wireless Commun.*, vol. 9, no. 9, Sept. 2010, pp. 2792–99.
[15] S. Goldwasser and S. Micali, "Probabilistic Encryption," *J. Computer and Sys. Sci.*, vol. 28, 1984, pp. 270–99.

Biographies

Wade Trappe [F] (trappe@winlab.rutgers.edu) is a professor in the Electrical and Computer Engineering Department at Rutgers University, and associate director of the Wireless Information Network Laboratory (WINLAB), where he directs WINLAB's research in wireless security. He has led several federally funded projects in the area of cybersecurity and communication systems, projects involving security and privacy for sensor networks, physical layer security for wireless systems, a security framework for cognitive radios, the development of wireless testbed resources (the ORBIT testbed, www.orbit-lab.org), and new RFID technologies. His experience in network security and wireless spans over 15 years, and he has co-authored a popular textbook in security, *Introduction to Cryptography with Coding Theory*, as well as several monographs on wireless security, including *Securing Wireless Communications at the Physical Layer and Securing Emerging Wireless Systems: Lower-layer Approaches*. He has served as an Editor for *IEEE Transactions on Information Forensics and Security*, *IEEE Signal Processing Magazine*, and *IEEE Transactions on Mobile Computing*. He served as the lead Guest Editor for the September 2011 Special Issue of *Transactions on Information Forensics and Security* on Using the Physical Layer for Securing the Next Generation of Communication Systems and served as the IEEE Signal Processing Society representative to the governing board of IEEE TMC. He is currently IEEE SPS Regional Director for Regions 1-6.

Physical Layer Security for Massive MIMO: An Overview on Passive Eavesdropping and Active Attacks

Dževdan Kapetanović, Gan Zheng, and Fredrik Rusek

ABSTRACT

This article discusses opportunities and challenges of physical layer security integration in MaMIMO systems. Specifically, we first show that MaMIMO itself is robust against passive eavesdropping attacks. We then review a pilot contamination scheme that actively attacks the channel estimation process. This pilot contamination attack not only dramatically reduces the achievable secrecy capacity but is also difficult to detect. We proceed by reviewing some methods from literature that detect active attacks on MaMIMO. The last part of the article surveys the open research problems that we believe are the most important to address in the future and give a few promising directions of research to solve them.

INTRODUCTION

During the recent past the field of MaMIMO systems has quickly emerged as one of the most promising techniques to boost system throughput of emerging and future communication systems. In MaMIMO, the vision is to equip the base station with an antenna array comprising 100+ antenna elements and a large number of independent transceiver chains. An impressive amount of research on the topic has been conducted, and MaMIMO will be integrated in the upcoming 5G standard [1, 2]. The advantages of MaMIMO are manifold, and to give a few we mention:

1. An array gain corresponding to the number of BS antenna elements.
2. A channel hardening effect, rendering stable and predictable channel conditions to users.
3. Nearly orthogonal channels from the BS to the users.
4. Simple signal processing at both the BS and at the users.

Another advantage of MaMIMO, not yet widely recognized, is that the potential of physical layer security (PLS) against passive eavesdropping attacks is increased dramatically. PLS is based on the following result: Consider a Gaussian wiretap channel [3] where a BS communicates with a legitime user (LU) in the presence of a passive eavesdropper (ED). Assume that the Shannon capacities from the BS to the LU and the ED are $\mathcal{C}_{\mathrm{LU}}$ and $\mathcal{C}_{\mathrm{ED}}$, respectively. Then, the *secrecy capacity* between the BS and the LU is $\mathcal{C}_{\mathrm{SC}} = \mathtt{max}\{\mathcal{C}_{\mathrm{LU}} - \mathcal{C}_{\mathrm{ED}}, 0\}$. This rate can be transmitted reliably and securely without any use of a formal crypto system. In conventional MIMO systems, the two capacities, $\mathcal{C}_{\mathrm{LU}}$ and $\mathcal{C}_{\mathrm{ED}}$, are of similar order of magnitude, rendering a fairly small secrecy capacity. With MaMIMO and passive eavesdropping, the situation changes dramatically. With standard time-division duplex (TDD) mode MaMIMO operations, and due to 1 and 2 above, the received signal power at the LU is several orders of magnitude larger than the received signal power at the ED. This generates a situation where the secrecy capacity is nearly the full capacity to the LU, that is, $\mathcal{C}_{\mathrm{SC}} \approx \mathcal{C}_{\mathrm{LU}}$ [4]. Altogether, MaMIMO enables excellent PLS, without any extra effort.

There are of course countermeasures that can be taken by the ED. First of all, it could place itself physically close to the LU so that the channels to the LU and the ED are highly correlated. In this case, 3 no longer holds true, and the secrecy capacity may be compromised. Another effective strategy that the ED can adopt is to exploit the weakness of the channel estimation phase in MaMIMO. By switching from a passive to an active mode, the ED can pretend to be the LU and send a pilot sequence of his own; this is the so called pilot contamination attack. The BS will then beamform signal power to the ED instead of the LU. To detect such a stealthy attack is challenging since there is a normal amount of pilot contamination already present in MaMIMO systems.

Active ED attacks are by no means unique to MaMIMO. On the contrary, jamming the BS is a well researched attack in conventional MIMO. See for example [5, 6] for two recent papers where the ED attacks the channel estimation phase. The ED can also combine passive eavesdropping and active jamming attacks. The strategies for countering such an attack in conventional

Dževdan Kapetanović did this work while at the University of Luxembourg.

Gan Zheng is with the University of Essex.

Fredrik Rusek is with Lund University.

0163-6804/15/$25.00

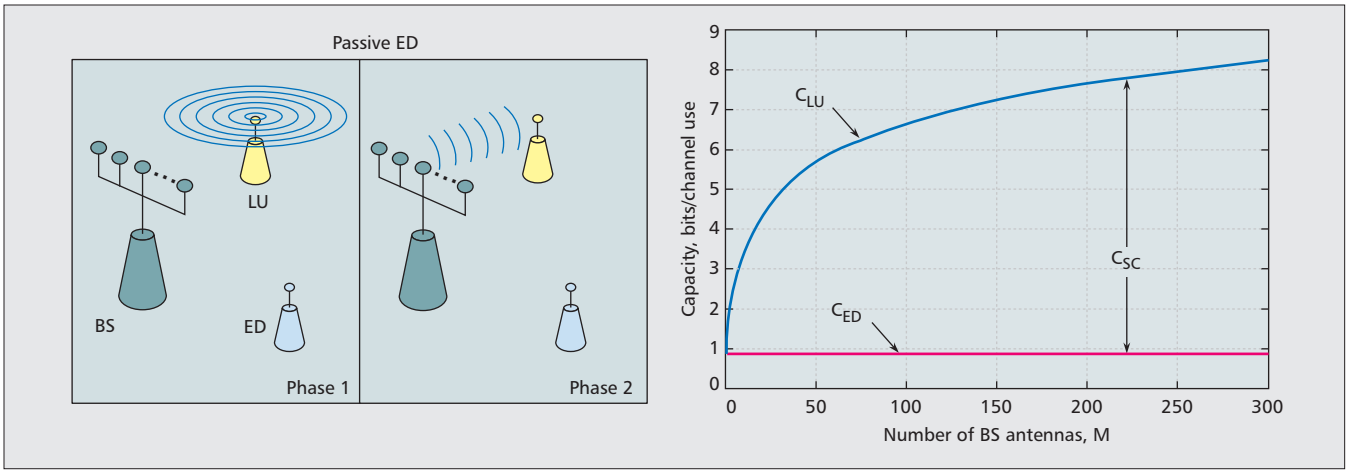


Figure 1. Left: A single-antenna passive ED in a single cell containing a single-antenna LU; Right: Example of secrecy capacity (length of the vertical line). The ED's capacity C_{ED} becomes independent of M. The secrecy capacity increases with M.

MIMO channels are discussed in [7]. A game-theoretic approach is taken in [8] to deal with a combination of passive and active attacks. However, none of the aforementioned papers explicitly dealt with MaMIMO systems, and the inefficiency of passive eavesdropping in MaMIMO was not observed.

Although MaMIMO has received huge attention, the existing literature on the combination of PLS and MaMIMO is scarce. This article will survey the opportunities that MaMIMO may bring for making PLS a reality, as well as discussing problems that must be tackled in the future. The article begins with a discussion of the benefits that MaMIMO brings to PLS in the presence of a passive ED. We then highlight that active attacks are more likely to occur for MaMIMO. Three detection methods to deal with active attacks are then briefly reviewed. We then outline important open problems to address. The article is concluded by a discussion of a few promising directions of future research. We believe that there are many interesting aspects of PLS integration in MaMIMO to be researched and hope that this survey will attract the attention of the research community to this exciting and open field.

Passive and Active Eavesdropping Attacks

For conceptual simplicity, we consider a single cell with an M-antenna BS, one single-antenna LU, and one single-antenna ED. The uplink channels from the LU and the ED to the BS are denoted as g_{LU} and g_{ED}, respectively. We assume a TDD system where channel reciprocity holds, and the corresponding downlink channels are g_{LU}^T and g_{ED}^T, where $(\cdot)^T$ denotes a transpose operation. Standard TDD MaMIMO involves two phases: the LU transmits a training symbol to the BS in the uplink; and relying on channel reciprocity, the BS performs channel estimation and beamforms the signal to the LU in the downlink using the uplink channel estimation $\hat{g}_U$ with proper scaling.

The ED aims to overhear the communication from the BS to the LU while at the same time being undetected. To this end, the ED can launch either a passive attack or an active attack, which will be reviewed below.

Passive Eavesdropping Attack

Let us now discuss the passive attack within a MaMIMO context as shown in Fig. 1. The key observation here is that the presence of a passive ED is not at all affecting the beamforming at the BS and has a negligible effect on the secrecy capacity. Intuitively, this is because MaMIMO has the capability to focus the transmission energy in the direction of the LU. This implies that the received signal strength at the ED is much less than that at the LU. In Fig. 1, the resulting ergodic capacities $\mathcal{C}_{LU}$ and $\mathcal{C}_{ED}$ are shown as functions of the number of BS antennas M with perfect channel estimation. It is assumed that g_{LU} and g_{ED} are independent and identically distributed (i.i.d) complex Gaussian vectors with equal mean powers, and the BS transmit power is normalized to 0 dB.

As can be seen from Fig. 1, the ED's capacity remains the same as M increases; this is so since the BS does not beamform in the ED's direction. However, the capacity to the LU is greatly increased for large values of M. In other words, the secrecy capacity is about half of the LU capacity with conventional MIMO ($M \approx 2 - 8$), while it constitutes more than 85 percent of $\mathcal{C}_{LU}$ already at $M = 100$. From this simple example, we can see that MaMIMO has excellent potential for the integration of PLS.

An Active Attack on Channel Estimation

The resilience of MaMIMO against the passive attack is based on the assumption that the uplink channel estimation $\hat{g}_{LU}$ is independent of the ED's channel g_{ED}. This motivates the ED to design active attacks on the channel estimation process to influence the BS's beamforming design. Next we describe such an attack based on the pilot contamination scheme in [9].

As illustrated in Fig. 2, during the uplink channel estimation, the LU transmits a pilot symbol to the BS. At the same time, the ED launches the attack by sending another pilot

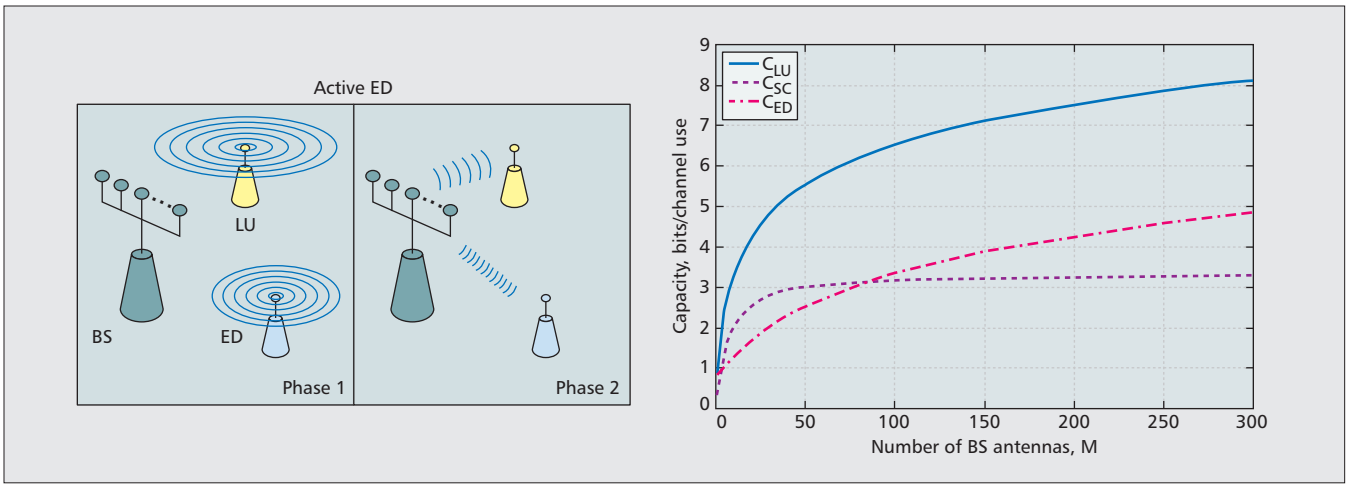


Figure 2. Left: Active attack on the channel estimation; Right: The resulting channel capacities and the secrecy capacity. The ED's training power is 10 dB weaker than the LU's.

symbol. In the worst case the ED is synchronized with the legitimate transmission, and this is possible by overhearing the signaling exchange between the BS and the LU.

The consequence of this attack, if left undetected, is that the promising PLS benefits of MaMIMO are lost. The difference from the passive attack is that the channel estimate $\hat{g}_{LU}$ becomes correlated with the ED's channel g_{ED}, and consequently the equivalent channel for the ED also improves as M increases. Even worse, if the ED uses higher training power, it dominates the training phase and the secrecy capacity may become zero.

For the same settings as in Fig. 1, we plot the ergodic channel capacities and the ergodic secrecy capacity in Fig. 2. The signal sent by the ED is 10 dB weaker than the training signal sent by the LU. As can be seen, both the ED's and the LU's channel capacities increase with M; however, the secrecy capacity remains constant for $M > 50$.

Although the situation is similar to the well-known pilot contamination problem in multi-cell systems [10], a notable difference is that the ED is out of control and therefore existing schemes for reducing pilot contamination in MaMIMO cannot be applied. In the remainder of this article we survey recent progress to detect the active attack and discuss possible future directions of research.

Detection Schemes

As described previously, detection of an active ED is crucial for secure MaMIMO communication. Now we will present arguments that show how peculiar and different ED detection is in MaMIMO systems.

Consider a detection scheme applied by the BS during uplink packet transmission that is based on successful packet reception. In systems such as LTE and WLAN, the uplink packet contains channel estimation pilots. If the ED attacks these pilots, one could argue that this would result in decoding errors and thus packet loss (due to a bad channel estimate). Indeed, this is what happens in conventional MIMO systems with few antennas (depending on the robustness of the used modulation and coding scheme, of course). Therefore, the BS would suspect the presence of strong interference, either coming from an ED or another user, and could take actions based on this. However, in MaMIMO the erroneous channel estimate does not typically result in a decoding error, assuming that ED's channel is uncorrelated with LU's channel. Hence, in contrast to a conventional MIMO system, a successful packet reception does not imply the absence of an ED in a MaMIMO system. As we have argued, if this erroneous channel estimate is left undetected and arises from an active ED, it can have a detrimental effect on the secrecy capacity if used in the subsequent downlink phase. Hence, the importance of effective detection schemes during channel estimation.

An appealing and conceptually simple detection strategy that can be used during channel estimation, which we will also argue against, is to let the BS first estimate the mean power or large scale fading β of the estimated uplink channel vector $\hat{g}_{LU}$, that is, $\beta = \mathbb{E}[\| \hat{g}_{LU} \|^2]$. The value β changes slowly over time/frequency so that a good estimation of it is feasible. The instantaneous received energy of a pilot observation at the BS converges to $\beta P + N_0$ as M grows large, where P is the power of the pilot symbol and N_0 is the noise density at the BS. Thus, if β and N_0 are known, the BS can compare the instantaneous received energy with $\beta P + N_0$. In the absence of an active ED the two quantities are close to each other, while an active ED is detected if the instantaneous received energy is much larger than $\beta P + N_0$.

Unfortunately, there is a simple countermeasure that the ED can take. Since the value of β changes slowly, the ED can adapt its transmit power in order to sabotage the estimation of β at the BS. The ED starts transmitting at low power, and increases the power over several coherence intervals of β. The ED is thereby emulating the natural change of the channel propagation environment, and the BS cannot distinguish the increased received power from a natural channel quality improvement. The lesson learned from this example is that detection

In MaMIMO, the erroneous channel estimate does not typically result in a decoding error, assuming that ED's channel is uncorrelated with LU's channel. Hence, in contrast to a conventional MIMO system, a successful packet reception does not imply absence of an ED in a MaMIMO system.

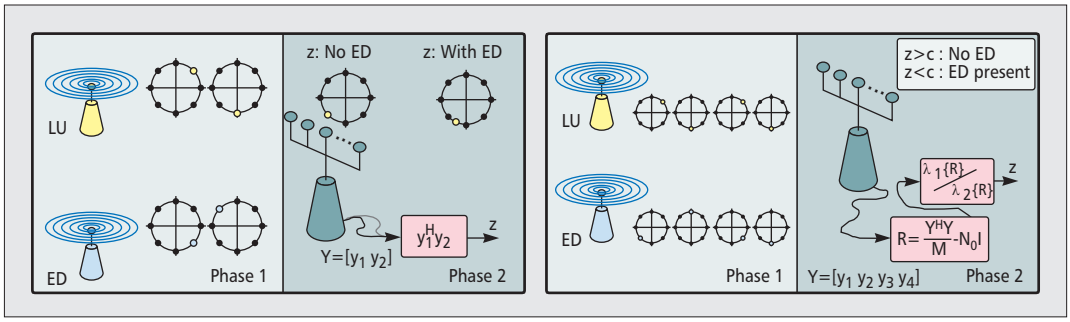


Figure 3. Left: Detection scheme 1a. LU first transmits two random PSK symbols. After processing at the BS side, the correlation of the two received signals should (roughly) be a valid PSK symbol if there is no active ED present. Notice that case 3 is not covered in the figure. Right: Detection scheme 1b. LU transmits (in this case) four random PSK symbols. The BS constructs a correlation matrix, and performs a test based on the ratio of the two largest eigenvalues of the matrix.

methods should not solely depend on the large scale fading parameter β and should preferably work without the knowledge of β. This section will describe two different schemes (one with two flavors) that can effectively detect such an attack in MaMIMO systems without the knowledge of β. Both schemes share two important features, simplicity and effectiveness, that are due to MaMIMO. Although ED's presence during the channel estimation phase removes the expected MaMIMO gain in secrecy capacity, as demonstrated previously, MaMIMO instead enables simple and effective schemes for detection of the ED.

Detection Scheme 1: LU Transmits Random Pilot Symbols

This scheme exploits controlled randomness in transmitting "random" pilots to detect an active ED. As first proposed in [11], the LU transmits a sequence of random phase-shift keying (PSK) symbols, which forms the key to detecting the ED at the BS. Below we discuss two variations of this scheme.

Scheme 1a: Two pilot symbols p_1 and p_2 transmitted by the LU are chosen independently from an N-PSK constellation. The BS receives the two pilot signals $\mathbf{y}_1$ and $\mathbf{y}_2$. The BS now forms the detection statistic z as the phase of $\mathbf{y}_1^H\mathbf{y}_2$, where $(.)^H$ is a conjugate transpose. Three scenarios are possible:

1. The ED is absent during both transmissions. In this case, the phase of z can be shown to converge to a phase of a valid PSK symbol as the number of antennas grows large.
2. The ED is present in both slots. As the number of antennas grows large, z converges with probability $1 - 1/N$ to a number that is not a valid phase of an N-PSK symbol.
3. The ED is present in only one of the slots. In this case, z still converges to a valid PSK phase. However, the received power at the BS will be biased toward a larger value during the slot where the ED is present. Hence, one can form the ratio $q = \|\mathbf{y}_1\|^2/\|\mathbf{y}_2\|^2$. If $q < \gamma_1$ or $q > \gamma_2$, for some thresholds γ_1, γ_2, then it is decided that the ED is present.

From 2, we know that as the antenna number M grows large, the probability of detection converges to $1 - 1/N$. Thus, it can be made arbitrarily close to 1 by increasing the alphabet size N while the false alarm probability converges to 0 due to 1. In order to use a large value of N, M must be quite large ($M > 200$). Note that this simple scheme is effective due to MaMIMO, which provides convergence of the scalar product to different values depending on the ED's presence. Moreover, for large M, this scheme is robust to knowledge of the noise power N_0. Namely, the noise is averaged out in the scalar product between the received signals when M is large.

Scheme 1b: The second random pilot scheme provides improvements to the above scheme in case of three or more observations. Given L observations $\mathbf{y}_1, \ldots, \mathbf{y}_L$, form the matrices

$$\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_L],$$

$$\mathbf{R} = \frac{\mathbf{Y}^H\mathbf{Y}}{M} - N_0\mathrm{I},$$

where $\mathbf{I}$ is an $L \times L$ identity matrix. When M becomes large, $\mathbf{R}$ converges to a rank-one matrix if the ED is absent, otherwise it converges to a full rank matrix with probability $1 - 1/N$. Based on this observation, the following detection rule is applied: if $\lambda_1\{\mathbf{R}\}/\lambda_2\{\mathbf{R}\}$ is greater than some threshold, ED is absent; otherwise it is present, where $\lambda_1\{\mathbf{R}\}$ and $\lambda_2\{\mathbf{R}\}$ are the largest and the second largest eigenvalues of $\mathbf{R}$.

We will see later that this scheme provides significant performance enhancement with only four observations. The scheme also takes care of the case when the ED is not present in all L slots. Again, noteworthy is the simplicity of this scheme, which is due to MaMIMO since it enables convergence of $\mathbf{R}$ to matrices of different rank depending on the ED's presence. Although performing significantly better than scheme 1a, scheme 1b requires a good estimate of N_0, since it is used to construct $\mathbf{R}$.

Detection Scheme 2: Cooperative Detection Scheme

The above random detection incurs the overhead of transmitting additional random sequences. We next briefly introduce a detection method that does not transmit additional pilot symbols [12].

As illustrated in Fig. 4, upon receiving the

training signal from the LU, the BS can apply a beamformer based on the received signal, and then transmit a pilot to the LU using the same beamformer. The beamformer is constructed in such a way that the received sample at the LU in the absence of an active ED equals an agreed value (between the BS and the LU) after a scaling with $1/M$. For simplicity, this agreed value can be taken as 1 in our discussion. In the case that there is an active ED, the LU will observe a much smaller quantity, and this forms our basis to detect the presence of the ED.

As before, simple beamforming is effective due to MaMIMO. Similar to the random pilot scheme 1b, the cooperative scheme also requires a decent estimate of N_0 for proper performance.

Comparison of Detection Schemes

We discuss the pros and cons of the introduced detection schemes and show a comparison of detection performance in Fig. 5 with 200 antennas at the BS. The channel vectors g_{LU} and g_{ED} are independent, with each element drawn from a complex Gaussian distribution with mean 0 and variance 1. The pilot powers from the LU and the ED are also equal. For the random pilot schemes, we use a QPSK alphabet. For the cooperative scheme, it is assumed that the noise power at the BS equals the noise power at the LU.

As shown in Fig. 5, the cooperative scheme performs best at moderate to high SNRs. The dotted curve illustrates the performance of random pilot scheme 1a. Its performance is inferior to the cooperative scheme at moderate to high SNRs, but significantly better at low SNRs. The dashed curve represents random pilot scheme 1b with four received slots, which exhibits very good performance compared to the random pilot scheme with two slots.

As depicted in Fig. 5, the biggest advantage of the cooperative scheme is its detection performance during the two exchange intervals. The drawback with the cooperative scheme is that the ED can cause problems during the whole frame exchange. For example, the ED might be very close to the LU, and thus contaminate any packets that the LU receives from the BS. The noise power at the LU will therefore be significantly higher than the noise power at the BS, causing a degradation in the detection performance at the LU.

The advantage of the random pilot schemes are their robustness to jamming of the LU by the ED, since the detection is performed at the BS. If the ED would increase its power toward the BS, it would instead result in yet easier detection of the ED. However, as illustrated in Fig. 4, the random pilot scheme requires more than two observations of the received signal in order to have a performance comparable to or better than the cooperative scheme.

Further Discussions and Future Directions

In this section we discuss limitations of the solutions introduced above and give promising future directions.

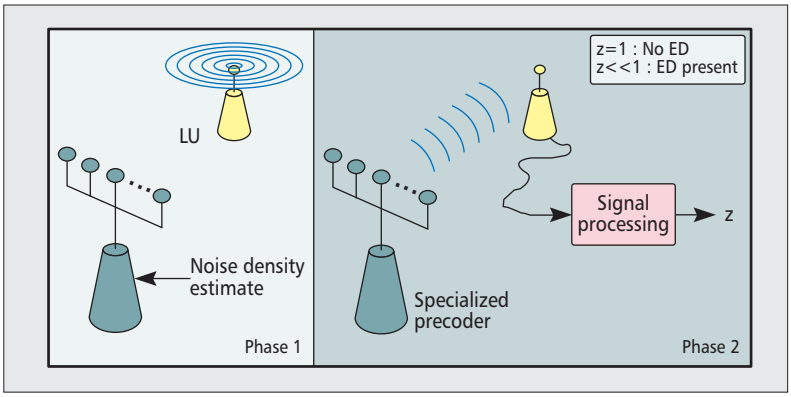


Figure 4. Overview of Detection scheme 2. In phase 1, the BS performs channel estimation. In the next phase, the BS applies a specialized beamformer (see [12]) which ensures that the received signal at the LU after processing becomes 1. If a smaller quantity is observed, there is an active ED present.

Limitations

The reason why only the active attack is effective in MaMIMO is that the channels to the LU and the ED are assumed to be largely uncorrelated. If the channels were correlated, or in the worst case the same, then there is no need for the ED to be active. This means that if the ED can position itself such that its channel to the BS is highly correlated with the channel from the LU to the BS, then the ED can be passive. Furthermore, in line-of-sight scenarios the beamforming at the BS is directional. If the ED is at the very same angle-of-departure from the BS, the ED will receive a highly correlated signal, although it may not be very close. We point out that this problem is alleviated in 3D beamforming scenarios. These considerations are important to take into account in further work to delineate suitable scenarios for combining PLS with MaMIMO.

Detection of Active Attacks in Multi-User and Multi-Cell Systems

We have so far been dealing solely with a single cell and a single user. The case of a multi-user scenario does not alter the situation much as the users can be allocated orthogonal resources for transmitting training symbols. Thus, there is virtually no interaction among users during the training phase. A multi-cell scenario, on the other hand, brings about radical changes. This is so since even without any active ED, the received signal at the BS corresponding to the LU's training signal is being interfered with by LUs from other cells; this is the so called pilot contamination problem of massive MIMO [2]. Consequently, in a multi-cell scenario, the problem of detecting the presence of an active ED changes into the much more challenging problem of distinguishing an active ED from LUs in other cells. In a multi-cell scenario, the above detection methods will all fail since they will detect the presence of the other cells' LUs, but are not sophisticated enough to distinguish these from an active ED.

The multi-cell active ED detection problem is a wide open research problem as, to the best of the authors' knowledge, there are no attempts in the existing literature to deal with it. Although

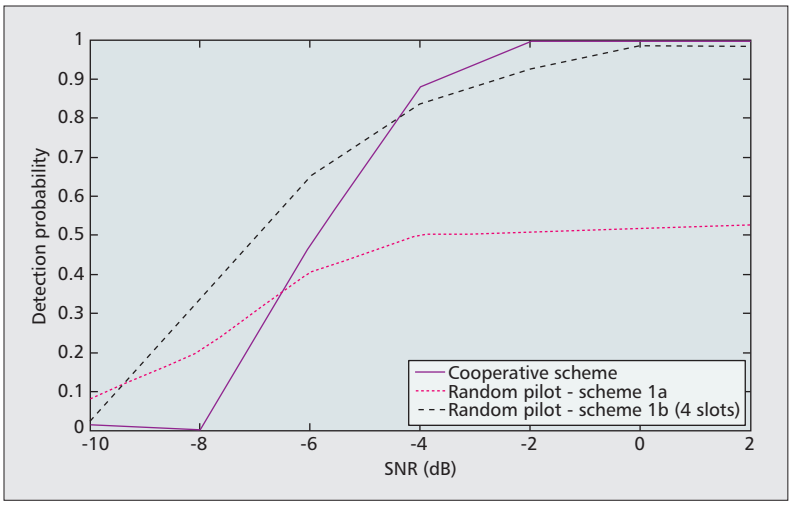


Figure 5. Comparison of the discussed detection schemes. The false alarm probability is 1 percent in all schemes. The random pilot schemes use random QPSK pilots.

not much is known, we briefly introduce two potential ways forward that can be pursued in future research.

Cooperative BSs: Cooperative BSs is a technology that has already found its way into the LTE standard, both in the form of the coordinated multipoint (CoMP) and as the less complex network assisted interference cancellation and suppression (NAICS). In these architectures, the BSs of different cells are connected via a backhaul network and exchange information. Consequently, there is a possibility to let the BSs jointly estimate the level of LU-induced pilot contamination. Even better, the cooperative BSs can jointly try to minimize the pilot contamination in the system. Then the detection methods can be applied verbatim. If the detection methods identify a suspiciously high level of pilot contamination, then the transmission can be terminated. In the case that there was in fact no active ED, but the LU-induced pilot contamination level was unusually high, the terminated transmission is not a major problem as the transmission is not very efficient whenever the pilot contamination is high.

Methods Based on Radio Propagation Characteristics: In a multi-cell case, pilot contamination is caused by LUs from other cells. To be effective, the ED should arguably be located within the serving cell. The radio propagation characteristics between users and a MaMIMO BS are today well understood, and a reasonable working assumption is that the statistical properties of incoming radio waves to the BS from LUs far away are different from those coming from a potential ED that is located much closer to the BS. This is so since signals from users in other cells are typically being reflected by a few dominant objects in the vicinity of the users, and these objects may have line-of-sight to the serving cell BS. Thus, the radio waves' angle-of-arrivals (AoAs) for users in other cells may be limited to a few possibilities, while the signal from a close by ED may reach the BS via interacting objects in the vicinity of the BS. This opens up the possibility to differentiate between pilot contamination from LUs and pilot contamination from an active ED.

The Angle-of-Arrival Database for Location-Aware Users: A particularly interesting direction of research may be to combine the previously outlined approach with the results from [13]. In [13], the authors discuss possibilities of systems where LUs can obtain position estimates of themselves. If the LUs report their positions and signal strengths to a central node, then a database of signal strength as a function of physical position can be constructed. In MaMIMO, this idea can be extended and utilized to detect active EDs. Rather than building a database of signal strengths, a database can contain the AoAs to the BS from LUs at certain positions. The detection of an active ED would then comprise the following steps:

- The LU reports their position to the BS.
- The LU transmits a training symbol.
- The BS requests the AoAs for the particular user position from the database.
- If the measured AoAs for the pilot observation does not match the input from the database, there is an active ED.

So far, the authors are not aware of any attempt in the literature to build an AoA database, and future research is needed to investigate the feasibility of this approach.

Again, if the ED can get physically close to the LU, the measured AoAs may match those in the database even in the presence of an active ED. However, recent measurement campaigns [14] with a 128-port antenna array and where eight users are located within a five-meter diameter circle show that the radio propagation characteristics are sufficiently different for the eight users (both in line-of-sight and non-line-of-sight) so that they can all be spatially separated. This indicates that a MaMIMO system may very well be capable of distinguishing an ED from a LU even when they are physically close.

Detection using Learning Mechanism

Machine learning is a powerful tool that comes naturally to mind for detecting active EDs. Machine learning comes in two forms: supervised learning and unsupervised learning. In supervised learning one must guarantee that there is no active ED present when training the machine. When this is not possible to guarantee, unsupervised learning should be used. However, we can foresee a number of problems that arise that must be dealt with. First, a fairly large amount of training data must be used in machine learning, which may cause unacceptably high overhead. In MaMIMO this is a particularly severe problem as the amount of training may grow with the antenna number. Second, the mobility of users is a problem since the channel quality and characteristics may vary dramatically over time and space.

Alternatively, device dependent radio metrics such as frequency and phase shift differences of radio signals can be used as unique fingerprints to detect active EDs [15]. This method is channel-invariant, but extracting the features is a

complicated task. More effort needs to be made to resolve these problems in order to use machine learning tools for detecting active EDs.

Conclusions

MaMIMO has shown great potential to boost system capacity, but the combination of PLS and MaMIMO is not well understood yet. In this article, we have reviewed both attacks and detection schemes. In particular, we have shown that although passive eavesdropping has little effect on secrecy capacity, an active attack on the channel estimation phase is harmful. We have presented three detection schemes that can effectively identify the active attack. The detection of active attacks in multi-cell and multi-user systems is especially relevant to the applications of MaMIMO, but is challenging and not much is known today. Some promising research directions to this end have been discussed in the article.

The detection of active attacks in multi-cell and multi-user systems is especially relevant to the applications of MaMIMO, but is challenging and not much is known today. Some promising research directions to this end have been discussed in the article.

References

[1] E. G. Larsson *et al.*, "Massive MIMO for Next Generation Wireless Systems," *IEEE Commun. Mag.*, Feb. 2014, pp. 186–95.

[2] F. Rusek *et al.*, "Scaling up MIMO," *IEEE Signal Process. Mag.*, vol. 30, no. 1, Jan. 2013, pp. 40–60.

[3] A. Khisti and G. W. Wornell, "Secure Transmission with Multiple Antennas: The Misome Wiretap Channel," *IEEE Trans. Info. Theory*, vol. 56, no. 7, July 2010, pp. 3088–3104.

[4] J. Zhu, R. Schober, and V. Bhargava, "Secure Transmission In Multi-Cell Massive MIMO Systems," accepted in *IEEE Trans. Wireless Commun*; available online: http://arxiv.org/abs/1405.7161.

[5] R. Miller and W. Trappe, "On the Vulnerabilities of CSI in MIMO Wireless Communication Systems," *IEEE Trans. Mobile Comput.*, vol. 11, no. 8, Aug. 2012, pp. 1386–98.

[6] S. Sodagari and T. C. Clancy, "On Singularity Attacks in MIMO Channels," *Wiley Trans. Emerging Telecommunication Technologies*, May 2013.

[7] A. Mukherjee and A. L. Swindlehurst, "Optimal Strategies for Countering Dual-Threat Jamming/Eavesdropping-Capable Adversaries in MIMO Channels," *IEEE Military Commun. Conf. (MILCOM)*, 2010, pp. 1695–1700.

[8] Q. Zhu *et al.*, "Eavesdropping and Jamming in Next-Generation Wireless Networks: A Game Theoretic Approach," *IEEE Military Commun. Conf. (MILCOM)*, 2011, pp. 119–24.

[9] X. Zhou, B. Maham, and A. Hjørungnes, "Pilot Contamination for Active Eavesdropping," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, Mar. 2012, pp. 903–07.

[10] J. Jose *et al.*, "Pilot Contamination and Precoding in Multi-Cell TDD Systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, Aug. 2011, pp. 2640–51.

[11] D. Kapetanovic *et al.*, "Detecting Pilot Contamination Using Random Training and Massive MIMO," *Proc. IEEE Personal, Indoor and Mobile Radio Communications (PIMRC)*, London, UK, Sept. 8–11, 2013, pp. 13–18.

[12] D. Kapetanovic, A. Al-Nahari, A. Stojanovic, and Fredrik Rusek, "Detection of Active Eavesdroppers in Massive MIMO," *IEEE Personal, Indoor and Mobile Radio Communications (PIMRC)*, Washington DC, Sept. 2-5, 2014.

[13] R. Di Taranto *et al.*, "Location-Aware Communications for 5G Networks," *IEEE Signal Proc. Mag.*, vol. 31, no. 6, Nov., 2014, pp. 102–12.

[14] J. Flordelis *et al.*, "Spatial Separation of Closely-Spaced Users in Measured Massive Multi-User MIMO channels," accepted for publication in *Proc. IEEE Int'l. Conf. Commun. (ICC)*, London, UK, 2015.

[15] N. Nguyen *et al.*, "Device Fingerprinting to Enhance Wireless Security Using Nonparametric Bayesian Method," *Proc. IEEE Conf. Computer Commun. (INFOCOM)*, Shanghai, Apr. 2011.

Biographies

Dževdan Kapetanović received his M.Sc. degree in computer science in 2007, and the Ph.D. degree in electrical engineering in 2012, both from Lund University, Lund, Sweden. From October 2012 to December 2013 he was a research associate at the Security and Trust (SnT) Center, University of Luxembourg, Luxembourg. Since January 2014 he has been an employee at Ericsson Research in Lund, Sweden. His research interests are communication theory and applied information theory.

Gan Zheng is currently a lecturer at the School of Computer Science and Electronic Engineering, University of Essex, UK. He received the B. E. and the M. E. degrees from Tianjin University, Tianjin, China, and a Ph.D. degree in electrical and electronic engineering from The University of Hong Kong, Hong Kong, in 2008. He worked as a research associate at University College London, UK, and the University of Luxembourg, Luxembourg. His research interests include cooperative communications, cognitive radio, physical-layer security, and full-duplex radio. He is the first recipient of the 2013 IEEE Signal Processing Letters Best Paper Award.

Fredrik Rusek received the M.Sc. degree in electrical engineering in 2002 and the Ph.D. degree in digital communication theory in 2007, both from Lund Institute of Technology. In 2007 he joined the Department of Electrical and Information Technology at Lund Institute, where he has held an associate professorship since 2012. He has been employed part time as an algorithm expert at Huawei Technologies, Lund, Sweden, since 2012. His research interests include modulation theory, equalization, wireless communications, and applied information theory.

Multi-Tier Network Secrecy in the Ether

Moe Z. Win, Liangzhong Ruan, Alberto Rabbachin, Yuan Shen, and Andrea Conti

ABSTRACT

Communications in the ether are highly susceptible to eavesdropping due to the broadcast nature of the wireless medium. To improve communication confidentiality in wireless environments, research efforts have been made to complement cryptography with physical layer security. A recent view of the role of interference, especially in multi-tier wireless networks, suggested that interference engineering can increase the level of communication confidentiality. The design of interference engineering strategies (IESs) requires a thorough characterization of concurrent effects of wireless emissions on legitimate and eavesdropping receivers. This article advocates IESs for achieving a new level of communication confidentiality in multi-tier wireless networks (namely multi-tier network secrecy) with different degrees of coordination among the tiers. Insights on how IES benefits wireless network secrecy are provided, guiding the design of such strategies for a new level of communication confidentiality.

INTRODUCTION

Network secrecy is essential for emerging wireless applications that entail the transmission of confidential information. However, single-tier and multi-tier (e.g., cognitive, overlaid, or heterogeneous) wireless networks are intrinsically non-secure due to the broadcast nature of the propagation medium, which allows eavesdropping of transmissions in the ether [1]. While contemporary cryptography-based security is widely used [2], such an approach does not exploit the physical properties of wireless networks and can be complemented by physical layer security [3].

Information-theoretic secrecy was first investigated for a wiretap channel, in which an eavesdropper attempts to intercept legitimate transmission based on its noisy observation [3]. Then it was studied for the case of fading channels [4], interference channels [5], multi-antenna links [6], multilevel networks [7], and eavesdroppers' collusion [8]. The effects of intrinsic network properties, such as node spatial distribution, wireless propagation medium, and aggregate network interference, on secrecy have been assessed in [9]–[11].

Modern wireless networks operate with resource sharing among many users, leading to a large number of interferers. While interference was traditionally considered deleterious, it was shown in [10] that network interference can be exploited to benefit the communication confidentiality (i.e., to create a network secrecy protection).[1] In particular, strong inter-tier interference brought by the overlaid structure was shown to enhance secrecy in multi-tier networks [12]–[14].

To explore the potential of interference exploitation, it is essential to characterize concurrent effects of emissions from legitimate transmitters (LTs) on both eavesdropping receivers (ERs) and legitimate receivers (LRs). Such characterization forms the basis of transmission strategy design so that aggregate interference at the ERs is stronger than that at the LRs, thus enhancing the level of network secrecy (i.e., creating network secrecy protection). Poisson point processes (PPPs) have been widely adopted to study aggregate interference generated by wireless networks with randomly scattered nodes [15]–[17]. In particular, network secrecy protection achieved by means of uncoordinated interference was demonstrated using homogeneous PPPs [10].

We envision that interference engineering strategies (IESs) can increase the secrecy protection in both single-tier and multi-tier networks by a proper coordination of LTs. Various degrees of coordination among LTs can be adopted, depending on the desired level of communication confidentiality and knowledge of the wireless environment. To develop versatile IESs, it is imperative to define network secrecy metrics and characterize the engineered interference in networks with stochastic topology.[2]

The goal of this article is to provide insights into how intrinsic properties of multi-tier networks affect their secrecy and to demonstrate that interference engineering can elevate multi-tier network secrecy. The key elements of this article are as follows:

- The concept of multi-tier network secrecy and the benefit of exploiting network interference for communication confidentiality.
- IESs with different degrees of coordination for the elevation of multi-tier network secrecy.

The proposed approach is based on the cross-fertilization of communication theory, probability theory, and stochastic geometry.

Moe Z. Win and Liangzhong Ruan are with the Massachusetts Institute of Technology.

Alberto Rabbachin was with MIT, and is now with the European Commission.

The views expressed in the article are the sole responsibility of the author and in no way represent the view of the European Commission and its services.

Yuan Shen was with MIT, and is now with Tsinghua University.

Andrea Conti is with the University of Ferrara.

[1] Interference was used to make conversation indecipherable in the Hall Pompeiana of Massimo Theater in Palermo at the end of the nineteenth century by intentionally creating interfering echoes generated by the shape of the hall [10].

[2] IESs were developed and their secrecy advantage was demonstrated using PPPs to model randomly scattered nodes [11].

0163-6804/15/$25.00 © 2015 IEEE

Multi-Tier Network Model

Consider a scenario composed of multiple tiers of legitimate and eavesdropping networks with stochastic topology (Fig. 1). In particular, nodes in the legitimate and eavesdropping networks of the kth tier are distributed in $\mathbb{R}^n$ according to independent homogeneous spatial PPPs $\mathbf{\Pi}_{\text{tx}}^{(k)}$, $\mathbf{\Pi}_{\text{rx}}^{(k)}$, and $\mathbf{\Pi}_{\text{e}}^{(k)}$ for LTs, LRs, and ERs, respectively, with densities $\lambda_{\text{tx}}^{(k)}$, $\lambda_{\text{rx}}^{(k)}$, and $\lambda_{\text{e}}^{(k)}$nodes per unit volume (e.g., nodes per square meter in a two-dimensional scenario).

The interference, generated by spatially scattered nodes reusing the same channel resources, is dominated by emitters in the vicinity of the receiver. Therefore, network interference has a heavy-tail behavior and cannot be modeled by Gaussian distribution. The behavior of network interference is well modeled by the family of *stable* distributions, capturing the effects of important network parameters that depend on transmission power, path loss exponent, fading distribution, and node spatial density [15].

Although the reuse of resources is common in modern networks, interferers are kept outside a region around the receiver. In this case, *truncated-stable* distributions serve as a better model than *stable* distributions [18]. By using the CF of the random processes involved, the statistics of aggregate interference at any given position in the network can be obtained.

Wireless Network Secrecy

The assessment of network secrecy requires the definition of proper metrics that account for the intrinsic properties of all the links in the legitimate and eavesdropping networks. In particular, the network secrecy throughput density (NSTD) τ_{ns} has been defined in [10] as the number of confidential information bits per second per Hertz per unit volume [cib/s/Hz/m$^{\text{n}}$]. It represents the average secrecy throughput originating from a unit volume with a secrecy rate per link of at least $R_{\text{s}}^{(k)}$ [cib/s/Hz] for the kth tier.[3]

The NSTD in the kth tier is given by

$$\tau_{\text{ns}}^{(k)} = \lambda_{\text{tx}}^{(k)} P_{\text{it}}(R_{\ell}^{(k)\star}) R_{\text{s}}^{(k)}$$

where $P_{\text{it}}(R)$ is the probability that a link can support information transmission at rate R, that is, $P_{\text{it}}(R) = \mathbb{P}\{\log\det(\mathbf{I} + \mathbf{S}_k\mathbf{Q}_k^{-1}) \geq R\}$, with $\mathbf{I}$, $\mathbf{S}_k$, and $\mathbf{Q}_k$ denoting the identity matrix, the covariance matrix of the signal at an LR, and the covariance matrix of the interference at an LR, respectively. For a given secrecy rate $R_{\text{s}}^{(k)}$ and maximum tolerable level $P_{so}^{(k)\star}$ of the secrecy outage probability (SOP), the secrecy-protection rate $R_{\ell}^{(k)\star}$ of the legitimate link is given by the solution of the following optimization problem:

$$\mathscr{P}_s: \quad \max_{R} \quad \mathbb{P}\{\log\det(\mathbf{I}+\mathbf{S}_k\mathbf{Q}_k^{-1}) \geq R\}$$

$$\text{s.t.} \quad \mathbb{P}\{R_{\text{e}}^{(k)} > R - R_{\text{s}}^{(k)}\} \leq P_{\text{so}}^{(k)\star}$$

where $R_{\text{e}}^{(k)}$ is a random variable denoting the rate of the strongest eavesdropping link.

Remark: $R_{\ell}^{(k)\star}$ is the rate that a legitimate link can support with maximum probability while guaranteeing the secrecy rate $R_{\text{s}}^{(k)}$ with SOP no greater than $P_{\text{so}}^{(k)\star}$. To determine the SOP, legitimate nodes only need to know the channel statistics (not the instantaneous channel state information) of the eavesdropping links.

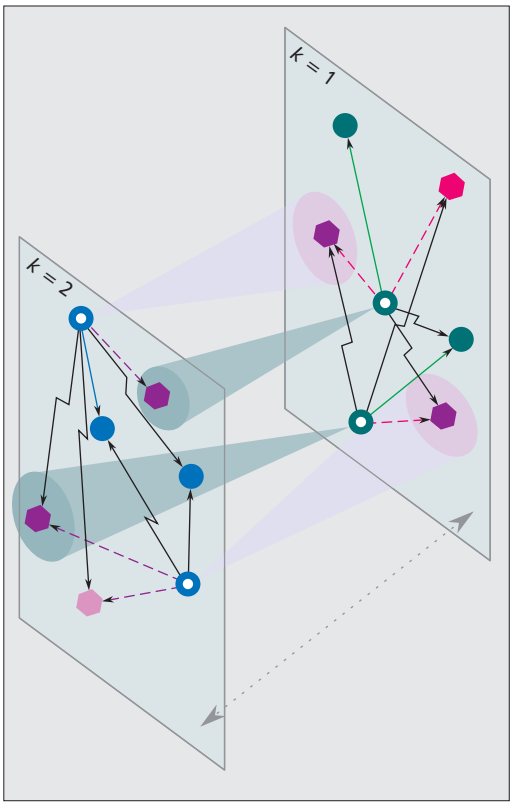


Figure 1. Multi-tier network with primary and secondary transmitters (green and blue empty circles), receivers (green and blue solid circles), and eavesdroppers (red and purple hexagons). Solid, dashed, and jagged lines represent the legitimate, eavesdropping, and interfering links, respectively. Cones represent the interference injected from one tier to another. The bidirectional dotted arrow depicts inter-tier coordination.

Interference Engineering

A key observation is that the broadcast nature of the wireless medium gives rise to opposing effects: on one hand, it makes the confidential information from an LT vulnerable to interception, and on the other hand, it enables other LTs to interfere with the ERs, thereby weakening their interception capability. Therefore, interference can be exploited to elevate multi-tier network secrecy. However, care must be taken in weakening the ERs since the LTs' emissions also affect unintended LRs. This calls for IESs that control the emission of the LTs for network secrecy protection. The IESs can be classified into two main categories, uncoordinated and coordinated, which have different complexity and performance.

Uncoordinated IESs design precoders,[4] separately at each LT, for generating desirable interference for network secrecy protection. These strategies typically require only *local* channel state information (between each LT and its nearby intended and unintended LRs). One

A key observation is that the broadcast nature of the wireless medium gives rise to opposing effects: on one hand, it makes the confidential information from an LT vulnerable to interception, and on the other hand, it enables other LTs to interfere with the ERs, thereby weakening their interception capability.

[3] Two kinds of averaging are involved in secrecy throughput evaluation: one is over all tier-k links with secrecy rate at least $R_{\text{s}}^{(k)}$ originating from LTs in a unit volume (to their corresponding LRs and ERs); the other is over all links originating from all LTs in all tiers that are interfering with those receivers.

[4] The precoder determines the phase and relative amplitude of the signal at each transmitting antenna according to some criteria.

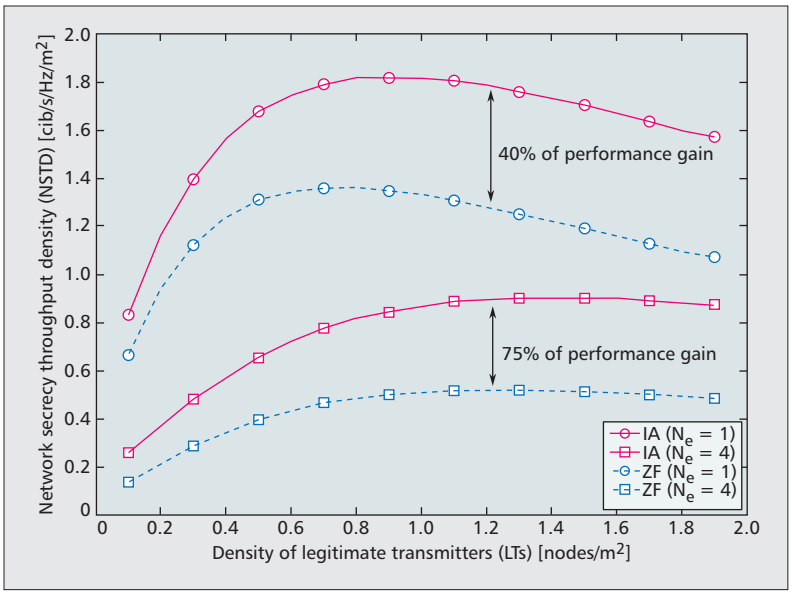


Figure 2. NSTD as a function of the density of the LTs for a single-tier network. The densities of the LTs and ERs are the same. All legitimate nodes are equipped with four antennas. The number of antennas at each ER, N_e, is specified in the figure.

example of uncoordinated strategies is zero-forcing (ZF) beamforming,[5] in which each LT designs the precoder so as to cancel its interference at a number of nearby LRs. In this way, ZF generates strong aggregate interference at the ERs and relatively mild interference at the LRs. Although uncoordinated strategies are easy to implement, they may be insufficient in the presence of strong eavesdropping networks (e.g., the ERs have capability to cancel interference or to collude), calling for coordinated strategies.

Coordinated IESs design precoders, jointly at multiple LTs, for generating sufficient interference to impede strong eavesdropping networks. These strategies typically require *global* channel state information, which is difficult to obtain in large networks. For instance, interference alignment (IA) restricts the interference from multiple LTs to a low-dimensional subspace at the LRs [19], thus reducing the level of aggregate interference at the LRs but not at the ERs. To address the issue of requiring global channel state information, one can either divide a large network into small clusters or design iterative methods to coordinate LTs' emissions into the ether. Nevertheless, coordinated IESs still require higher computational complexity and communication overhead than uncoordinated strategies.

Figure 2 shows the NSTD as a function of the density of LTs for single-tier networks in $\mathbb{R}^2$ with transmission signal-to-noise ratio (SNR) equal to 30 dB at each LT. Spatial IA[6] and ZF are considered with different numbers N_e of receiving antennas at the ERs. It can be observed that the NSTD achieved with IA is significantly higher than that with ZF. Note also that the NSTD decreases with larger N_e, since the ERs' interference cancellation capability is higher when they are equipped with more antennas. On the other hand, the relative performance gain of IA compared to ZF increases from 40 to 75 percent when N_e grows from 1 to 4. This shows the necessity of coordinated strategies in the presence of strong ERs. For a given target level of communication confidentiality, it is important to choose an appropriate IES according to the capabilities of the legitimate and eavesdropping networks.

In a network with multiple tiers, the IESs can also be classified according to the presence or absence of inter-tier coordination. Intuitively, inter-tier coordination designs precoders, jointly among the tiers, thereby providing better network secrecy protection at the cost of additional computational complexity and communication overhead. For instance, consider a cognitive network with primary and secondary tiers. A secondary LT, with channel state information of the links to nearby primary LRs, can design its precoder to avoid interfering with these LRs. At the same time, this secondary LT causes additional interference at the nearby ERs, thus improving the level of communication confidentiality in the primary network.

We now describe four IESs that can be employed by a cognitive network with primary and secondary tiers.

IA-Coordinated Jamming (CJ): The primary LTs and LRs jointly design their precoders and decoders to achieve spatial IA. Then the secondary LTs adjust their precoders according to the channel state information of the links to nearby primary LRs, as well as to the decoders of these receivers for avoiding interference. This strategy requires both inter-tier and intra-tier coordination.[7]

ZF-CJ: The primary LTs separately design their precoders to transmit without creating interference on nearby primary LRs, which adopt arbitrary decoders. Then the secondary LTs adjust their precoders according to the channel state information of the links to nearby primary LRs, as well as to the decoders of these receivers for avoiding interference. This strategy requires only inter-tier coordination.

IA-Uncoordinated Jamming (UJ): The primary LTs and LRs jointly design their precoders and decoders to achieve spatial IA. The secondary LTs adopt precoders independent of the channel state information of the links to primary LRs and of the decoders of these receivers. This strategy requires only intra-tier coordination.

ZF-UJ: The primary LTs separately design their precoders to transmit without creating interference on nearby primary LRs, which adopt arbitrary decoders. The secondary LTs adopt precoders independent of the channel state information of the links to primary LRs and of the decoders of these receivers. This strategy requires no coordination.

Figure 3 shows the NSTD of the primary network as a function of the density of secondary LTs for a cognitive network in $\mathbb{R}^2$ with transmission SNR equal to 30 dB at each LT. The four IESs are employed by primary and secondary nodes. It can be observed that employing an IES with some coordination improves the NSTD with respect to one with no coordination. Note also that when the density of the secondary LTs increases, the NSTD of the two IESs with inter-tier interference coordination increases, while

5 For brevity, ZF beamforming is hereafter referred to as ZF.

6 Spatial IA exploits the finite dimensional signal space provided by the multiple antennas at each node without adopting temporal symbol extension.

7 In this case, the inter-tier coordination is single-directional, that is, the secondary LTs adjust their strategy according to that of the primary users, but not the other way around. Note that this single-directional inter-tier coordination requires only *local* channel state information.

that of the two IESs without inter-tier interference coordination decreases. From a network secrecy perspective, this shows that inter-tier interference coordination changes the relationship of the primary and secondary users from *competitive* to *mutually beneficial*. It can also be observed that the IA-CJ strategy provides the highest level of communication confidentiality as it uses both inter-tier and intra-tier coordination, thus fully exploiting the capability of all LTs to create more interference at the LRs than at the ERs. On the other hand, when the density of the secondary LTs is high, the contribution of inter-tier coordination is significantly larger than that of intra-tier coordination. Hence, in this regime, an IES with ZF-CJ may be sufficient to provide a good trade-off between performance and complexity.

Final Remark

This article demonstrates that interference engineering can enhance the level of secrecy in single-tier and multi-tier wireless networks by reducing the capabilities of eavesdropping networks. The IESs for multi-tier networks can be classified according to the presence of inter-tier or intra-tier coordination. Our analysis reveals how IES can reduce the eavesdropping capability, providing guidelines for the design of efficient IESs with different degrees of coordination. It has also been shown that IESs provide mutually beneficial relationships among the tiers, leading to a new paradigm of multi-tier wireless networks with interference engineering.

Acknowledgment

This research was supported, in part, by the National Science Foundation under Grant CCF-1116501, the MIT Institute for Soldier Nanotechnologies, the Copernicus Fellowship, and the FP7 European project CONCERTO under Grant 288502.

References

[1] M. Z. Win *et al.*, "Cognitive Network Secrecy with Interference Engineering," *IEEE Network*, vol. 28, no. 5, Sept./Oct. 2014, pp. 86–90.
[2] W. Diffie and M. E. Hellman, "New Directions in Cryptography," *IEEE Trans. Inf. Theory*, vol. 22, no. 6, Nov. 1976, pp. 644–52.
[3] A. D. Wyner, "The Wire-Tap Channel," *Bell Sys. Tech. J.*, vol. 54, no. 8, Oct. 1975, pp. 1355–87.
[4] Y. Liang, H. V. Poor, and S. Shamai, "Secure Communication over Fading Channels," *IEEE Trans. Info. Theory*, vol. 54, no. 6, June 2008, pp. 2470–92.
[5] X. Tang *et al.*, "Interference Assisted Secret Communication," *IEEE Trans. Inf. Theory*, vol. 57, no. 5, May 2011, pp. 3153–67.
[6] F. Oggier and B. Hassibi, "The Secrecy Capacity of the MIMO Wiretap Channel," *IEEE Trans. Info. Theory*, vol. 57, no. 8, Aug. 2011, pp. 4961–72.
[7] J. Lee *et al.*, "Distributed Network Secrecy," *IEEE JSAC*, vol. 31, no. 9, Sept. 2013, pp. 1889–1900.
[8] P. C. Pinto, J. O. Barros, and M. Z. Win, "Secure Communication in Stochastic Wireless Networks — Part II: Maximum Rate And Collusion," *IEEE Trans. Inof. Forensics Security*, vol. 7, no. 1, Feb. 2012, pp. 139–47.
[9] X. Zhou *et al.*, "On the Throughput Cost of Physical Layer Security in Decentralized Wireless Networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, Aug. 2011, pp. 2764–75.
[10] A. Rabbachin, A. Conti, and M. Z. Win, "Wireless Network Intrinsic Secrecy," I*EEE/ACM Trans. Net.*, vol. 23, no. 1, Feb. 2015, pp. 56–69.

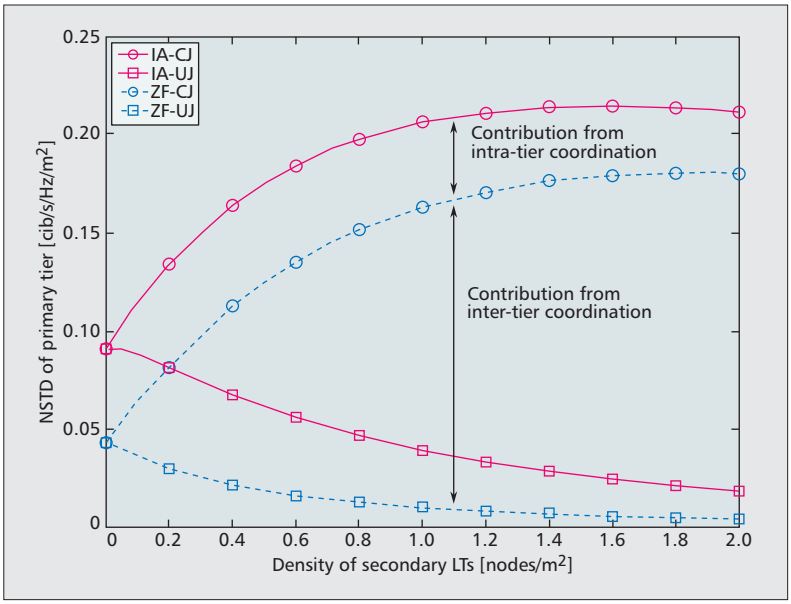


Figure 3. NSTD as a function of the density of the secondary users. The density of the primary LTs, primary LRs, and ERs are all 0.2 nodes/m^2. All legitimate nodes are equipped with four antennas, and each ER is equipped with six antennas.

[11] L. Ruan, V. K. Lau, and M. Z. Win, "Generalized Interference Alignment — Part II: Application to Wireless Secrecy," *IEEE Trans. Signal Processing*, to appear, 2015.
[12] Y. Liang *et al.*, "Capacity of Cognitive Interference Channels with and without Secrecy," *IEEE Trans. Info. Theory*, vol. 55, no. 2, Feb. 2009, pp. 604–19.
[13] Z. Gao *et al.*, "Security and Privacy of Collaborative Spectrum Sensing in Cognitive Radio Networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, Dec. 2012, pp. 106–12.
[14] Z. Ho, E. Jorswieck, and S. Engelmann, "Information Leakage Neutralization for The Multi-Antenna Non-Regenerative Relay-Assisted Multi-Carrier Interference Channel," *IEEE JSAC*, vol. 31, no. 9, Sept. 2013, pp. 1672–85.
[15] M. Z. Win, P. C. Pinto, and L. A. Shepp, "A Mathematical Theory of Network Interference and its Applications," *Proc. IEEE*, vol. 97, no. 2, Feb. 2009.
[16] M. Haenggi *et al.*, "Stochastic Geometry and Random Graphs for the Analysis and Design of Wireless Network," *IEEE JSAC*, vol. 27, no. 7, Sept. 2009, pp. 1029–46.
[17] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic Geometry for Modeling, Analysis, and Design of Multi-Tier and Cognitive Cellular Wireless Networks: A Survey," *IEEE Commun. Surveys and Tutorials*, vol. 15, no. 3, Aug. 2013, pp. 996–1019.
[18] A. Rabbachin *et al.*, "Cognitive Network Interference," *IEEE JSAC*, vol. 29, no. 2, Feb. 2011, pp. 480–93.
[19] L. Ruan, V. K. Lau, and M. Z. Win, "The Feasibility Conditions for Interference Alignment in MIMO Networks," *IEEE Trans. Signal Processing*, vol. 61, no. 8, Apr. 2013, pp. 2066–77.

Biographies

Moe Z. Win [S'85, M'87, SM'97, F'04] (moewin@mit.edu) is a professor at the Massachusetts Institute of Technology (MIT). Prior to joining MIT, he was with AT&T Research Laboratories for five years and with the Jet Propulsion Laboratory for seven years. His current research topics include network localization and navigation, network interference exploitation, intrinsic wireless secrecy, adaptive diversity techniques, and ultra-wide bandwidth systems. He is an elected Fellow of the AAAS and the IET, and was an IEEE Distinguished Lecturer.

Liangzhong Ruan [S'10, M'14] (lruan@mit.edu) received a Ph.D. degree in EECS from Hong Kong University of Science and Technology in 2013., and a B.E. degree in EE from Tsinghua University in 2007. Between 2012 and 2013, he was with MIT as a visiting graduate student. He is currently a postdoctoral associate at MIT. His research interests include interference management, intrinsic wireless secrecy, and quantum entanglement distillation.

ALBERTO RABBACHIN [S'03, M'07] (a.rabbachin@ieee.org) was a postdoctoral fellow at MIT. He is now a project officer at the European Commission. His research interests involve communication theory and stochastic geometry applied to real-problems in wireless networks including network secrecy, cognitive radio, and interference exploitation. He serves as an Editor for *IEEE Communications Letters*. He received the International Outgoing Marie Curie Fellowship, the European Commission JRC best young scientist award, and the IEEE William R. Bennett Prize in the Field of Communications Networking.

YUAN SHEN (S'05-M'14) (shenyuan_ee@tsinghua.edu.cn) received a Ph.D. degree and an S.M. degree in EECS from MIT in 2014 and 2008, respectively, and a B.E. degree in EE from Tsinghua University in 2005. He is currently an associate professor with the EE Department at Tsinghua University. His research focuses on network localization and navigation, inference techniques, and intrinsic wireless secrecy. He was a recipient of the Marconi Society Young Scholar Award and IEEE Communications Society Fred W. Ellersick Prize. He is Secretary (2015–2017) for the IEEE Communications Society Radio Communications Committee.

ANDREA CONTI [S'99, M'01, SM'11] (a.conti@ieee.org) is an associate professor at the University of Ferrara. His research interests involve theory and experimentation of wireless systems and networks including network localization, adaptive diversity communications, cooperative relaying techniques, and network secrecy. He has been elected Chair of the IEEE Communications Society Radio Communications Committee and is an IEEE Distinguished Lecturer. He received the HTE Puskás Tivadar Medal, the IEEE Fred W. Ellersick Prize, and the IEEE Stephen O. Rice Prize in the Field of Communications Theory.

Physical Layer Key Generation in Wireless Networks: Challenges and Opportunities

Kai Zeng

ABSTRACT

Physical layer key generation that exploits reciprocity and randomness of wireless fading channels has attracted considerable research attention in recent years. Although theoretical study has shown its potential to generate information-theoretic secure keys, great challenges remain when transforming the theory into practice. This article provides an overview of the physical layer key generation process and discusses its practical challenges. Different passive and active attacks are analyzed and evaluated through numerical study. A new key generation scheme using random probing signals, and combining user generated randomness and channel randomness, is introduced as a countermeasure against active attacks. The numerical results show that the proposed scheme achieves higher security strength than existing schemes using constant probing signals under active attacks. Future research topics on physical layer key generation are discussed.

INTRODUCTION

With the proliferation of the Internet of Things (IoT), diversified wireless devices need to establish secure communications on the fly. One common way to secure the communication between wireless devices is to generate a symmetric key between them and use it to encrypt/decrypt the message. One conventional mechanism to generate a shared secret key between two parties is the Diffie-Hellman (D-H) key exchange protocol. However, the computation overhead of D-H protocol is significant due to expensive exponential operation, which is undesirable for resource constrained devices such as embedded sensors, wearable devices, RFIDs, and so on. Furthermore, with the ever increasing computing power of attackers, D-H protocol has to increase the key length in order to maintain a certain level of security strength, which in turn aggravates the computation overhead.

An alternative way to generate a shared secret key between wireless devices is to exploit the reciprocity of the random fading channel [1–5]. This mechanism is generally called physical layer key generation, in which wireless devices measure highly correlated wireless channel characteristics (e.g., channel impulse responses or received signal strengths) and use them as shared random sources to generate a shared key. In theory, in a rich multipath scattering environment, a passive attacker who is more than a half-wavelength away from the legitimate users will obtain uncorrelated channel measurements, and thus cannot infer much information about the generated key. The physical layer key generation mechanisms do not require expensive computation and have the potential to achieve information-theoretic security, in the sense that the secrecy of the generated key is not dependent on the hardness of a computational problem but relies on the physical laws of the wireless fading channels.

Due to its attractive features of lightweight and information-theoretic security, physical layer key generation has gained considerable attention in recent years. A typical key generation process includes channel probing, randomness extraction, quantization, reconciliation, and privacy amplification. Although theoretical study provides guideline on designing physical layer key agreement protocols, there are still significant challenges remaining to achieve an efficient and secure-proven key generation scheme in practice. The major challenges lie in the difficulty of measuring the information leaked to eavesdroppers, tackling channel measurement correlations, reducing reconciliation overhead, and deciding on the compression ratio in the privacy amplification stage.

In this article, we give an overview of the general key generation process, and discuss its practical challenges and possible solutions under passive attacks in the following section. The active attacks against physical layer key generation are summarized after that, and a new key generation scheme using random probing signals and combining user generated randomness and channel randomness is presented. The security strength of the proposed key generation scheme

The author is with George Mason University.

0163-6804/15/$25.00 © 2015 IEEE

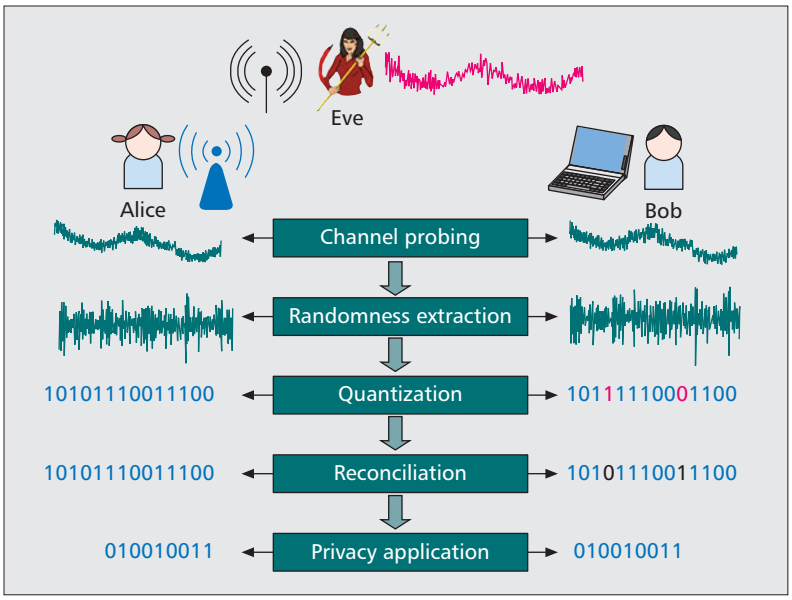


Figure 1. Secret key generation model and steps under passive attack.

and the traditional one using constant probing signals are analyzed and compared under both passive and various active attacks. We then discuss future research directions of physical layer key generation. Conclusions are given in the final section.

Physical Layer Key Generation under Passive Attacks

We first introduce the key generation process, and discuss the practical challenges and possible solutions under passive attacks. We illustrate an application scenario in Fig. 1 where two legitimate users, Alice (an access point) and Bob (a laptop), aim to generate a shared secret key using channel measurements. There is a passive attacker, Eve, who can overhear all the transmissions from Alice and Bob.

Key Generation Primitives

Alice and Bob generally apply the following five steps, illustrated in Fig. 1, to generate a key: channel probing, randomness extraction, quantization, information reconciliation, and privacy amplification [2, 6].

Channel Probing: This is used to collect channel measurements by Alice and Bob. The channel measurements can be channel state information (CSI), received signal strength (RSS), or phase. In this step, Alice and Bob exchange channel probing signals with each other. One channel probing contains a pair of bidirectional channel probing with a short lag of time assuming a half-duplex radio. The received signals are usually modeled as the transmitted sounding signal timing (in the frequency domain) channel gain plus noise. Alice and Bob observe highly correlated received signals due to channel reciprocity.

Randomness Extraction: The received signals at Alice and Bob may contain deterministic parts that can be determined or inferred by the attacker. For example, in Fig. 1, the received signals at Alice and Bob have the same fluctuation pattern on a large scale. This fluctuation is determined by the distance between Alice and Bob. If Eve is close to one of them, she will also observe this large-scale change. Therefore, Alice and Bob should not use this large-scale component to generate shared keys. Otherwise, the key will be easily determined by the attacker. Alice and Bob need to extract randomness caused by channel fading to generate shared keys by removing the large-scale component. A moving window average method can be used to extract the small-scale randomness [6].

Quantization: This is used to quantize the extracted random channel measurements into bits.

Information Reconciliation: This is a form of error correction carried out between Alice and Bob in order to ensure that the keys generated separately on both sides are identical. Due to imperfect reciprocity, the extracted bits at Alice and Bob sides after quantization are usually not identical, although they may be highly similar. This imperfection mainly comes from the fact that Alice and Bob cannot measure the channel at the same time due to the half-duplex property of the radio. Furthermore, the noises at Alice's and Bob's sides are usually independent. During reconciliation, parity bit information may be exchanged to correct errors, and a certain amount of bit information will be revealed to Eve.

Privacy Amplification: This is a method for eliminating Eve's partial information about the key and the correlation among the bits. Eve's partial information comes from eavesdropping during both the probing and reconciliation phases.

Although there has already been intensive study of physical layer key generation under passive attacks [2–4, 6, 7], significant challenges and open issues remain to design an efficient and security-proven key generation scheme in practice. In the following subsections, we discuss these practical challenges and possible solutions.

Difficulty in Estimating Leaked Information in Practice

The secret key capacity is defined as the conditional mutual information between Alice and Bob given Eve's observation [7]. In theory, we can compute various bounds of key capacity given the assumption of knowing eavesdropping CSI. However, in practice, it is very hard to estimate how much information is leaked to a passive eavesdropper. Experimental work has demonstrated that there is a strong correlation in measurements observed by passive eavesdroppers located significantly greater than a half-wavelength away from legitimate devices [8]. It may be due to a poor multipath scattering environment or interference. Therefore, there is not a clear safeguard distance to ensure the secrecy of the generated key. Furthermore, it is hard to know the locations or number of passive eavesdroppers in practice, which introduces difficulty in estimating the leaked information.

This is probably the most critical open issue

hindering the design of a security-proven physical layer key generation scheme in practice. More investigation is needed to address this issue.

One possible solution can be adding noise into the channel to jam the eavesdroppers, creating ambient erasure channels to the eavesdropper [9]. How to guarantee a predictable limit amount of information leaked to eavesdroppers deserves further study.

Reconciliation Overhead

The reconciliation overhead can be significant if the bit agreement ratio is low before the reconciliation. The cause of bit disagreement comes from the imperfect reciprocity of the measured channel characteristics, which is mainly caused by the time lag between the bidirectional channel measurements in the channel probing phase. The reconciliation process is essentially an error correction process with information exchange between Alice and Bob. Existing approaches include the Cascade algorithm and low density parity check (LDPC). It has been shown that if Cascade is used to reconcile two bitstrings having a 10 percent bit mismatch, the number of exposed bits can be around 60 percent.

In order to minimize the reconciliation overhead, it is very important to achieve a high bit agreement ratio before reconciliation. One straightforward way to achieve a high bit agreement ratio is to reduce the time lag between bidirectional channel probings. However, this time lag is constrained by the antenna tx/rx turnaround time and the medium access control protocol. Giving higher access priority to the probing frames can be helpful.

A second way to reduce the bit disagreement is to use lower-level quantization at the cost of reduced key length. It has been found that lower-level quantization is more robust against noise, but the key length is reduced significantly [6].

A third way to reduce the reconciliation cost is to preprocess the measured data. A low-pass Savitzky-Golay filter is applied to process the measured RSS indicator (RSSI) traces to reduce the maximum frequency of changes in received signal power arising from motion in a small-scale fading environment [3]. A fractional interpolation filtering mechanism is also proposed to enhance the bit agreement ratio [5]. A Farrow filter is used to estimate the channel measurement at simultaneous instants halfway between the original non-simultaneous measurements.

Spatial and Temporal Correlation

In practice, there are always spatial and temporal correlation between the channel measurements, which will lead to correlated bits in the generated key. Before quantization, we need to de-correlate the channel measurements. Existing solutions include applying discrete Karhunen-Loeve transform (KLT) to convert the measured channel vectors into uncorrelated components [3]. Note that uncorrelated is not the same as independent. Although it holds for Gaussian random vectors, they are not equivalent in general. The KLT guarantees zero covariance between transformed elements, but not higher-order cross-moments. A sophisticated attacker may utilize higher-order cross-moments to predict partial bits in the key.

Privacy Amplification

In the privacy amplification phase, Alice and Bob compress the bitstrings obtained after reconciliation to their real entropy. A universal hash function can be applied. However, in practice, it is nontrivial to decide the compression ratio. As discussed previously, it is very hard to accurately estimate how much information is leaked to eavesdroppers during the channel probing phase. Without knowing the leaked information, it is difficult to decide the compression ratio. Furthermore, due to spatial and temporal correlation of the channel measurements, the generated bits may have inherent correlations, which reduce the entropy of the generated key. In order to estimate the entropy of a bitstring, we usually need a large number of bits, which may not be obtainable in practice. Consider that when generating a 128-bit key, there are 2^{128} possible permutations of the bits. A true random bit generator should generate each of the 2^{128} permutations with equal probability. To estimate the entropy of a finite short bitstring, we may apply the concept of approximate entropy and use the measure of Lempel-Ziv complexity [4].

The reconciliation overhead can be significant if the bit agreement ratio is low before the reconciliation. The cause of bit disagreement comes from the imperfect reciprocity of the measured channel characteristics, which is mainly caused by the time lag between the bidirectional channel measurements in the channel probing phase.

Physical Layer Key Generation Under Active Attacks

Existing works on physical layer key generation mainly focus on security analysis and protocol design under passive attacks. However, the study of physical layer key agreement techniques under active attacks is largely open.

The existing active attacks can be classified into three categories:

- A disruptive jamming (DJ) attack, which aims to disrupt the key generation process and reduce the key generation rate of legitimate users
- A manipulative jamming (MJ) attack, which injects a signal to manipulate the channel measurements and subsequently compromise a portion of the key
- A channel manipulation (CM) attack, which aims to control the wireless channel between Alice and Bob, thus allowing the attacker to infer the generated key

Disruptive Jamming Attack

Disruptive jamming attacks are proposed against physical layer key generation to disrupt the channel probing process [10]. The attacker injects a jamming signal into the channel in order to minimize the key generation rate between legitimate users.

The injected signals affect the received signals at Alice and Bob, which will introduce inconsistency of the channel measurements, thus reducing the key generation rate. As found in [10], under active attacks, key generation efficiency degrades rapidly with adversarial signal power and external signal interference.

How to achieve a robust secret key generation scheme under active attack is still an open

The methodology of integrating user-introduced randomness and channel reciprocity to generate keys is analogous to D-H key exchange. The independent random signals generated at both sides are analogous to the random numbers generated in D-H protocol.

issue. Existing anti-jamming mechanisms such as channel hopping and spread spectrum frequency hopping may be applied to alleviate the jamming effect. It will also be helpful if Alice and Bob can determine or detect which probing frames are contaminated by the jamming signal and discard those contaminated frames or quit the key generation process.

Manipulative Jamming Attack

Instead of disrupting the channel probing process and reduce the key generation rate, a manipulative jamming attacker aims to largely control the channel measurements at legitimate users, thus compromising the generated key.

A manipulative attack is introduced in [11]. To avoid key disagreement, which may lead to attack detection, the attacker waits for injection opportunities when it detects similar RSSs from Alice and Bob, which indicates similar channel gains in the attacking channels. Then the attacker emits a random attacking signal into the channel, which will cause similar RSSs at Alice and Bob. This attack requires the destruction of the legitimate probing frames sent by Alice and Bob by reactive jamming. Under this attack, a portion of received signals at Alice and Bob are controlled by the active attacker. Although Alice and Bob may achieve a high key generation rate, a portion of the shared key bits are compromised without Alice's and Bob's awareness. It is considered a more adverse case than a disruptive jamming attack.

Channel Manipulation Attack

A smart attacker can also try to manipulate and control the channel between Alice and Bob, and largely control the generated key. It was demonstrated in [2] that an adversary can use planned movements in a relative static environment causing desired and predictable changes in the channel between Alice and Bob. For instance, when there is a line-of-sight path between Alice and Bob, an attacker in the middle can block this path to cause an RSS drop at Alice's and Bob's sides. When the attacker moves away, the RSS increases. Then an attacker can randomly block and unblock the line-of-sight path between Alice and Bob to cause a random RSS drop and increase, thus controlling the channel variations and making the generated key predictable.

One possible way to avoid this attack is to run the RSS-measurement-based key extraction scheme only in rich multipath environments where multiple random moving objects are present so that the attacker's movement alone will not be able to change the channel predictably.

Defending against Active Attacks

In this subsection, we present a new key generation scheme to defend against the above mentioned active attacks. The fundamental reason why these active jamming attacks can be successful lies in the fact that Alice and Bob only use the channel measurements to generate the key. If an attacker can manipulate the channel measurements, she can manipulate or infer the generated key. We propose to integrate user-generated randomness into the channel probing, and generate a shared key based on channel reciprocity and combination of user generated randomness and channel randomness.

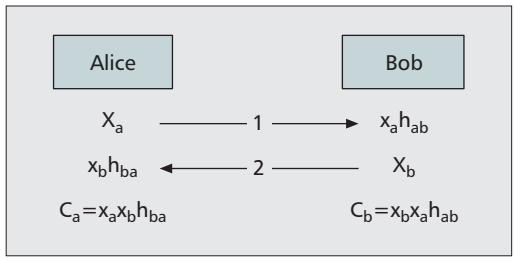


Figure 2. Secret key generation with user-introduced randomness.

The basic idea is illustrated in Fig. 2. When probing the channel, instead of transmitting a constant sounding signal, Alice and Bob transmit independent random signals x_a and x_b. For easy understanding, we ignore the noise in the following explanation.

When there is no attack, Alice and Bob receive $x_b h_{ba}$ and $x_a h_{ab}$, respectively. Note that since x_a and x_b are random signals, neither Alice nor Bob can decode this signal or estimate the channel. At both ends, Alice and Bob multiply the received signal by the local generated random signal to compose a shared randomness. Alice obtains $C_a = x_a x_b h_{ba}$, and Bob obtains $C_b = x_b x_a h_{ab}$. If reciprocity holds, say $h = h_{ab} = h_{ba}$, $C_a = C_b$, which can be used to generate a shared key.

Under the manipulative jamming attack, assuming the original channel probing signal is destroyed by the attacker, what Alice and Bob receive will be the attacking signals with the same power. Alice and Bob will multiply locally generated random signals by the received attacking signal. Alice and Bob can still generate keys from the measurements when $x_a = x_b$ at a 50 percent chance if x_a and x_b are binary. The generated keys, however, will not be compromised by Eve since she has no knowledge of x_a or x_b.

Under the channel manipulation attack, although the attacker can control the channel h, it will not lead to a compromised key. Under this attack, Alice and Bob are still able to generate a shared key based on $x_a x_b h$, which is different from h.

The methodology of integrating user-introduced randomness and channel reciprocity to generate keys is analogous to D-H key exchange. The independent random signals generated at both sides are analogous to the random numbers generated in D-H protocol. Users multiplying received signals by locally generated random signals to achieve agreement is analogous to users conducting discrete exponential operation on the received numbers with locally generated random numbers in D-H protocol. Note that all four steps after channel probing illustrated in Fig. 1 are still needed in order to generate a shared key for the proposed method.

Security Analysis under Passive and Active Attacks

We now compare the security strength of the traditional key generation scheme that uses constant probing signals with the proposed one that

exploits random probing signals under both passive and active attacks. Two metrics are used:

- *Key generation rate* measuring how much secret mutual information is shared between Alice and Bob given the information obtained by Eve
- *Leaked information* measuring mutual information among Alice's received signals, Bob's received signals, and Eve's overheard (for passive attacks) or injected signals (for active attacks)

For passive attacks, we assume a case when Eve is close to Alice. We set the correlation coefficient between legitimate channel and eavesdropping channel as 0.9. For disruptive jamming cases, we assume the attacker injects a constant signal. For manipulative jamming cases, we assume that the attacker always chooses the best attacking moment when the attacking channels have equal gain. For channel manipulation cases, we assume that the attacker can control the channel coefficient h.

For simplicity, we assume that the channel measurements are independent and channel reciprocity holds. We only consider the in-phase component of signals. We assume that all the channel coefficients follow zero mean Gaussian with variance of 10. All the measurement noises follow zero mean Gaussian with unit variance. Therefore, the SNR of each channel is 10 dB by default. For passive attacks, we examine the performance under different SNRs of the legitimate channel h. For active attacks, we fix the SNR of the legitimate channel as 10 dB and change the power (SNR) of the attacking signals.

The key generation rate and leaked information for different cases are shown in Figs. 3 and 4. Each point on the curve represents the numerical result with 10^6 samples. P represents a passive attack, DJ represents a disruptive jamming attack, MJ represents a manipulative jamming attack, CM represents a channel manipulation attack, C represents a key generation scheme with constant probing frame, and R represents key generation scheme integrating user-generated randomness.

From Fig. 3, we can see that under passive attacks, the proposed scheme using random probing signals achieves a similar key generation rate to the scheme using constant probing signals.

Under disruptive jamming attack, the proposed scheme achieves a much higher key generation rate than does the constant probing scheme. An interesting observation is that under disruptive jamming attack, when the SNR of the attacking signal is increased from 0 to 10 dB, the key generation rate of DJ-R is decreased, while when it keeps increasing, the key generation rate is increased. The reason behind this observation is that when the attacking signal is weak, the legitimate channel gain dominates the channel measurement. Thus, the attacking signal is considered noise, which will degrade the key generation rate. When the attacking signal becomes stronger, it dominates the channel measurement and makes channel measurements at Alice and Bob more similar, which leads to a higher key generation rate.

It can be seen that the key generation rate is nearly zero under all the active jamming attacks when constant probing signals are used. However, under active attacks, the proposed scheme achieves a much higher key generation rate. The key generation rate of CM-R is comparable to that achieved under P-R and P-C.

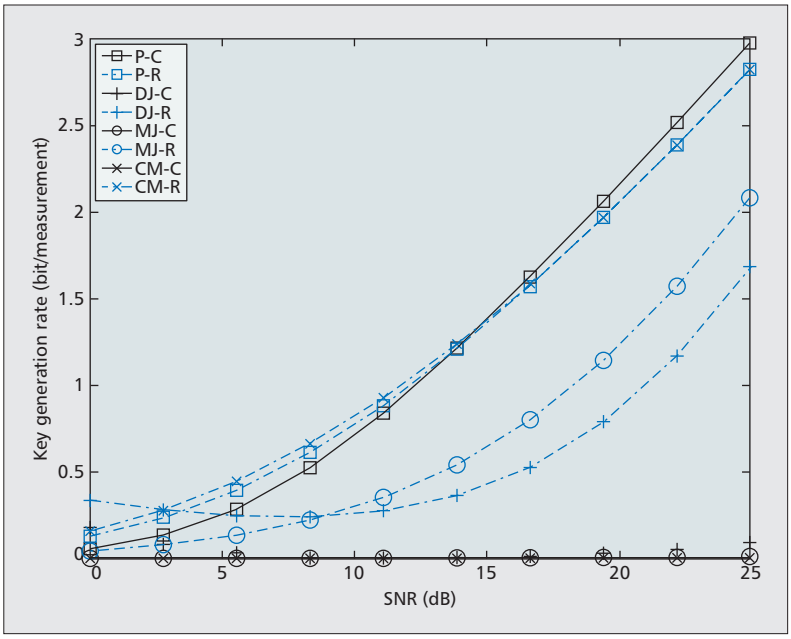


Figure 3. Key generation rate under different attack models.

The fundamental reason is that the proposed scheme integrates user-generated randomness and channel randomness to generate the key. The attacker has no control over the user-generated randomness, and thus cannot gain much information about the generated key. When the attacking signal becomes stronger, it actually helps to increase the key generation rate, since it introduces more randomness and higher reciprocity into the channel.

From Fig. 4, we can see that a large amount of information is leaked to an active attacker under manipulative jamming and channel manipulation attacks when using constant probing signals, which make the key generation rate nearly zero. When the attacking signal is stronger, more information is leaked. The leaked information here is actually the compromised key information. If Alice and Bob are not aware of these two kinds of attacks, they would generate a key that is largely controlled by the attacker.

When using random probing signals, the amounts of leaked information to both passive and active attackers under different attacking models are small, which indicates a strong security strength of the proposed scheme. It can be noted that although the scheme is proposed to counter active attacks, it also improves the security strength under passive attacks. As shown in Fig. 4, under passive attacks, the leaked information using random probing signals is much less than that using constant probing signals.

Future Research Topics

We have discussed various practical challenges of physical layer key generation and analyzed its security strength under both active and passive

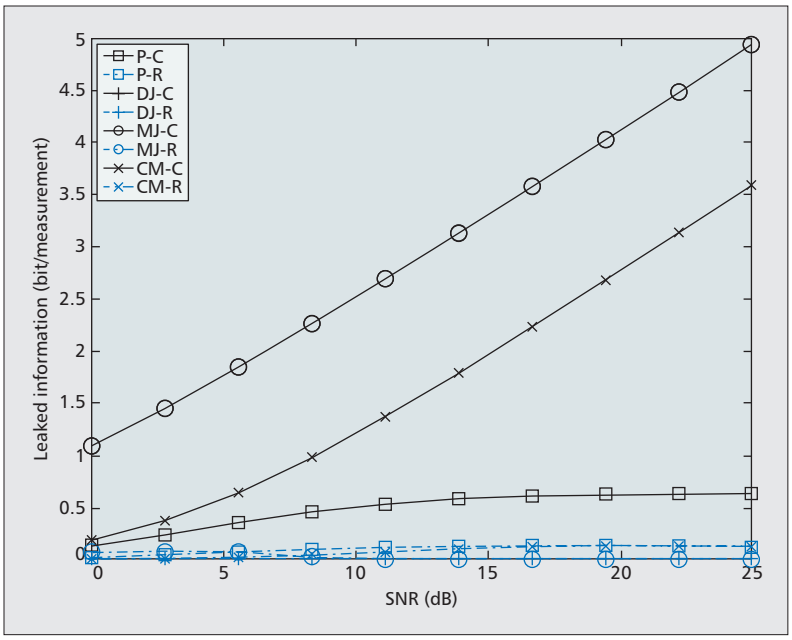


Figure 4. Leaked information under different attack models.

attacks. A lot of research issues and opportunities remain for further investigation. We discuss some of them in the following subsections.

Multi-User Case

Most of the existing key generation schemes are designed for two-user scenarios. Although the multi-user case is common in practice, the research on multi-user key generation is largely open. Both theoretical analysis on the secret key capacity and the design of practical protocols need further investigation. Users may need to generate pair-wise keys or group keys [12] under various constraints of time delay, energy consumption, spectrum efficiency, and so on. How to schedule the channel probing in a multi-user scenario is an interesting issue. In general, when a user probes the channel faster, it can achieve higher key generation rate [4]. However, the entropy of each measurement is reduced due to increased correlation between the consecutive channel measurements. Each user may alternatively probe the channel in order to fairly share the medium. The broadcast nature of the wireless channel can also be exploited to reduce the probing overhead if one user needs to generate a shared key with multiple neighbors.

Multihop Networks

Multihop networks such as mesh networks, mobile ad hoc networks, and sensor networks require secure end-to-end communication via multiple hops. An intuitive extension of the single-hop physical layer key generation to multihop networks is to generate a pair-wise secret key on each link and generate an end-to-end shared secret through per-hop encryption. However, an intermediate node or a forwarder may not be trustworthy or be compromised. There may be multiple paths between two nodes. The security strength, energy consumption, and delay on each path may be different. Generating an end-to-end secret key at the physical layer with consideration of joint medium access control and routing design under untrustworthy intermediate nodes is worth further investigation. Multipath diversity can also be exploited to enhance the security strength of the generated key.

Full-Duplex Radio

Almost all the existing physical layer key generation research is based on the assumption of half-duplex radio. With the practical implementation of full-duplex radio becoming reality, physical layer key generation can be made more efficient. A straightforward benefit is the decreased time lag between two bidirectional channel probings, which will increase the reciprocity of the channel measurements. When Alice and Bob transmit the probing signal at the same time, a passive attacker may only overhear a superposed signal, which prevents the attacker gaining CSI, thus enhancing security. However, an advanced attacker with multiple antennas may perform beamforming to separate the probing signals to gain CSI. Furthermore, the attacker can also jam and listen at the same time with full-duplex capability, which may increase the chance of key compromise for manipulative jamming attacks. The security strength of the proposed key generation scheme with random probing signals is expected to be maintained as long as the attacker cannot figure out the random probing signals. Near-field communication that exploits inductive coupling naturally supports full-duplex communication, which can also be exploited to achieve low-cost high-throughput key generation [13].

Millimeter-Wave Communication and Massive MIMO

With the evolution to 5G cellular communications, future wireless devices will be equipped with 60 GHz millimeter-wave radios with tens of antennas. The propagation property of millimeter-wave is different from those at lower frequency, and is more like a beam. This unique propagation property may be exploited to enhance physical layer key generation security since the channel may get decorrelated within a very short distance, thus improving security strength under eavesdropping attacks. With many antennas, the key generation rate is expected to be significantly increased with the increase of antenna spatial diversity. On the other hand, the channel coherence time, which is usually inversely proportional to the carrier frequency, will be significantly small, which introduces challenges on ensuring reciprocity of bidirectional channel measurements. Full-duplex technology may be integrated with millimeter-wave communication to achieve high channel reciprocity.

Integrating Other Physical/Physiological Random Sources

Other than using physical layer information to generate shared keys, we can integrate other random sources, such as readings from co-located sensors [14] or physiological information [15]. The IoT will consist of various devices equipped with sensors and wireless interfaces. How to exploit the properties of sensor readings and

channel randomness to bootstrap device-to-device secure communication and at the same time protect user privacy is an interesting issue.

Conclusion

In this article, we provide an overview of the physical layer key generation process and point out its practical challenges and possible solutions. We analyze the security strength of physical layer key generation under different attacking models. In order to achieve secure physical layer key generation under active attacks, we propose the use of random probing signals to hide the channel state information, and combine user generated randomness and channel randomness to generate a shared secret key. Numerical results show that the proposed scheme achieves much higher security strength than the existing scheme using constant probing signals. Future research directions on physical layer key generation under multi-user multihop scenarios and new communication technologies including full-duplex radio and millimeter-wave communications are discussed. It is also promising to integrate other physical or physiological information to generate shared secret keys.

Acknowledgments

This research was partially supported by the NSF CAREER award under Grant Number (CNS-1502584) and NSF EARS program grant under Grant Number (CNS-1464487).

References

[1] J. Hershey, A. Hassan, and R. Yarlagadda, "Unconventional Cryptographic Keying Variable Management," *IEEE Trans. Commun.*, vol. 43, no. 1, Jan 1995, pp. 3–6.
[2] S. Jana *et al.*, "On the Effectiveness of Secret Key Extraction from Wireless Signal Strength in Real Environments," *Proc. 15th ACM Annual Int'l. Conf. Mobile Computing and Networking '09*, 2009, pp. 321–32.
[3] N. Patwari *et al.*, "High-Rate Uncorrelated Bit Extraction for Shared Secret Key Generation from Channel Measurements," *IEEE Trans. Mobile Computing*, vol. 9, no. 1, Jan. 2010, pp. 17–30.
[4] Y. Wei, K. Zeng, and P. Mohapatra, "Adaptive Wireless Channel Probing for Shared Key Generation Based on Pid Controller," *IEEE Trans. Mobile Computing*, vol. 12, no. 9, Sept. 2013, pp. 1842–52.
[5] S. Ali, V. Sivaraman, and D. Ostry, "Eliminating Reconciliation Cost in Secret Key Generation for Body-Worn Health Monitoring Devices," *IEEE Trans. Mobile Comp.*, vol. 13, no. 12, Dec. 2014, pp. 2763–76.
[6] K. Zeng *et al.*, "Exploiting Multiple-Antenna Diversity for Shared Secret Key Generation in Wireless Networks," *Proc. IEEE INFOCOM '10*, 2010, pp. 1837–45.
[7] U. M. Maurer, "Secret Key Agreement by Public Discussion from Common Information," *IEEE Trans. Info. Theory*, vol. 39, no. 3, May 1993, pp. 733–42.
[8] M. Edman, A. Kiayias, and B. Yener, "On Passive Inference Attacks against Physical-Layer Key Extraction," *Proc. 4th Euro. Wksp. System Security*, 2011, pp. 8:1–8:6.
[9] K. Argyraki *et al.*, "Creating Secrets Out of Erasures," *Proc. 19th Annual Int'l. Conf. Mobile Computing and Networking*, 2013, pp. 429–40.
[10] M. Zafer, D. Agrawal, and M. Srivatsa, "Limitations of Generating a Secret Key Using Wireless Fading under Active Adversary," *IEEE/ACM Trans. Networking*, vol. 20, no. 5, Oct. 2012, pp. 1440–51.
[11] S. Eberz *et al.*, "A Practical Man-In-The-Middle Attack on Signal-Based Key Generation Protocols," *Proc. ESORICS, LNCS*, S. Foresti, M. Yung, and F. Martinelli, Eds., vol. 7459. Springer, 2012, pp. 235–52.
[12] H. Liu *et al.*, "Group Secret Key Generation via Received Signal Strength: Protocols, Achievable Rates, and Implementation," *IEEE Trans. Mobile Computing*, vol. 13, no. 12, Dec. 2014, pp. 2820–35.
[13] R. Jin et al., "Practical Secret Key Agreement for Full-Duplex Near Field Communications," *Proc. 9th ACM Symp. Info., Comp. and Commun. Security*, 2014, pp. 217–28.
[14] V. M. Prabhakaran, K. Eswaran, and K. Ramchandran, "Secrecy via Sources and Channels," *IEEE Trans. Info. Theory*, vol. 58, no. 11, Nov. 2012, pp. 6747–65.
[15] K. K. Venkatasubramanian *et al.*, "Pska: Usable and Secure Key Agreement Scheme for Body Area Networks," *IEEE Trans. Info. Tech. Biomed.*, vol. 14, no. 1, Jan. 2010, pp. 60–68.

Biography

Kai Zeng is an assistant professor in the Department of Electrical and Computer Engineering at George Mason University. He received his Ph.D. degree in electrical and computer engineering from Worcester Polytechnic Institute (WPI) in 2008. He was a postdoctoral scholar in the Department of Computer Science at University of California, Davis (UCD) from 2008 to 2011. He worked in the Department of Computer and Information Science at the University of Michigan — Dearborn as an assistant professor from 2011 to 2014. He was a recipient of the U.S. National Science Foundation Faculty Early Career Development (CAREER) award in 2012. He won the Excellence in Postdoctoral Research Award at UCD in 2011 and Sigma Xi Outstanding Ph.D. Dissertation Award at WPI in 2008. He is an Editor of *IEEE Transactions on Wireless Communications*. His current research interests are in cyber-physical system security and privacy, physical layer security, network forensics, and cognitive radio networks.

The IoT will consist of various devices equipped with sensors and wireless interfaces. How to exploit the properties of sensor readings and channel randomness to bootstrap device-to-device secure communication and at the same time protect user privacy is an interesting issue.

Distributed Inference in the Presence of Eavesdroppers: A Survey

Bhavya Kailkhura, V. Sriram Siddhardh Nadendla*, and Pramod K. Varshney*

ABSTRACT

The distributed inference framework comprises a group of spatially distributed nodes that acquire observations about a POI and transmit computed summary statistics to the fusion center. Based on the messages received from the nodes, the FC makes a global inference about the POI. The distributed and broadcast nature of such systems makes them quite vulnerable to different types of attacks. This article focuses on efficient mitigation schemes to mitigate the impact of eavesdropping on distributed inference and surveys the currently available approaches along with avenues for future research.

INTRODUCTION

Distributed inference networks (DINs) have attracted much recent attention due to a variety of applications in civilian and military domains. These include surveillance, environment monitoring, cognitive radio networks, and cyber physical systems. DINs employ a group of sensing entities that collaborate to sense and make inferences about a given point of information (POI). In the traditional framework of centralized inference networks, nodes transmit raw observations to the fusion center (FC). These transmissions are not attractive in practice due to their large bandwidth and energy requirements. Therefore, DINs have been proposed, where the nodes transmit compressed observations that are obtained by processing original observations into a finite alphabet set.

In this article, we denote the POI with a variable $\theta \in \Theta$, where Θ is the set of possible states the phenomenon can take. Consider a distributed network, as shown in Fig. 1, comprising N sensors and an FC, which makes inferences about the POI. We assume that the ith node makes an observation Y_i and compresses it into a symbol v using a quantizer γ_i. The compressed symbol v_i is then transmitted to the FC through a channel, which is represented as a function $C_F^i(\cdot)$. We denote the received symbols at the FC as $u_i = C_F^i(v_i)$, corresponding to the ith sensor's transmission. The FC uses the fusion rule Γ_{FC} to integrate the symbols $u = \{u_1, \ldots, u_N\}$ into a global inference $\hat{\theta}_{\mathrm{FC}} \in \Theta$ about the unknown phenomenon θ.

Although the problem of distributed inference encompasses a broader set of problems, in this article, we focus our attention on two fundamental problems: *distributed detection* and *distributed estimation*. The fundamental difference in the two problems lies in the definition of the set Θ. In the case of *distributed detection*, $\Theta = \{0, 1\}$, and in the case of *distributed estimation*, Θ is a continuous set. A practical application of distributed detection is a cognitive radio (CR) network where the secondary users are interested in vacant primary user (PU) channels. On the other hand, examples of distributed estimation include location estimation and surveillance using spatially distributed sensors.

Despite their many benefits, the distributed and broadcast nature of the communication links makes DINs susceptible to a breach in confidentiality, which is an important problem especially when the network is a part of a larger cyber-physical system. In a fundamental sense, there are two motives for any eavesdropper (Eve),[1] *selfishness* and *malice*, to compromise the confidentiality of a given DIN. For instance, some of the nodes within a CR network may selfishly take advantage of the FC's inferences and may compete against the CR network in using the PU's channels without paying any participation costs to the network moderator. Therefore, in the recent past, there has been a lot of interest in addressing confidentiality in DINs. While our focus in this article is on physical layer secrecy of DINs, it should be mentioned that physical layer secrecy issues have also been considered for other systems such as communication networks [2].

To set the notations, we represent the channel between the ith sensor and Eve as a function $C_E^i(\cdot)$. The symbol corresponding to the ith node received at Eve is denoted by $w_i = C_E^i(v_i)$ (Fig. 1). In other words, the total information leakage is a function of $\mathbf{w} = \{w_1, \cdots, w_N\}$. Similar to the FC, we assume that Eve uses a decision rule Γ_E to integrate the symbols $\mathbf{w}$ into its own global inference $\hat{\theta}_E$. Several metrics have been proposed in the literature to quantify secrecy or the information leakage to Eve. Some of them include equivocation, Kullback-Leibler divergence (KLD) , Fisher Information (FI), and probability of error (P_E). Ideally, the goal is to minimize this information leakage to the maximum extent possible. For example, if KLD or conditional FI is the chosen metric, *perfect secre-*

The authors are with Syracuse University.

** These authors contributed equally to this work.*

[1] In this article, we only survey passive eavesdropping attacks, in which Eve just intercepts the ongoing communication between sensors and the FC, and does not send any signals over the communication channels. In certain cases, Eve can take a more active approach by transmitting falsified information to the FC in addition to eavesdropping. These attacks, in the context of distributed inference, are referred to as Byzantine attacks [1].

0163-6804/15/$25.00 © 2015 IEEE

cy is achieved when KLD or conditional FI at Eve becomes zero.

In this article, we survey the state-of-the-art approaches proposed to address secrecy in the context of DINs. We first introduce a taxonomy where we present a survey of the state of the art on secrecy in DINs. Then we specifically focus on distributed detection and estimation frameworks, respectively, where we present a detailed account of how secrecy is addressed in each of these frameworks. Finally, we present some important open problems while designing a secure DIN in the presence of eavesdroppers.

Approaches to Mitigate Threats on Confidentiality

There are fundamentally four approaches to address secrecy in the context of DINs, which we discuss next.

Design of Sensor Quantizers and Fusion Rule

In this approach, the network designer takes advantage of the difference in the channels (C_F^i, C_E^i), for all $i = 1, \cdots, N$, while designing sensor quantizers and the fusion rule. We denote by $\gamma = \{\gamma_1, \cdots, \gamma_N\}$ the vector of all sensor quantizers in the DIN. We assume that the quantizer γ_i at the ith sensor lies within the set $\mathbb{R}_i$, $i = 1, \cdots, N$. Similarly, we denote the set of decision rules at the FC and Eve as $\mathbb{R}_{FC}$ and $\mathbb{R}_E$, respectively.

Without loss of generality, we denote the performance metric at the FC and Eve as Ω_{FC} and Ω_E, respectively. Consider a scenario where the network has a tolerable upper bound on the amount of information leaked to Eve. Mathematically, this can be quantified in terms of a constraint α on Eve's performance metric Ω_E. Then one approach for distributed inference system design is below.

Problem 1: *Find* (γ, Γ_{FC}) *such that* Ω_{FC} *is maximized while satisfying the constraints*:
- $\max_{\Gamma_E \in \mathbb{R}_E} \Omega_E$ *lies below a tolerable value* α.
- *Quantizers satisfy* $\gamma_i \in \mathbb{R}_i$, $\forall i = 1, \cdots, N$.
- *Fusion rule at the FC satisfies* $\Gamma_{FC} \in \mathbb{R}_{FC}$.

Note that error exponents are asymptotic performance metrics at the FC and Eve that represent exponential decay rates of P_E of their respective "optimal" detectors. Therefore, if the performance metric chosen is an error-exponent such as KLD (for Neyman-Pearson detection setup) or Chernoff Information (for Bayesian detection setup), Problem 1 becomes independent of the fusion rules Γ_{FC} and Γ_E at the FC and Eve, respectively, and reduces to the design of sensor quantizers alone.

Stochastic Encryption

Through another approach where the network is designed within the tolerable bounds on information leakage to Eve, one can pursue a more active approach where the sensors flip their decisions randomly in order to confuse Eve. In this case, the FC is assumed to have better knowledge about the sensors than Eve, since the FC either deterministically knows the flipping sensors or has knowledge about the flipping probability, about which Eve is completely ignorant. This introduces a significant difference in the channels (C_F^i, C_E^i), $i = 1, \cdots, N$, thus reducing the information leakage to Eve.

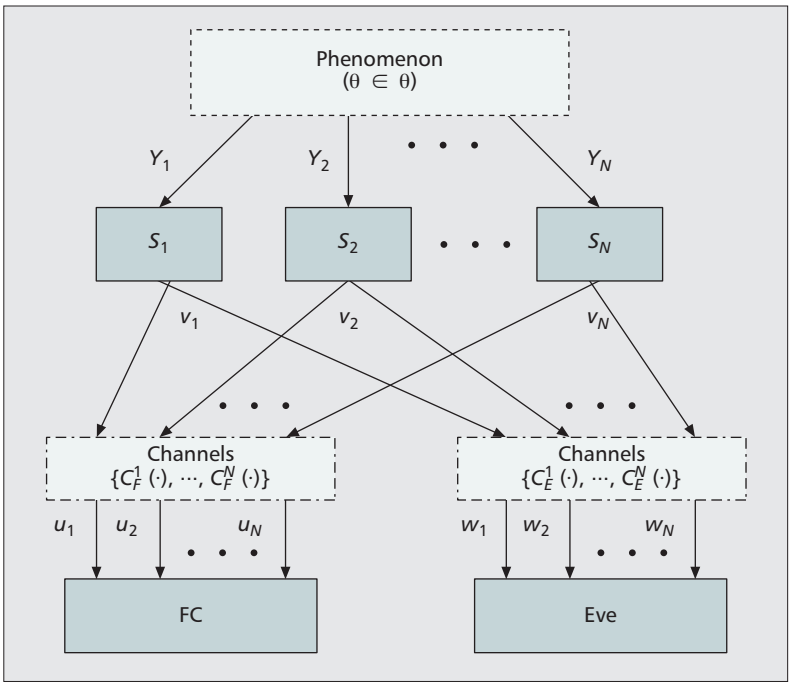


Figure 1. Distributed inference network in the presence of an eavesdropper.

Let the alphabet set of the compressed symbols v_i at the ith sensor be denoted as $\mathcal{A}$, where the size of $\mathcal{A}$ is denoted by M. In other words, the ith sensor employs an M-ary quantizer to compress the observation Y_i into one of the M symbols. Let us denote the flipping probability matrices as $\mathcal{P} = \{P_1, \cdots, P_N\}$, where P_i denotes the flipping probability matrix at the ith sensor, which can be interpreted as pre-shared keys between the nodes and the FC. Note that P_i is a stochastic matrix for any $i = 1, \cdots, N$, since all of its row elements sum up to unity. The basic problem in this case can be stated as below.

Problem 2: *Find* $\mathcal{P} = \{P_1, \cdots, P_N\}$ *such that* Ω_{FC} *is maximized while satisfying the constraints:*
- $\max_{\Gamma_E \in \mathbb{R}_E} \Omega_E$ *lies below a tolerable value* α.
- P_i *is a row-stochastic matrix, for all* $i = 1, \cdots, N$.

Note that several variants of this problem can be investigated depending on the amount of knowledge the FC has regarding the stochastic encryption process. For example, one may consider that the FC has complete knowledge about the flipping probability matrices $\mathcal{P}$, but does not know exactly whether or not the sensor messages are flipped. In this case, the FC can improve the secrecy performance at the expense of detection performance. On the other hand, the ideal scenario is the case where the FC acquires exact instantaneous knowledge regarding which sensor messages are flipped. This can be done by spending energy in the mechanism that facilitates communication between the FC and the flipping sensors.

The ideal scenario is the case where the FC acquires exact instantaneous knowledge regarding which sensor messages are flipped. This can be done by spending energy in the mechanism that facilitates communication between the FC and the flipping sensors.

Artificial Noise Injection

Another approach, similar to the case of stochastic encryption, is the addition of artificial noise to sensor transmissions. Note that both stochastic encryption and the addition of artificial noise to sensor transmissions are data falsification schemes that are employed to confuse Eve.

In this article, we denote the artificial noise added to the ith sensor's transmissions as η_i. Then the ith sensor transmits $\mathbf{x}_i$ to the FC and Eve, where $\mathbf{x}_i = v_i + \eta_i$. Let $f_i(\eta_i)$ denote the distribution of η_i. Also, let $\mathcal{F} = \{f_1(\eta_1), \cdots, f_N(\eta_N)\}$ denote the set of artificial noise distributions employed by all the sensors in the network. Then the problem can be stated as follows.

Problem 3: *Find $\mathcal{F} = \{f_1(\eta_1), \cdots, f_N(\eta_N)\}$ such that Ω_{FC} is maximized while satisfying the constraints:*

- $\max_{\Gamma_E \in \mathbb{R}_E} \Omega_E$ *lies below a tolerable value* α.
- $f_i(\eta_i)$ *is a probability density function of* η_i, *for all* $i = 1, \cdots, N$.

MIMO Beamforming

In order to ensure minimal performance loss at the FC as a trade-off to attaining the secrecy constraint at Eve, another alternative approach is to use multiple-input multiple-output (MIMO) beamforming, where the sensor messages are directed toward the FC. In this case, we assume that the sensors are equipped with multiple antennas to transmit their messages to the FC. The beamforming mechanism is designed in such a way that some of the available energy is invested in the beams directed toward the FC, while the nulls are directed toward Eve.

We denote the number of antennas at the ith sensor as L_i. Therefore, the ith sensor constructs a vector $\mathbf{x}_i$ based on the symbol v_i and transmits it to the FC and Eve, respectively. Based on the channel gains at the FC and Eve, this $\mathbf{x}_i$ is designed to appear very noisy at Eve, and simultaneously have significant information about the compressed symbol v_i at the FC. For example, let $\mathbf{x}_i$ be constructed as $\mathbf{x}_i = \mathbf{b}_i v_i$, where $\mathbf{b}_i$ is the beamforming gain vector of the ith sensor's signal. Assuming that both the FC and Eve have only a single antenna, the resulting received symbols at the FC and Eve are given by u_i and w_i, respectively. Let n_{FC_i} and n_{E_i} denote the noise at the FC and Eve, respectively. Then $u_i = \mathbf{h}_i^T \mathbf{x}\mathbf{i} + n_{FC_i} = v_i \mathbf{h}_i^T \mathbf{b}_i + n_{FC_i}$ and $w_i = \mathbf{g}_i^T \mathbf{x}\mathbf{i} + n_{E_i} = v_i \mathbf{g}_i^T \mathbf{b}\mathbf{i} + n_{E_i}$. Let the beamforming matrix be denoted as $B = [\mathbf{b}_1 \cdots \mathbf{b}_N]$. Since all practical sensors are energy-constrained, we assume that the total energy available at the ith sensor is denoted by E_i. Then the design problem can be formally stated as follows.

Problem 4: *Find $B = [\mathbf{b}_1 \cdots \mathbf{b}_N]$ such that ΩFC is maximized while satisfying the constraints:*

- $\max_{\Gamma_E \in \mathbb{R}_E} \Omega_E$ *lies below a tolerable value* α.
- $\mathbf{b}_i$ *is chosen such that the total transmit energy is within the prescribed limit* E_i *for all* $i = 1, \cdots, N$.

In general, it is assumed that the FC is more powerful than the sensors in terms of available resources (e.g., hardware/software and available energy). Also, all of the above approaches can be combined together to design a system in a holistic manner and attain better performance in terms of Ω_{FC}, given a tolerable Eve's constraint α.

Secrecy in Distributed Detection

In this section, we provide a survey on how secrecy is addressed within the framework of classical and compressive detection networks respectively. In both these frameworks, we organize the survey according to the four different approaches listed previously.

Classical Distributed Detection

First, we consider the first approach, where the distributed detection network (i.e., sensor quantizers and fusion rule) is optimized while satisfying the secrecy constraints at Eve. Nadendla *et al.* made the first attempt in 2010 in [3], where they considered an unconstrained differential secrecy problem. Let us denote KLDs at the FC and Eve by D_{FC} and D_E, respectively. Problem 1 in their setup reduces to the design of sensor quantizers alone, with $\Omega_{FC} = D_{FC} - D_E$ and $\alpha = \infty$. It was assumed that the channel state information (CSI) is completely known at both the FC and Eve. The authors showed that in the case of an eavesdropper with noisier channels, the optimal local detectors are always on the boundaries of the achievable region of a sensor's ROC and therefore are likelihood-ratio tests (LRTs). The authors also considered Problem 1 with $\Omega_{FC} = D_{FC}$ and $\Omega_E = D_E$, in which case the structure of an optimal local detector was conjectured to be an LRT-based test based on numerical results.

In 2009, Marano *et al.* [4] considered the problem of designing optimal decision rules for a sensor network where the sensors perform censoring in order to save energy. It was assumed that the eavesdropper does not have access to the sensors' transmitted data, but can monitor the transmission activity of the channel and exploit the busy/idle state of the channel to detect the hypothesis. KLD was used as the performance metric for both the FC and Eve, and a censoring strategy was developed in order to maximize the KLD of FC while ensuring that the KLD of Eve was zero (perfect secrecy). Although their framework of censoring sensor networks is more general, they assumed that Eve can only determine whether or not an individual sensor transmits its decision. In reality, Eve can extract more information than merely determining the presence or absence of transmission, and hence can make a reasonably good decision based on its reception.

Li *et al.* investigated the problem of Bayesian distributed detection in 2014 with two nodes in the network in the presence of an eavesdropper [5], where Eve has access to only one of the sensor's transmissions. Here, Ω_{FC} and Ω_E were assumed to be negative expected detection costs at the FC and Eve, respectively. The authors proved that LRT-based tests were optimal at the sensors if the network is designed to minimize the expected detection cost at the FC such that the minimum average cost at Eve is no greater than a prescribed non-negative value a.

Li *et al.* [6] also investigated the detection

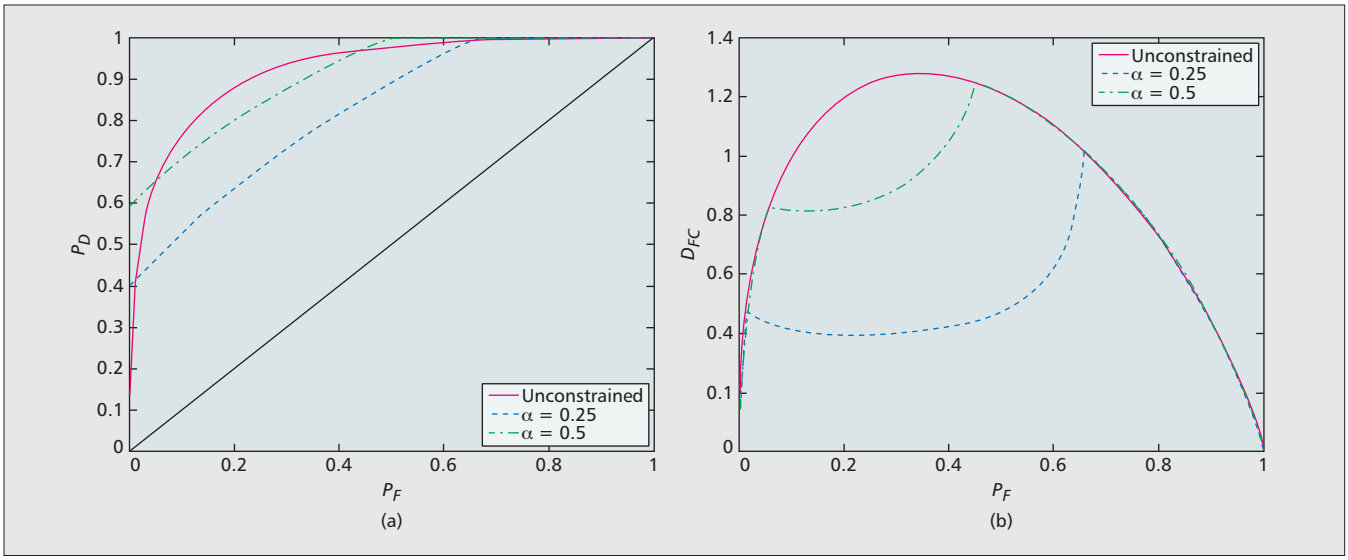


Figure 2. Sensor performance in the presence of a constraint, $D_E \leq \alpha$, where $\rho_e = 0.1$ [7]: a) a sensor's ROC in the presence of Eve; b) D_{FC} as a function of local false alarm probability.

problem under the Neyman-Pearson setup for the same network as in [5]. The sensor quantizers and the fusion rule were designed to maximize the FC's probability of detection (Ω_{FC}) in the presence of constraints on false alarm probabilities at the FC and Eve, along with the probability of detection at Eve (Ω_E). Note that the false alarm constraints at both the FC and Eve are captured by the feasibility sets $\mathbb{R}_{FC}$ and $\mathbb{R}_E$, respectively. Here, the authors proved that the optimal local quantizer is a deterministic LRT, while the fusion rule may still be a randomization between two or more LRTs. Later, in 2014, Nadendla *et al.* [7] investigated a more general framework in with N sensors. They proved the conjecture in [3] in the context of binary symmetric channels[2] (BSCs) between the sensors, FC, and Eve. An algorithm was also presented to find optimal thresholds for the likelihood ratio quantizers when the sensor observations are corrupted by additive Gaussian noise. Figure 2 depicts the FC's performance in terms of both probability of detection (P_D) and KLD at the FC as a function of false alarm probability (P_F) in the presence of tolerable limits on Eve's KLD. Note that the optimal quantizer is always on the intersection of the ROC and Eve's constraint curve. The authors also showed that the network with non-identical sensors and channels can be designed by solving N sequential problems, where the order of this sequence is dictated by the quality of the corresponding sensor's channel.

Next, we survey the literature that addresses the second mitigation approach, where a stochastic cipher is employed to confuse Eve regarding the true phenomenon. Soosahabi *et al.* [8] employ J-divergence as the performance metric for both the FC and Eve and design a network that guarantees perfect secrecy. This is achieved by fixing $\alpha = 0$ in Problem 2. Probabilistic ciphers were also studied in [9], where the performance metric chosen was the error probability in the case of both FC and Eve. Note that both [8, 9] assume the existence of an underlying key exchange mechanism that is secure from Eve. Alternatively, channel-aware stochastic ciphers use seeds that are obtained by exploiting randomness in the channel-gains between the node and the FC. For example, Jeon *et al.* [10] proposed a type-based multiple access (TBMA) protocol for a distributed detection network with a multiple access channel (MAC). Here, some of the nodes in the network are selected to deliberately transmit interfering signals so as to minimize degradation in the FC's detection performance, while simultaneously preventing Eve from identifying the sensors generating interference. Note that the above scheme requires full CSI at the sensors, and therefore may be impractical in some scenarios. In order to alleviate this problem, efforts such as [11] have been made in the literature, where Jeon *et al.* designed a secure transmission strategy for the local nodes in a parallel distributed detection network, in which the FC first broadcasts known symbols and two thresholds to let the nodes measure their channel condition. Depending on the received symbols, the nodes are divided into three groups, non-flipping, flipping, and dormant. The non-flipping set of sensors quantize the sensed data and transmit them to the FC, while the flipping sensors transmit flipped decisions in order to confuse Eve. The sensors within the dormant set sleep in order to conserve energy and have an energy-efficient sensor network with longer lifetime.

Finally, there have been efforts to design a hybrid mitigation approach that combines the effects of both the first and second approaches. In this regard, Nadendla [12] considered the problem of Bayesian distributed detection in the presence of an eavesdropper, where the nodes use identical threshold quantizers to make their binary decisions and encrypt them before transmission using a simple probabilistic cipher. Cipher parameters and threshold were optimized jointly so as to ensure an acceptable probability of error at the FC while maximizing the probability of error at Eve.

[2] The BSC model is used here as it is able to model many practical channels.

Collaborative Compressed Detection

In scenarios where the POI is a high dimensional signal vector, the collaborative compressed detection (CCD) framework has been proposed. In contrast to the conventional detection framework, in CCD, the detection problem is solved directly in the compressive measurement domain. More specifically, the CCD framework comprises a group of spatially distributed nodes that require observations regarding the high-dimensional ($K \times 1$) signal vector to be detected. Nodes compress their observations using a $M \times K$ low-dimensional ($M \ll K$) random projection operator ϕ. Each node i sends an unquantized (or quantized) version of compressed observation vector Y_i to the FC, where a global decision is made.

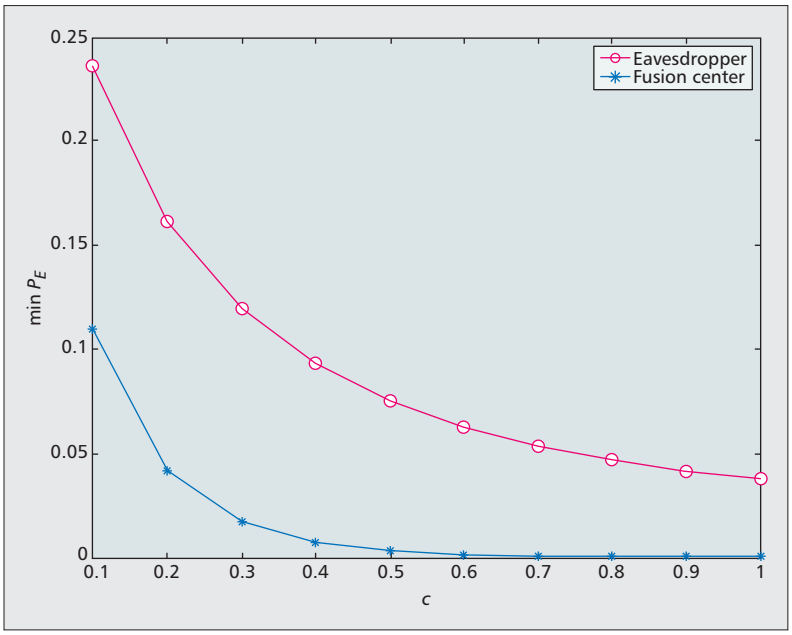


Figure 3. Minimum error probability as a function of compression ratio c where local sensor threshold $\lambda = 1$, $\beta = 0.2$, SNR = 10 dB, and $N = 10$ [14].

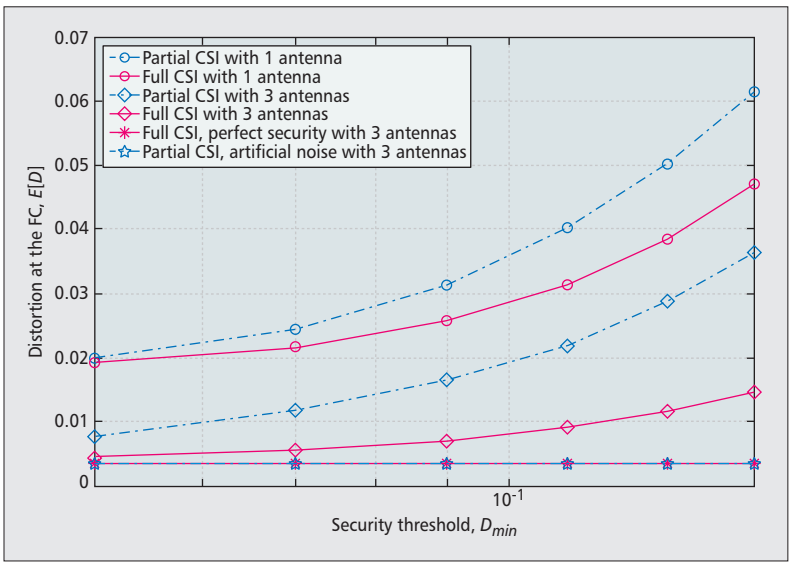


Figure 4. Performance comparison between full CSI, partial CSI and artificial noise in a multiple-antenna system [15].

First, we focus our attention on the first approach, where nodes do not quantize their observations, and the FC receives compressed observation vectors, $\mathbf{Y} = [Y_1, \cdots, Y_N]$. Kailkhura *et al.* [13] considered the problem of collaborative signal vector detection using unquantized compressive measurements under a physical layer secrecy constraint $\Omega_E \leq \alpha$. To counter Eve, the authors considered a variant of the third mitigation approach by proposing to use β fraction of cooperative nodes that assist the FC by injecting artificial noise (adding or subtracting a constant vector D_i from their observation vector Y_i) in the system to confuse the eavesdroppers. The authors employed deflection coefficient d_i as the performance metric for both the FC and Eve; thus, $\Omega_{FC} = d_{FC}$, and $\Omega_E = d_E$. The problem of determining optimal system parameters (i.e., compression ratio c and noise injection parameters (β, D_i)) that maximize d_{FC} while ensuring perfect secrecy at the eavesdropper (information of the eavesdropper is exactly zero, that is, $a = 0$) was also considered.

Kailkhura *et al.* [14] extended the CCD framework to the case where compressive measurements were quantized to one bit using the LRT. The performance metric was assumed to be P_E. The authors considered a hybrid mitigation approach that combines the features of both the second and third mitigation approaches. They proposed to use B out of N cooperating trustworthy nodes that assist the FC by providing flipped decisions (stochastic enciphering with

$$P_i = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

for all $i = 1, \cdots, B$) to Eve to achieve perfect secrecy. The authors considered the problem of designing optimal system parameters (fusion rule, compression ratio c, and fraction of data falsifying nodes $\beta = B/N$) such that P_E at the FC is minimized while ensuring perfect secrecy. In Fig. 3, the minimum P_E (over all the fusion rules), at both the FC and the eavesdropper, is plotted as a function of c for the equal prior case. It can be seen from Fig. 3 that the detection performance, at both the FC and the eavesdropper, is a monotonically increasing function of c; that is, detection performance is better with less compression. This suggests that compression improves security performance at the expense of detection performance.

Secrecy in Distributed Estimation

In this section, we survey the state of the art on how breaches in confidentiality are mitigated in distributed estimation networks. First, we survey the fourth mitigation approach in the context of distributed estimation networks. For example, Guo *et al.* [15] considered the problem of estimating a single point Gaussian source in the presence of Eve, where the sensor observations are transmitted using an amplify-and-forward technique over a slow-fading orthogonal MAC. Two different scenarios were addressed within this framework: one where there are multiple

nodes, with each node having a single transmit antenna, and another where a single node has multiple antennas. Through appropriate power allocation at the sensors, the network is designed to achieve the minimum mean square error (MSE) regarding the POI in each of the above mentioned scenarios while guaranteeing MSE at Eve to be greater than a threshold α. As shown in Fig. 4, the authors plot the distortion (MSE) performance at the FC with respect to the security threshold $\alpha = D_{min}$, for a one-antenna case and a three-antenna case, respectively. For comparison, the system performance is depicted under four settings, namely partial CSI, full CSI, full CSI with perfect secrecy, and partial CSI with artificial noise. First, due to the channel knowledge of both the FC and Eve, it is not surprising to see that the performance of full CSI is superior to the performance of partial CSI, and the gap keeps increasing as we increase the secrecy threshold. Another important observation is the small gap between the MSE in the perfect secrecy setting and the MSE in the setting with artificial noise. Similar performance was also obtained for the multiple nodes network, where each node has only one transmit antenna.

Next, we survey how stochastic encryption is used as a mitigation scheme in distributed estimation networks. Aysal *et al.* [16] considered the problem of distributed estimation of a deterministic signal in the presence of an Eve, where each node collects a noisy observation, performs binary quantization, and transmits the 1-bit decision to the FC. The authors assume that both the FC and Eve pursue maximum likelihood estimation in the presence of a stochastic cipher, for which bias, variance, and MSE were derived in closed form. In the context of symmetric ciphers, where

$$P_i = \begin{pmatrix} 0 & p \\ p & 0 \end{pmatrix}$$

for all $i = 1, \cdots, N$, the behavior of Eve's bias and MSE and FC's Cramer-Rao lower bound (CRLB) are characterized in Fig. 5. Note that as $p \to 0$, Eve's bias increases, Eve's MSE increases, and the CRLB decreases. On the other hand, as p tends to unity, Eve's bias decreases, Eve's MSE decreases, and the CRLB decreases. In other words, choosing a smaller p is better as it results in a significant amount of bias and MSE at Eve, with a marginal increase in the estimation variance at the FC. In the case where

$$P_i = \begin{pmatrix} 0 & p_0 \\ p_i & 0 \end{pmatrix}$$

for all $i = 1, \cdots, N$ with $p_0 \neq p_1$, the effect of varying p_0 and p_1 on the FC's CRLB, Eve's bias, and Eve's MSE are summarized in Fig. 6. In their numerical results, the authors also demonstrated that asymmetric ciphers (i.e., ciphers with asymmetric flipping probability matrices) produce greater bias and MSE than the symmetric ciphers.

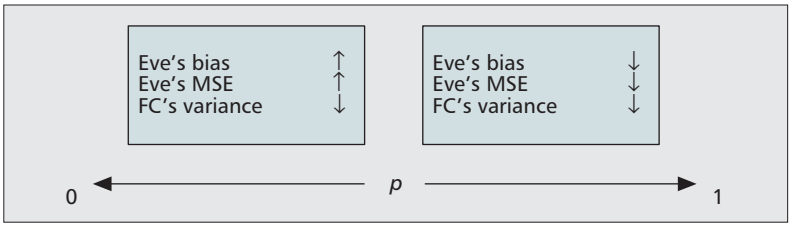


Figure 5. The effect of varying p on the FC's CRLB (variance of the optimal ML estimator), and the bias and MSE of Eve [16].

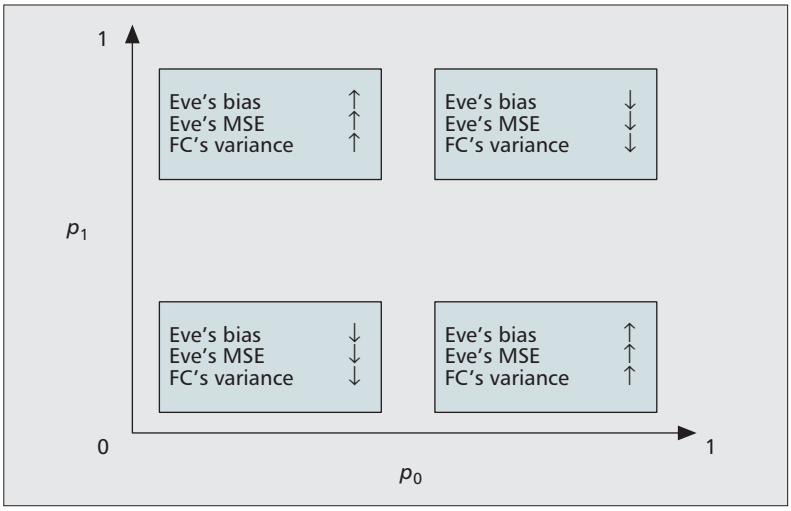


Figure 6. The effect of varying p_0 and p_1 on the FC's CRLB, Eve's bias, and Eve's MSE [16].

Summary and Open Problems

Despite the increasing attention on secure distributed inference in the presence of eavesdroppers, research in this area is still at an early stage. So far, four different approaches have been proposed to mitigate breaches in confidentiality in the context of DINs. However, all of these four approaches rely on an important underlying assumption: that Eve's channels C_E^i, for all $i = 1, \cdots, N$, are completely known at the FC, which may not be true in practice. In fact, there is no work in the context of inference networks on how one can acquire the information about a passive Eve's channel. This is a hard problem to solve because there is no feedback from Eve to any of the nodes in the network regarding its presence or activity. An alternative to this roadblock is to assume that Eve's channel belongs to a set $\mathcal{C}$, and investigate the best and worst case performance at Eve over a class $\mathcal{C}$. Information regarding this set $\mathcal{C}$ can be obtained from the scene where the network is deployed.

Also, the designers may extend the aforementioned four fundamentally different approaches into several hybrid approaches by considering two or more of these approaches together to create a more sophisticated and improved system in terms of FC performance for a given tolerable constraint on Eve. Although there have been a few attempts in this direction, one can still envision many such hybrid mechanisms where the designer may accumulate the benefits of each of these approaches. For example, including an energy consumption constraint for secure system design is an important open problem. Of course, there is always a need for new approaches that are fundamentally different from any of the four approaches listed in this article.

Recently, other architectures such as tree, tandem and flat (without the FC) have attracted much interest due to their practical significance. These architectures pose serious issues in terms of secrecy and secure communications. Studying the secrecy performance of such systems in the context of distributed inference is worth exploring in the future.

Recently, other architectures such as tree, tandem, and flat (without the FC) have attracted much interest due to their practical significance. These architectures pose serious issues in terms of secrecy and secure communications. Studying the secrecy performance of such systems in the context of distributed inference is worth exploring in the future.

References

[1] A. Vempaty, L. Tong, and P. Varshney, "Distributed Inference with Byzantine Data: State-of-the-Art Review on Data Falsification Attacks," *IEEE Signal Processing Mag*., vol. 30, no. 5, 2013, pp. 65–75.

[2] Y. Liang, H. Vincent Poor, and S. Shamai (Shitz), "Information Theoretic Security," *Found. Trends Commun. Info. Theory*, vol. 5, Apr. 2009, pp. 355–580.

[3] V. S. S. Nadendla, H. Chen, and P. K. Varshney, "Secure Distributed Detection in the Presence of Eavesdroppers," *Proc. 44th ASILOMAR*, 2010, Nov. 2010, pp. 1437–41.

[4] S. Marano, V. Matta, and P. K. Willett, "Distributed Detection with Censoring Sensors under Physical Layer Secrecy," *IEEE Trans. Signal Pro*cessing, vol. 57, no. 5, May 2009, pp. 1976–86.

[5] Z. Li, T. J. Oechtering, and K. Kittichokechai, "Parallel Distributed Bayesian Detection with Privacy Constraints," *Proc. IEEE ICC*, June 2014, pp. 2178–83.

[6] Z. Li, T. J. Oechtering, and J. Jalden., "Parallel Distributed Neyman-Pearson Detection with Privacy Constraints," *Proc. IEEE ICC Wksps.*, 2014, June 2014, pp. 765–70.

[7] V. Sriram Siddhardh Nadendla and Pramod K. Varshney, "Design of Binary Quantizers for Distributed Detection Under Secrecy Constraints," Oct. 2014.

[8] R. Soosahabi *et al.*, "Optimal Probabilistic Encryption for Secure Detection in Wireless Sensor Networks," *IEEE Trans. Info. Forensics and Security*, vol. 9, no. 3, Mar. 2014, pp. 375–85.

[9] R. Soosahabi and M. Naraghi-Pour, "Scalable PHY-Layer Security for Distributed Detection in Wireless Sensor Networks," *IEEE Trans. Info. Forensics and Security*, vol. 7, no. 4, Aug. 2012, pp. 1118–26.

[10] H. Jeon *et al.*, "Secure Type-Based Multiple Access," *IEEE Trans. Info. Forensics and Security*, vol. 6, no, 3, Sept. 2011, pp. 763–74.

[11] H. Jeon *et al.*, "Channel Aware Encryption and Decision Fusion for Wireless Sensor Networks," *IEEE Trans. Info. Forensics and Security*, vol. 8, no. 4, Apr. 2013, pp. 619–25.

[12] V. S. S. Nadendla, *Secure Distributed Detection in Wireless Sensor Networks via Encryption of Sensor Decisions*, Master's thesis, LA State Univ., 2009.

[13] B. Kailkhura, T. Wimalajeewa, and P. K. Varshney, "On Physical Layer Secrecy of Collaborative Compressive Detection," *Proc. 48th ASILOMAR*, 2014.

[14] B. Kailkhura *et al.*, "Distributed Compressive Detection with Perfect Secrecy," *Proc. 2nd Int'l. Wksp. Compressive Sensing in Cyber-Physical Systems*, 2014.

[15] X. Guo, A. S Leong, and S. Dey, "Estimation in Wireless Sensor Networks with Security Constraints," 2014.

[16] T. C. Aysal and K. E. Barner, "Sensor Data Cryptography in Wireless Sensor Networks," *IEEE Trans. Info. Forensics and Security*, vol. 3, no. 2, June 2008, pp. 273–89.

Biographies

Bhavya Kailkhura [S'12] received an M.S. degree in electrical engineering from Syracuse University, New York. Since 2012, he has been pursuing a Ph.D. degree with the Department of Electrical Engineering and Computer Science, Syracuse University. His research interests include high-dimensional data analysis, signal processing, machine learning, and their applications to solve inference problems with security and privacy constraints.

V. Sriram Siddhardh Nadendla [S] received his B.E. degree in electronics and communication engineering from Sri Chandrasekharendra Saraswati Viswa Mahavidyalaya (SCSVMV University), Kanchipuram, India, in 2007 and his M.S. degree in electrical engineering from Louisiana State University, Baton Rouge, in 2009. He is currently pursuing his Ph.D. in electrical and computer engineering in the Department of Electrical Engineering and Computer Science, Syracuse University. His current research interests are in security, distributed inference, wireless communications and networks, resource allocation, game theory, and mechanism design. During 2007–2009, he worked as a teaching assistant in the Department of Electrical and Computer Engineering, Louisiana State University. Since August 2009, he has been a research assistant at the Center for Advanced Systems Engineering (CASE), Syracuse University. He was a research intern at ANDRO Computational Solutions, LLC, Rome, New York, in summer 2013. He serves as a reviewer of *IEEE Transactions on Wireless Communications* and *IET Communications*. During his stay at SCSVMV University, he was awarded the University Gold Medal along with his Bachelor's degree. In addition to annual merit scholarships, he was also a recipient of the Dr. S. Subbulakshmi Endowment Cash Prize and Dr. S. Suryanarayanan Endowment Cash Prize for securing the highest marks in the university examinations, SCSVMV(DU), during 2003–2007.

Pramod K. Varshney [S'72, M'77, SM'82, F'97] received a B.S. degree in electrical engineering and computer science (with highest honors), and M.S. and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana-Champaign in 1972, 1974, and 1976, respectively. From 1972 to 1976, he held teaching and research assistantships with the University of Illinois. Since 1976, he has been with Syracuse University, where he is currently a Distinguished Professor of Electrical Engineering and Computer Science and director of CASE: Center for Advanced Systems and Engineering. He served as associate chair of the department from 1993 to 1996. He is also an adjunct professor of radiology at Upstate Medical University, Syracuse. His current research interests are in distributed sensor networks and data fusion, detection and estimation theory, wireless communications, image processing, radar signal processing, and remote sensing. He has published extensively. He is the author of *Distributed Detection and Data Fusion* (Springer-Verlag, 1997). He has served as a consultant to several major companies. He was a James Scholar, a Bronze Tablet Senior, and a Fellow while at the University of Illinois. He is a member of Tau Beta Pi and is the recipient of the 1981 ASEE Dow Outstanding Young Faculty Award. He was elected to the grade of Fellow of the IEEE in 1997 for his contributions in the area of distributed detection and data fusion. He was Guest Editor of the January 1997 Special Issue on Data Fusion of *IEEE Proceedings*. In 2000, he received the Third Millennium Medal from the IEEE and the Chancellor's Citation for exceptional academic achievement at Syracuse University. He is the recipient of the IEEE 2012 Judith A. Resnik Award and a Doctor of Engineering degree honoris causa from Drexel University in 2014. He is on the Editorial Boards of the *Journal on Advances in Information Fusion* and *IEEE Signal Processing Magazine*. He was President of the International Society of Information Fusion in 2001.

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

Wireless Physical Layer Authentication via Fingerprint Embedding

Paul L. Yu, Gunjan Verma, and Brian M. Sadler

ABSTRACT

Authentication is a fundamental requirement for secure communications. In this article, we describe a general framework for fingerprint embedding at the physical layer in order to provide message authentication that is secure and bandwidth-efficient. Rather than depending on channel or device characteristics that are outside of our control, deliberate fingerprint embedding for message authentication enables control over performance trade-offs by design. Furthermore, low-power fingerprint designs enhance security by making the authentication tags less accessible to adversaries. We define metrics for communications and authentication performance, and discuss the trade-offs in system design. Results from our wireless software-defined radio experiments validate the theory and demonstrate the low complexity, practicality, and enhanced security of the approach.

FINGERPRINTING RADIO WAVEFORMS

We begin with an overview of intrinsic and intentionally embedded fingerprinting, and discuss the relationship with identification, communications, secrecy, and authentication. We then introduce a method of embedding fingerprints for wireless authentication that overcomes the deficiencies of using intrinsic fingerprints to identify the transmitter. Furthermore, the method has very small bandwidth requirements compared to traditional message authentication codes, making it a natural candidate for bandwidth-constrained environments such as mobile ad hoc networks (MANET).

INTRINSIC FINGERPRINTS

A fingerprint is literally the impression of a fingertip, but more broadly is a characteristic that identifies. This is often associated with an intrinsic property of uniqueness, or at least uniqueness viewed as a realization of a random process with structure. For example, identifying humans via biometrics now includes fingerprints, iris scans, DNA, voice features, and behavioral patterns [1]. Several investigators have considered applying these ideas to radio transmissions, including identification of radios based on signal transients [2], and study of vulnerability of these methods to impersonation [3]. In wired communications, identification of Ethernet cards has been demonstrated [4]. To be of practical use, fingerprints should be easily measurable with a sensor that is convenient and technologically feasible, and be robust to measurement noise. In addition, security features such as tamper resistance may be desirable, but these are not necessarily inherent in the fingerprint.

While these examples illustrate fingerprints derived from intrinsic structure, one may also derive a fingerprint from intrinsic randomness. For example, some wireless physical layer security techniques exploit the fact that a realization of a fading communications channel between any two agents is unique. Therefore, the channel realization may be used as a means of identifying the transmitter [5]. However, the channel properties are outside of our control, and can be noisy and rapidly time-varying, placing limits on the ability to systematically design for security.

EMBEDDED DEVICE FINGERPRINTS

Control of performance can be achieved by purposefully embedding a fingerprint in a device in a designed way so that each device can be uniquely identified. Now, in addition to the above characteristics of goodness, a good fingerprint will have strong security features, including the ability to defeat cloning and tampering. That is, a good fingerprint is not only unique and identifiable, but also hard to copy (spoof) or remove. Thus, for example, a manufacturer can label and recognize each individually manufactured device, while at the same time making it difficult to produce a counterfeit that cannot be differentiated from the genuine original. Needless to say, this has important implications for commercial enterprise.

Fingerprint embedding into devices is, of course, dependent on the specifics of the device, and so can take many forms. In an effort to make fingerprints uncloneable, some manufacturers have purposefully injected randomness into the manufacturing process to create unique characteristics [6]. Such an intrinsic fingerprint

The authors are with the U.S. Army Research Laboratory.

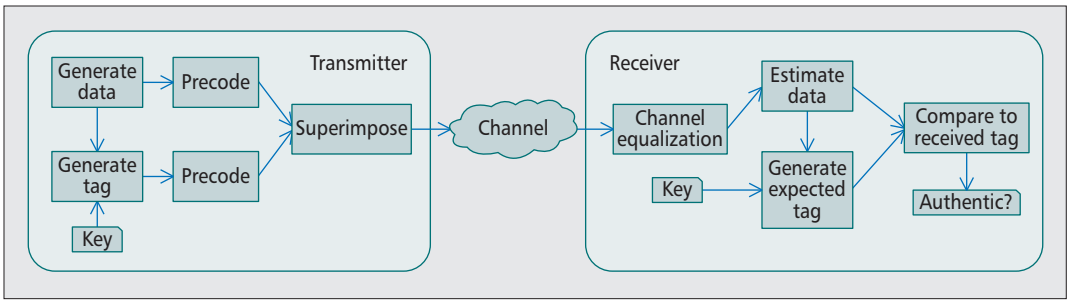


Figure 1. Physical layer authentication system diagram. The transmitter generates a data-dependent authentication tag and superimposes it with the data. The receiver estimates the data and generates the corresponding expected authentication tag. A matched filter test is performed to determine the presence of the tag in the signal and validate the transmitter's identity [8, Fig. 1].

Authentication defends against message tampering, and enables trust to be established between users. This is especially critical in wireless communications using an open and shared medium, where an adversary can eavesdrop, spoof, and jam.

can be measured and cataloged, like a serial number. The inherent randomness that is exploited for the intrinsic fingerprint may be uncontrollable at the micro-scale, and thus it may be very difficult or impossible to manufacture a similar device (a clone) with a prespecified fingerprint.

Embedded Fingerprints for PHY Authentication

Because a good fingerprint uniquely and securely identifies its source, it is an ideal candidate to convey authentication, the process of validating the identity of a message source while rejecting impersonation attacks by an adversary. Authentication defends against message tampering, and enables trust to be established between users. This is especially critical in wireless communications using an open and shared medium, where an adversary can eavesdrop, spoof, and jam.

Conventional authentication schemes transmit both the data message and a separate authentication message, referred to as a tag [7]. Each authentication tag is dependent on the associated data, and a secret key that is shared only between the transmitter and receiver. The tag generator employed at the transmitter is often a cryptographic hash function, the input of which is the key and the packet message, and its output is the tag. Such functions are highly nonlinear and difficult to invert, so an adversary cannot easily recover the key given the message and the associated tag.

In the conventional approach, the authentication tag is appended to the message, and both are transmitted at the same power. This has two disadvantages. One, the tag reduces spectral efficiency because it is time-multiplexed with the data. Two, the tag is available at high fidelity at the receiver.

Consequently, in this case, authentication security is solely predicated on encryption via the hash function, and in principle is susceptible to discovery if an adversary has sufficient computational resources. This motivates the use of an embedded fingerprint to carry the authentication tag.

An embedded fingerprint can be designed so that its bandwidth requirements are low, and its recovery is difficult for the adversary [8–10]. That is, we can arbitrarily decrease the ability of the adversary to observe the authentication tag by lowering the power of its embedding. As we show later, this leads to uncertainty about a secret key that is not readily defeated by an increase in the adversary's computational ability. Thus, embedding provides additional security and, unlike the conventional authentication approach, does not solely rely on cryptographic security. However, there is a design trade-off, because lowering the tag embedding power also weakens the ability of the intended receiver to authenticate valid packets. As we show below, good performance trade-offs are readily achieved in a software-defined radio (SDR) operating in a fading environment. The design trade-offs are fully characterized in [8–10].

Embedded Authentication Framework

The authentication system is diagrammed in Fig. 1. The transmitter (Alice) generates the authentication tag using the message and a shared secret symmetric key. The fingerprint is embedded in the transmission by adding the low-power tag to the message signal. The receiver (Bob) decodes the message and locally generates the expected authentication tag with his copy of the secret key. Finally, Bob validates the message as authentic if he detects the presence of the tag in the received signal.

Authentication System: Transmitter

For ease of presentation, we consider the case where Alice and Bob are each using single-antenna wireless devices communicating over a single carrier frequency [9]. The method readily generalizes to multi-carrier [10] and multi-antenna multiple-input multiple-output (MIMO) cases [8]. Alice has a message S to give to Bob, with whom she shares a secret key k. She first generates an authentication tag using a tag generating function, $T = g(S, k)$, which is based on a cryptographic hash function.[1] Then she transmits the weighted superposition of the message and tag signals,

$$X = \rho_S S + \rho_T T.$$

In order to make it difficult for the adversary to recover information about the authentication

[1] A cryptographic hash function is easy to compute, but infeasible to invert [7]. Furthermore, it is infeasible to find two messages that result in the same hash. RIPEMD-160 and SHA-2 are two examples.

tag, and to minimize self-interference, we set $\rho_S \gg \rho_T$ (i.e., the fingerprint has proportionally very low power). For any value of $\rho_T \in [0,1)$ we choose ρ_S so that the expected power of X remains constant, so we can regard ρ_S and ρ_T as power allocation percentages between the message and the fingerprint.

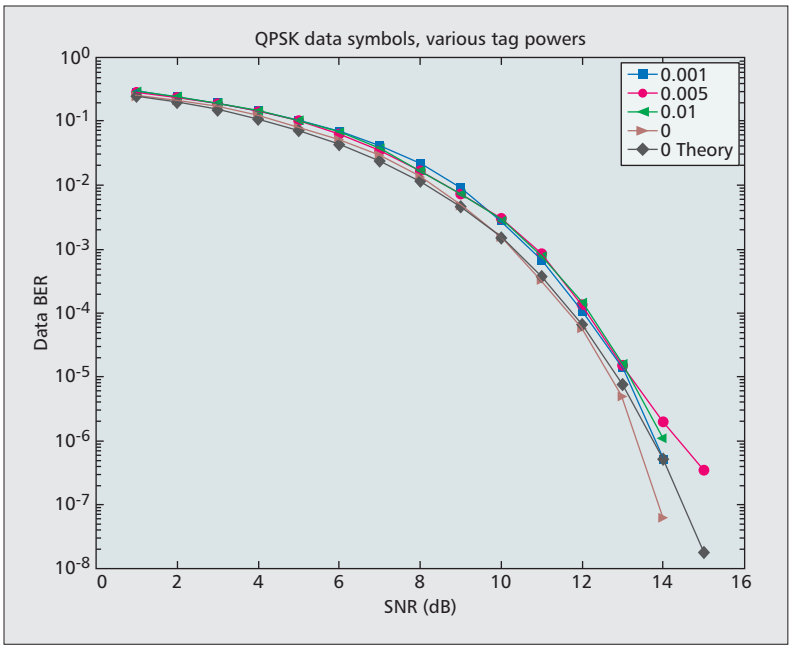


Figure 2. SDR experiment results. Low-powered authentication has minimal impact on the data bit error rate. Tag powers range from 0.1 to 1 percent of the transmit power. The 0 tag power case corresponds to the data only situation where no authentication is transmitted. The results show good agreement with the overlaid theoretical additive white Gaussian noise (AWGN) BER curve labeled 0 Theory.

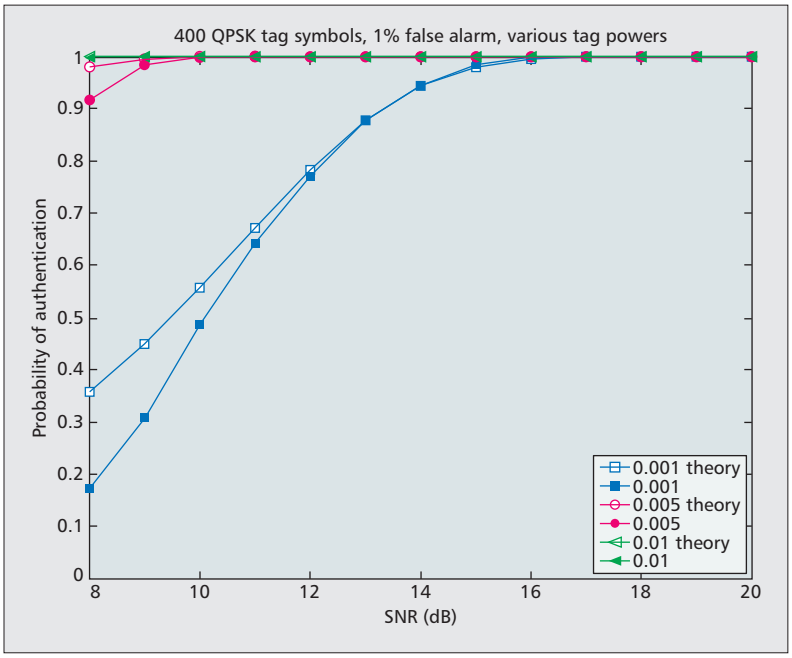


Figure 3. SDR experiment results. Authentication probability for various tag powers (from 0.1 to 1 percent of the transmit power). Packets contain 400 QPSK symbols, and the authentication false alarm probability is 1 percent. Higher-powered tags have high authentication performance.

Authentication System: Receiver

The receiver processing and authentication steps are shown in Fig. 1. Bob reverses the effect of the channel through equalization, and proceeds to demodulate and decode the message as usual. The fingerprint need not be embedded in pilot symbols so as to avoid interference with channel estimation and equalization. Because the authentication tag T is data-dependent, the receiver is unable to remove it prior to decoding the data, and thus it acts as interference during data recovery. However, by keeping the tag power relatively low, we show that message recovery is minimally impacted.

Referring again to Fig. 1, having estimated the data, the receiver can now proceed to complete the authentication process. Bob uses his shared secret key and the received data to generate the tag he expects to see. After removing the recovered data from the received signal to form a residual signal, the receiver determines the presence or absence of the authentication tag via a matched filter test comparing the expected tag and the residual signal [9, eq. 20].

The performance of the authentication test is determined by the energy of the authentication tag, which is under our design control through the tag length and the tag power allocation. Analysis and experiments show that for even moderate message packet sizes, the correlation test statistic is well approximated as having a Gaussian distribution, so the test threshold is readily set to achieve desired false alarm or detection probability [8–10].

The received tag is in the residual signal, and so is noisy. We rely on the matched filter test to overcome the noise to achieve reliable authentication. Cryptographic hash functions are used so that tags generated for distinct messages are generally far apart in Hamming distance, even when the messages are close in Hamming distance. This feature ensures strong resistance to spoofing attacks [7].

Theory and Experiments

We quantify the performance of our embedded fingerprinting approach by considering the effect of the fingerprint on the data demodulation, and the ability of the receiver to authenticate packets. In the following we present single-antenna single-carrier experiments using NI-USRP SDRs [11, 12], and compare these with theoretical predictions [9]. We use two USRP1 devices under MATLAB control [13], at a frequency of 2.39 GHz, employing quadrature phase shift keying (QPSK) modulation for the data and the tag.[2] The two radios are placed about 15 ft apart in an office building with many scatterers in the scene. By scaling the power of the transmitter, various receive signal-to-noise ratio (SNR) levels are attained, and between 25,000 and 30,000 packets are transmitted at each level under test.

[2] In this experiment we employ symbol synchronous messages and tags, which simplifies the implementation. However, the tag may be inserted into the transmit waveform in a variety of ways. For example, it need not be additive, which is a topic for further research.

Impact on Data BER

The superimposed tag takes power away from the data signal and acts as interference to data modulation. When the tag is superimposed at low power, the interference to data demodulation may be modeled as an increase in noise (i.e., as a decrease in data SNR). For example, suppose that a given channel realization yields 10 dB received SNR. If the tag uses 1 percent of the power, the data SNR becomes 9.94 dB. Hence, the data bit error rate (BER) is essentially unchanged at such a low authentication power, and the interference caused by the tag is minimal. Simulation and experimental results confirm this to be the case [8–10].

Figure 2 shows the impact of the authentication on the data BER for an over-the-air SDR experiment. For the tag power allocations tested (0.1, 0.5, and 1 percent), the BER curves are, for practical purposes, coincident. The theoretical and experimental additive white Gaussian noise BER curves (with no tag present) are also shown for comparison, validating the experimental results. We next show that while the change in data BER is essentially negligible for small tag powers, these low tag power levels are sufficient for robust authentication.

Authentication Performance

Figure 3 shows the experimental and theoretical authentication performance for various tag powers. The packets contain 400 QPSK data symbols and a corresponding 400 QPSK symbol tag, with tag power ranging from 0.1 to 1 percent of the transmit power. The authentication test threshold was set to achieve a 1 percent false alarm probability. This figure shows the effect of changing the tag power while holding the packet length constant. As previously discussed, authentication performance is improved by increasing the tag energy. Additional experiments show that modifying the packet length also yields very good agreement between theory and experiment. Other design variations include spreading the tag over multiple packets.

Security

In this section we quantify the benefit of transmitting a low-power tag in terms of how well it preserves the key's secrecy. Although Eve does not initially have the secret key k shared by Alice and Bob, she gains key information by observing their communications [14]. If she has complete knowledge of k and the tag generating function, she is able to impersonate Alice by generating legitimate tags for her messages. The protection of the key is therefore crucial to the security of the authentication system.

In the following we quantify the effort required for the adversary to learn the secret key from embedded fingerprints. We assume that Eve, just like Bob, is able to recover the data from her observations without error. Furthermore, we assume that Eve knows the tag generating function that defines the dependence of the tag on both the message and the secret key. Knowledge of the tag generation function implies knowledge of the set of possible keys, for example, the set of integers between 0 and $2^{32} - 1$, corresponding to 32-bit keys.

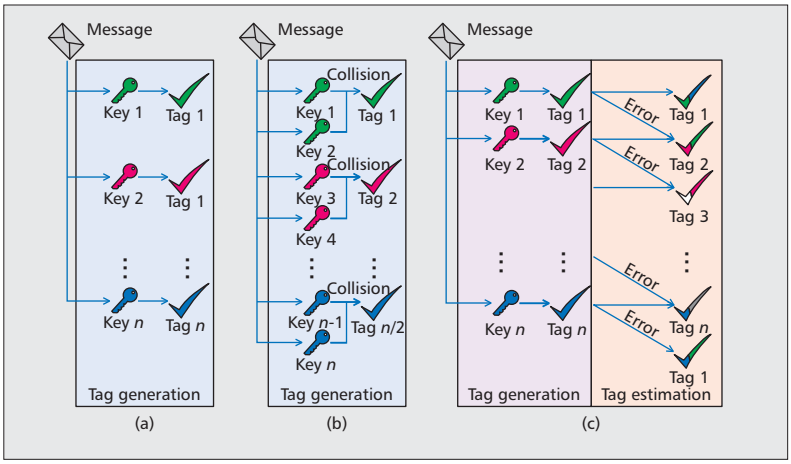


Figure 4. Illustration of key equivocation with embedded fingerprints. The tag generation is considered for a single fixed message for each panel: a) each key maps the message to a unique tag; knowledge of the message and tag leaves no uncertainty as to which key was used, and hence key equivocation is zero (key equivocation = 0 bits, noise-free tags); b) exactly two keys collide to map the message to the same tag, leading to uncertainty about which key (of the same color) was used even when the message and tag are known without error (key equivocation = 1 bit, noise-free tags; c) although the tag mapping is unique, exactly two tags can result in the same tag estimate, again leading to uncertainty about which tag was actually transmitted (key equivocation = 1 bit, noisy tags). Hence, positive key equivocation arises in both b and c.

The key equivocation, or conditional entropy [15], is a measure of Eve's uncertainty about the secret key given her observations and presumed infinite computational resources. This bounds the ability of the adversary to attack the authentication system [14]. Equivocation is non-negative and bounded from above by the entropy of the secret key. Zero key equivocation implies that Eve has no uncertainty about the key. Having no uncertainty means that there is only one key that fits the observations, although it may require a great deal of computation to ascertain its value. Zero key equivocation is the worst case for security because brute-force attacks are guaranteed to succeed (although the time required may be very long).

Figure 4 illustrates how positive key equivocation arises and how noise increases protection of the secret key. We consider the set of possible tags that can be generated from a fixed message and a set of n keys. In Fig. 4a, each key maps the message to a unique tag. Knowledge of the message and tag (as in a conventional scheme) leaves no uncertainty as to which key was used, and hence key equivocation is zero. In Fig. 4b, each key collides with exactly one other key to map the message to the same tag. Therefore even with knowledge of the message and tag, there is uncertainty about which of the colliding keys was used (e.g., key 1 vs. key 2). Finally, in Fig. 4c, we consider the effect of tag estimation on the zero key equivocation system in Fig. 4a, as occurs when the tag is embedded and only a noisy observation is available to the adversary. Estimation errors lead to uncertainty about which tag was transmitted, and hence which key was used.

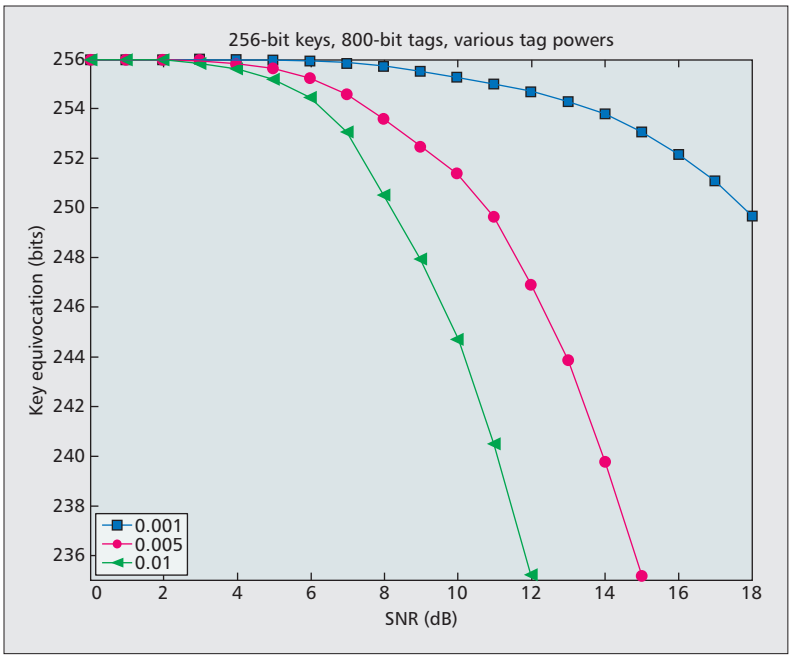


Figure 5. SDR experiment results. Key equivocation for various tag powers (from 0.1 to 1 percent of the transmit power). Low-powered tags have high key equivocation.

Intuitively, the more uncertainty the adversary has about which tag is present in the signal, the more uncertainty she has about the secret key. It then follows that lowering the tag power to present only a very noisy tag to the adversary results in high key equivocation. Although the same signal (modulo channel effects) is presented to both the intended receiver and adversary, it should be noted that the two parties have very different goals. The intended receiver has a detection problem: deciding if the tag corresponding to the received data and his key is present (corresponding to 1 bit of information). The eavesdropper has the much harder estimation problem: determining the transmitted tag and then the secret key (requiring $\log(n)$ bits of information). Because Bob is only trying to make a 1-bit decision, we can set the authentication power quite low without significantly degrading his authentication performance, while at the same time severely limiting the ability of Eve to extract key information.

Figure 5 shows experimental results for the key equivocation at various tag powers. The key equivocation is calculated based on the observed tag BER for each SNR. This figure depicts the case where key k has 256 bits and authentication tag T has 800 bits. We assume that there is zero key equivocation in the noiseless case, that is, each (message, tag) pair is associated with a unique key. In terms of equivocation this is the worst case scenario, so typical results will be better than those shown in Fig. 5.

Note that higher received SNR decreases the key equivocation. Intuitively, a cleaner observation leads to less uncertainty of the tag and hence the key. For the scenarios of interest (low tag power), the key equivocation is seen to be very high as a proportion of its 256-bit maximum.

Also note that lower tag power increases the key equivocation. As with the effect of received SNR, reducing the tag power reduces the ability of the receiver to make an accurate estimate of the tag. In this example, a large increase in key equivocation is apparent when reducing the tag power from 1 to 0.1 percent. As shown in Fig. 3, reducing the tag power does impact the ability of the intended receiver to authenticate properly. Hence, a design balance is sought to achieve the desired authentication performance while maintaining a high level of security.

Conclusions

Fingerprint embedding provides a flexible framework for message authentication, increasing security by obscuring the authentication tag to the adversary and saving transmission bandwidth. The method is readily adapted to software-defined radio, with low complexity. SDR experiments validate the theory, which enables controlled design of desired system trade-offs. There is ample room for the designer to choose the appropriate operating point that balances authentication probability and key equivocation because the impact on data BER is shown to be so slight for tag powers as high as 1 percent. The tag power and length are two parameters than can be chosen to satisfy the design requirements for arbitrary modulation schemes for the fingerprint. For example, suppose that the message is 400 symbols in length and the design requires > 99 percent authentication probability, > 250 bits of key equivocation, and $< 10^{-3}$ message BER at 10 dB SNR. Then, from Figs. 2, 3, and 5, we see that setting $\rho_T = 0.5$ percent and using 400 QPSK symbols for the tag satisfy the requirements. Alternatively, one may also decrease ρ_T and increase the length of the tag to satisfy the same requirements.

References

[1] A. K. Jain, A. A. Ross, and K. Nandakumar, *Introduction to Biometrics*, Springer, 2011.
[2] B. Danev and S. Čapkun, "Transient-Based Identification of Wireless Sensor Nodes," *Proc. ACM/IEEE IPSN*, 2009.
[3] B. Danev *et al.*, "Attacks on Physical-Layer Identification," *Proc. 3th ACM Conf. Wireless Network Security*, 2010, pp. 89–98.
[4] R. Gerdes *et al.*, "Physical-Layer Identification of Wired Ethernet Devices," *IEEE Trans. Info. Forensics Security*, vol. 7, no. 4, Aug. 2012, pp. 1339–53.
[5] L. Xiao *et al.*, "Using the Physical Layer for Wireless Authentication in Time-Variant Channels," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, July 2008, pp. 2571–79.
[6] U. Ruhrmair, S. Devadas, and F. Koushanfar, "Security Based on Physical Unclonability and Disorder," *Introduction to Hardware Security and Trust*, M. Tehranipoor and C. Wang, Eds. Springer, 2011.
[7] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography*, CRC Press, 2001.
[8] P. Yu and B. Sadler, "MIMO Authentication via Deliberate Fingerprinting at the Physical *Layer," IEEE Trans. Info. Forensics Security*, vol. 6, no. 3, 2011, pp. 606–15.
[9] P. Yu, J. Baras, and B. Sadler, "Physical-Layer Authentication," *IEEE Trans. Info. Forensics Security*, vol. 3, no. 1, Mar. 2008, pp. 38–51.
[10] —, "Multicarrier Authentication at the Physical Layer," *Proc. Int'l. Symp. on a World of Wireless, Mobile and Multimedia Networks*, 2008, pp. 1–6.
[11] Ettus Research — Product Detail, Mar. 2013; https://www.ettus.com/product/details/USRP-PKG
[12] G. Verma, P. Yu, and B. Sadler, "Physical Layer Authentication via Fingerprint Embedding Using Software-Defined Radios," *IEEE Access*, vol. 3, 2015, pp. 81–88.

[13] G. Verma and P. L. Yu, "A MATLAB Library for Rapid Prototyping of Wireless Communications Algorithms with the USRP Radio Family," U.S. Army Research Lab., tech. rep., 2013.
[14] U. Maurer, "Authentication Theory and Hypothesis Testing," *IEEE Trans. Info. Theory*, vol. 46, no. 4, July 2000, pp. 1350–56.
[15] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley-Interscience, 1991.

Biographies

Paul L. Yu [M] received B.S. degrees in mathematics and computer engineering, and a Ph.D. degree in electrical engineering. He is with the U.S. Army Research Laboratory (ARL), where his work is in the areas of signal processing, wireless communications, and network science.

Gunjan Verma received B.S. degrees in mathematics and computer science, and a B.A. degree in economics from Rutgers University, and an M.S. degree in applied mathematics from Johns Hopkins University. He is with the U.S. ARL, where his work spans information flow in complex networks, signal processing, and validation and prototyping of wireless communications algorithms using software-defined radios.

Brian M. Sadler [S'81, M'81, SM'02, F'07] received B.S. and M.S. degrees from the University of Maryland, College Park, and his Ph.D. degree from the University of Virginia, Charlottesville, all in electrical engineering. He is a Fellow of the ARL in Adelphi, Maryland. He is an Associate Editor for *EURASIP Signal Processing*, was an Associate Editor for *IEEE Transactions on Signal Processing* and *IEEE Signal Processing Letters*, and has been a Guest Editor for several journals including *IEEE JSTSP*, *IEEE JSAC*, *IEEE Signal Processing Magazine*, and the *International Journal of Robotics Research*. He is a member of the IEEE Signal Processing Society Sensor Array and Multi-Channel Technical Committee, and Co-Chair of the IEEE Robotics and Automation Society Technical Committee on Networked Robotics. He received Best Paper Awards from the Signal Processing Society in 2006 and 2010. He has received several ARL and Army R&D awards, as well as a 2008 Outstanding Invention of the Year Award from the University of Maryland. His research interests include information science, networked and autonomous systems, sensing, and mixed-signal integrated circuit architectures.

ENERGY HARVESTING COMMUNICATIONS: PART 2






Chau Yuen *Maged Elkashlan* *Yi Qian* *Trung Q. Duong* *Lei Shu* *Frank Schmidt*

Over the last decade, energy harvesting has emerged as a promising approach to enable self-sufficient and self-sustaining operation for devices in energy-constrained networks by scavenging energy from the ambient environment to power up devices.

In particular for wireless sensor networks, if sensors, which spread throughout a home or factories, in buildings, or even outdoors, are powered by energy harvesting, there are no batteries to replace and no labor costs associated with replacing them. For a cellular network, energy harvesting can be used to provide power in many elements of a telecom network, saving considerable costs in electricity supply, and providing low maintenance monitoring. As another important focus, RF energy is currently broadcast from billions of radio transmitters around the world. The ability to harvest RF energy from ambient or dedicated sources enables wireless charging of low-power devices and has significant benefits for product design, usability, and reliability.

This Feature Topic focuses on energy harvesting related issues in communications, through presenting a holistic view of research challenges and opportunities in the emerging area of energy harvesting communications. We hope this Feature Topic is able to help readers obtain better understanding of some key issues in energy harvesting, and drive more research interests. The Feature Topic has three parts, with 10 accepted papers presented in Part 1, nine accepted papers included in Part 2, and four accepted papers to appear in Part 3.

The second part of this Feature Topic starts with an article "Toward Self-Sustainable Cooperative Relays:State-of-the-Art and the Future," by Kuang-Hao Liu *et al.*, examines the current progress in energy harvesting relays with special emphasis on wireless power transfer through RF signals that carry both information and energy at the same time.

The article "RF-Powered Cellular Networks: Key Challenges and Solution Techniques," by H. Tabassum *et al.*, surveys the related research advancements in RF-powered cellular networks and their limitations, as well as design considerations for RF-powered cellular networks that can potentially tackle the major challenges and open up new research directions.

The article "Wireless Energy Harvesting in Interference Alignment Networks," written by Nan Zhao *et al.*, presents an overview of wireless energy harvesting in interference alignment networks, and provides a unified framework to jointly study wireless energy harvesting and interference alignment.

The article "A Survey of Energy Harvesting Communications: Models and Offline Optimal Policies," by Yejun He *et al.*, reviews the different ways of harvesting the ambient energy in energy harvesting communications and the models of energy harvesting communications.

The article "Cutting the Last Wires for Mobile Communication by Microwave Power Transfer," by Kaibin Huang *et al.*, provides an introduction to wireless powered communications by describing the key features of wireless powered communications, shedding light on a set of frequently asked questions, and identifying the key design issues and discussing possible solutions.

The article "Energy Harvesting Small Cell Networks: Feasibility, Deployment and Operation," by Yuyi Mao *et al.*, conducts a comprehensive study of energy harvesting small cell networks, and investigates important aspects, including the feasibility analysis, network deployment, and network operation issues.

The article "Wireless Energy Harvesting for Internet of Things," written by Pouya Kamalinejad *et al.*, summarizes enabling technologies for efficient wireless energy harvesting units, analyzes the lifetime of wireless energy harvesting enabled Internet of Things devices, briefly studies future trends in the design of efficient wireless energy harvesting systems, and specifies research challenges that lie ahead.

The article "Joint Wireless Information and Energy Transfer in Massive Distributed Antenna Systems," by Fangchao Yuan *et al.*, discusses the research opportunities in the joint wireless information and energy transfer in massive distributed antenna system.

The article "When Telecommunication Networks Meet Energy Grids: Cellular Networks with Energy Harvesting

and Trading Capabilities," written by Davide Zordan *et al.*, presents recent developments in energy harvesting, the way future energy markets are expected to evolve, and the new fundamental trade-offs that arise when energy can be traded.

BIOGRAPHIES

CHAU YUEN (yuenchau@sutd.edu.sg) received his B. Eng and Ph.D. degrees from Nanyang Technological University, Singapore, in 2000 and 2004, respectively. He was a postdoctoral fellow at Lucent Technologies Bell Labs, Murray Hill, New Jersey, during 2005. He was a visiting assistant professor at Hong Kong Polytechnic University in 2008. During the period of 2006–2010, he worked at the Institute for Infocomm Research, Singapore, as a senior research engineer. He joined Singapore University of Technology and Design as an assistant professor in June 2010. He serves as an Associate Editor for *IEEE Transactions on Vehicular Technology* and was awarded Top Associate Editor for three consecutive years. In 2012, he received the IEEE Asia-Pacific Outstanding Young Researcher Award. He has held positions on several conference organizing committees, and is on Technical Program Committees of various international conferences.

MAGED ELKASHLAN received his Ph.D. degree in electrical engineering from the University of British Columbia, Canada, in 2006. From 2006 to 2007, he was with the Laboratory for Advanced Networking at the University of British Columbia. From 2007 to 2011, he was with the Wireless and Networking Technologies Laboratory at the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia. He also held an adjunct appointment at the University of Technology Sydney, Australia, between 2008 and 2011. In 2011, he joined the School of Electronic Engineering and Computer Science at Queen Mary, University of London, United Kingdom, as an assistant professor. His research interests include millimeter wave communications, energy harvesting, cognitive radio, and wireless security. He currently serves as an Editor for *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Vehicular Technology*, and *IEEE Communications Letters*. He received Best Paper awards at IEEE ICC '14, International Conference on Communications and Networking in China in 2014, and IEEE VTC-Spring 2013. He received the Exemplary Reviewer Certificate of *IEEE Communications Letters* in 2012.

YI QIAN [M'95, SM'07] is an associate professor in the Department of Electrical and Computer Engineering, University of Nebraska-Lincoln (UNL). Prior to joining UNL, he worked in the telecommunications industry, academia, and the government. Some of his previous professional positions include serving as a senior member of scientific staff and technical advisor at Nortel Networks, a senior systems engineer and technical advisor at several startup companies, an assistant professor at the University of Puerto Rico at Mayaguez, and a senior researcher at the National Institute of Standards and Technology. His research interests include information assurance and network security, network design, network modeling, simulation and performance analysis for next generation wireless networks, wireless ad hoc and sensor networks, vehicular networks, smart grid communication networks, broadband satellite networks, optical networks, high-speed networks, and the Internet. He has a successful track record in leading research teams and publishing research results in leading scientific journals and conferences. Several of his recent journal articles on wireless network design and wireless network security are among the most accessed papers in the IEEE Digital Library. He is the current Chair of the Communications and Information Security Technical Committee in the IEEE Communications Society. He is an IEEE Distinguished Lecturer.

TRUNG Q. DUONG received his Ph.D. degree in telecommunications systems from Blekinge Institute of Technology (BTH), Sweden, in 2012, and then continued working at BTH as a project manager. Since 2013, he has joined Queen's University Belfast, United Kingdom, as a lecturer (assistant professor). He held visiting positions at Polytechnic Institute of New York University and Singapore University of Technology and Design in 2009 and 2011, respectively. His current research interests include cooperative communications, cognitive radio networks, green communications, physical layer security, massive MIMO, cross-layer design, mmWave communications, and localization for radios and networks. He has been a TPC chair for several IEEE international conferences and workshops, including most recently the IEEE GLOBECOM '13 Workshop on Trusted Communications with Physical Layer Security. He currently serves as an Editor for *IEEE Communications Letters* and Wiley *Transactions on Emerging Telecommunications Technologies*. He served as Lead Guest Editor of the Special Issue on Location Awareness for Radios and Networks of the *IEEE Journal on Selected Areas in Communications*, Lead Guest Editor of the Special Issue on Secure Physical Layer Communications of *IET Communications*, Guest Editor of the Special Issue on Green Media: Toward Bringing the Gap between Wireless and Visual Networks of *IEEE Wireless Communications*, Guest Editor of the Special Issue on Millimeter Wave Communications for 5G of *IEEE Communications Magazine*, Guest Editor of the Special Issue on Cooperative Cognitive Networks of the *EURASIP Journal on Wireless Communications and Networking*, and Guest Editor of the Special Issue on Security Challenges and Issues in Cognitive Radio Networks of the *EURASIP Journal on Advances Signal Processing*. He was awarded the Best Paper Award at IEEE VTC-Spring '13 and the Exemplary Reviewer Certificate of *IEEE Communications Letters* in 2012.

LEI SHU [M] received his B.Sc. degree in computer science from South Central University for Nationalities, China, in 2002, his M.Sc. degree in computer engineering from Kyung Hee University, Korea, in 2005, and his Ph.D. degree from the Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland, in 2010. Until March 2012, he was a specially assigned researcher in the Department of Multimedia Engineering, Graduate School of Information Science and Technology, Osaka University, Japan. He is a member of IEEE IES, IEEE ComSoc, EAI, and ACM. In October 2012, he joined Guangdong University of Petrochemical Technology, China, as a full professor. In 2013, he started to serve Dalian University of Technology as a Ph.D. supervisor in the College of Software, Beijing University of Posts and Telecommunications as a Master's supervisor in information and communication engineering, Wuhan University as a Master's supervisor in the College of Computer Science, guest professor at Tianjin University of Science and Technology, and a guest researcher at Guangzhou Institute of Advanced Technology, Chinese Academy of Sciences. Meanwhile, he is also working as vice-director of the Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis, China. He is the founder of the Industrial Security and Wireless Sensor Networks Lab. HIs research interests include wireless sensor networks, multimedia communication, middleware, fault diagnosis, and security. He has published over 230 papers in related conferences, journals, and books in the area of sensor networks. Currently, his H-index is 21 in Google Citation. Total citations of his papers by other people are more than 1600. He developed an open source wireless sensor networks simulator, NetTopo, to evaluate and demonstrate algorithms. NetTopo has been downloaded more than 3420 times over the past three years, and is widely used by international researchers and students. He was awarded the MASS 2009 IEEE TCs Travel Grant and the Outstanding Leadership Award of EUC 2009 as Publicity Chair, GLOBECOM 2010, and ICC 2013, the ComManTel 2014 Best Paper Award, and the Outstanding Service Award of IUCC 2012 and ComcomAP 2014. He also received a few more awards from the Chinese government: Top Level Talents in "Sailing Plan" of Guangdong Province, China, and Outstanding Young Professor of Guangdong Province, China. He has been serving as Editor-in-Chief for *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, and Associate Editor for *IEEE Access*, ACM/Springer *Wireless Networks*, *Journal of Network and Computer Applications*, *Transactions on Emerging Telecommunications Technology*, and several other publications. He has served as Co-Chair for many international conferences. He has obtained more than 4 million RMB in research grants since October 2012.

FRANK SCHMIDT is a pioneer in energy harvesting and the visionary in the management team of EnOcean. As chief technology officer he is responsible for the overall technical orientation, patent related activities, as well as the relationship management with educational, research and scientific organizations. Before joining EnOcean he was at the Central Research Department of Siemens AG, where he created self-powered wireless sensor technology as early as 1995. He has been granted more than 40 patents for his energy harvesting inventions and is the author of numerous technical publications in this field. He is a physicist and studied at the Technical University of Chemnitz, Germany.

Toward Self-Sustainable Cooperative Relays: State of the Art and the Future

Kuang-Hao Liu and Phone Lin

ABSTRACT

Cooperative relaying has been recognized as a promising technique to exploit spatial diversity gain using simple hardware, but it may barely fall in the class of green communications because extra power is consumed for information relaying. Recent advances in energy harvesting devices have paved the way for self-sustainable relays that power themselves by replenishing ambient energy without wires. This article examines the current progress of EH relays with special emphasis on wireless power transfer through RF signals that carry both information and energy at the same time. In this context, information relaying and EH are two tightly coupled processes, and practical constraints of EH devices pose new challenges to self-sustainable relays. This article addresses these issues and points out future research directions.

INTRODUCTION

In the course of developing future wireless communication systems, achieving high energy efficiency has been considered as important as providing high-speed reliable communications. Among numerous potential technologies, cooperative relaying has been recognized as a promising solution, whereby intermediate nodes act as relays to assist the communication between two distant nodes. By dividing a long communication link into two shorter ones, cooperative relaying can effectively mitigate severe path loss and shadowing effects without the need to increase transmission power. As an alternate form of automatic repeat request (ARQ), cooperative relaying also enhances transmission reliability via spatial diversity. To implement cooperative relaying, two options have been widely considered: standalone relay stations, which dedicate their resources to forward signals, and altruistic nodes, which share their resources to help forward others' signals. Traditional ways of using these relays, referred to as *conventional* relays, commonly focus on achieving high diversity gain at the cost of extra energy consumption. Conventional relays also rely on fixed power supply or replaceable batteries, increasing the burden of wide deployment.

In this article we address the issue of self-sustainable cooperative relays that are free from frequent battery replacement or power cables and do not add extra energy cost to perform tasks. These features are attractive to low-power sensors or embedded systems and are made possible thanks to the recent advances in energy harvesting (EH) devices [1]. An EH device can harvest energy from ambient sources such as solar, wind, and RF. Such devices intelligently manage themselves to perform designated functionalities using the harvested energy. While replacing traditional relays by EH devices appears to be attractive for self-sustainable relays, the fundamental difference between EH relays and traditional elays poses new challenges that hinder the immediate application of EH devices to cooperative communication networks. The goal of this article is to bridge the gap between conventional relays with fixed power supplies and EH relays with self-sustainable energy. Among various ambient energy sources, we focus on wireless power transfer (WPT) considering a unique characteristic of RF signals; that is, they inherently carry both information and energy. Hence, it is a natural choice for relays to charge themselves by extracting energy from the signals they overhear.

This article is organized as follows. A brief introduction to the physical constraints of EH devices is first given, followed by an overview of existing receiver designs for EH devices. Then we extend our discussion to multi-relay networks, where relay selection (RS) is discussed, as it has been recognized as the simplest but most effective approach to exploit cooperative diversity gain. A performance evaluation is conducted that compares two well-known RS schemes with a new RS scheme tailored for EH relays. The conclusion highlights some key issues that deserve further investigations.

PHYSICAL CONSTRAINTS

Deploying self-sustainable EH devices in communication networks faces numerous challenges due to the physical constraints of energy sources and EH hardware. Below, we list those constraints relevant to EH relays based on WPT.

- Causality: The energy harvested at present can only be used in future transmission.
- Half-duplex: EH and information transmission cannot take place concurrently.

Kuang-Hao Liu is with National Cheng Kung University.

Phone Lin is with National Taiwan University.

0163-6804/15/$25.00 © 2015 IEEE

- Energy storage: The energy storage is limited in size.
- Energy conversion: The energy picked by the harvester needs to be converted to DC voltage. Consequently, a certain amount of energy may be lost during energy conversion.

The first constraint is imposed by physical laws; the last three are subject to hardware capabilities. Their impacts on the system design for WPT are explained as follows.

The *causality constraint* sets the upper limit on the transmission power level of EH devices, and thus is the major factor that limits achievable transmission rates of EH devices based on WPT. To efficiently utilize the harvested energy, the transmission policy needs to be carefully designed. The transmission policy specifies how the energy source regulates its transmission power in feeding EH devices according to the channel variations due to propagation loss and multi-path fading in order to maximize the efficiency of WPT. The transmission policy should avoid overwhelming (aggressively charging) or underutilizing (slowly charging) the energy storage, both resulting in low utilization of harvested energy. However, the energy arrival rate and magnitude are often unknown a priori. Most likely, only statistical information is available, which makes it challenging to design the transmission policy. On the other hand, the transmission policy of EH devices needs to intelligently decide the transmission power and duty cycle according to the channel conditions and/or the status of the energy storage.

The *half-duplex constraint* arises from circuit limitation. Generally the energy harvester is responsible for picking up the RF energy and converting it to DC voltage, but it cannot perform information decoding (including down-conversion, analog-digital conversion, and demodulation). Hence, the receiver is not able to harvest energy when it is receiving signals. We note that here the half-duplex constraint differs from that which appeared in conventional half-duplex communications, which prohibits a device from transmitting and receiving at the same time.

In terms of *energy storage*, supercacitors or rechargeable batteries are commonly used to store the harvested energy. By storing and accumulating the harvested energy, the transmit power level is no longer limited to the harvested energy per round and thus loosens the causality constraint. However, the energy storage is limited in reality so energy will be wasted when the harvested rate exceeds the consumption rate. As to *energy conversion efficiency*, typical values for WPT range from 15 percent to 40 percent [2], depending on the environment (e.g. indoor or outdoor) and the implementation of the energy harvester. This again emphasizes the necessity for energy storage because only partial energy picked by the energy harvester is accessible.

To embrace the inherent energy of RF signals, the key enabler lies in an efficient receiver design, which allows for performing energy harvesting and information decoding from the same signal. In the next section we focus on the receiver design for EH devices by first reviewing the state-of-the-art progress. Another important issue, namely resource allocation for EH wireless systems, is not explicitly covered in this article but a rigorous treatment can be found in [3].

Receiver Architecture

As mentioned above, an EH receiver needs to perform two different tasks: EH and information decoding. Depending on how these two tasks are arranged (parallel, orthogonal, or in combination), below we introduce four types of receiver architecture for EH devices.

Ideal Receiver

To extract the power radiated by the RF signals, an energy harvester circuit consists of an impedance matching network to ensure maximum power transfer as well as a rectifier circuit to convert the RF signal to a DC voltage [4]. This energy harvester circuit is different from that of the information decoder, which performs down-conversion and sampling to decode the baseband signal. An ideal receiver assumes a perfect integration between the energy harvester and the information decoder without any power loss [5].

This energy harvester circuit is different from that of the information decoder, which performs down-conversion and sampling to decode the baseband signal. An ideal receiver assumes a perfect integration between the energy harvester and the information decoder without any power loss.

Separated Receiver

Since it is difficult to implement the ideal receiver, a common practice is to maintain two dedicated circuits that separately perform EH and information decoding. Below we introduce two classes of separated receiver design for EH nodes widely considered in the literature [6. 7].

In the first class of separated receiver architecture, the energy harvester and information decoder perform their tasks in different time, leading to the so-call time-switching (TS) receiver. As to the second class of separated receiver architecture, referred to as power-splitting (PS), EH and information decoding can take place in parallel. This is realized by a splitter that divides the incoming energy flow into two segments, one feeding into the energy harvester and the other into the information decoder. In comparing the implementation complexity of these two receivers, TS can be readily used with off-the-shelf devices but requires accurate time synchronization to instantly switch the receiver operation. On the other hand, PS needs an efficient power splitter to perform energy harvesting and information decoding simultaneously.

When these two types of receiver architecture are applied to EH relays, the cooperative relaying protocol can be designed as follows: define T as one transmission cycle, which consists of two periods, including the EH period and the information transmission period. For the TS receiver, denote θ as the fraction of time occupied by the energy harvester within a transmission cycle. Then the remaining time is equally shared by the source and the relay to send and forward the signal, respectively, as shown in Fig. 1a. The energy harvested during θT time will be used by the relay to forward the source signal to the destination node in $(1 - \theta)T/2$ time. One can see an obvious trade-off between energy harvesting and information delivery, called the rate-energy trade-off. The more time spent on collecting energy, the less time left for information transmission that decreases the achievable rate, and vice-versa.

For conventional relays, CSI-based RS has been proven to be optimal in terms of DMT. The best relay in CSI-based RS is defined as the one with the superior end-to-end channel condition because this relay offers the highest achievable rate among the others.

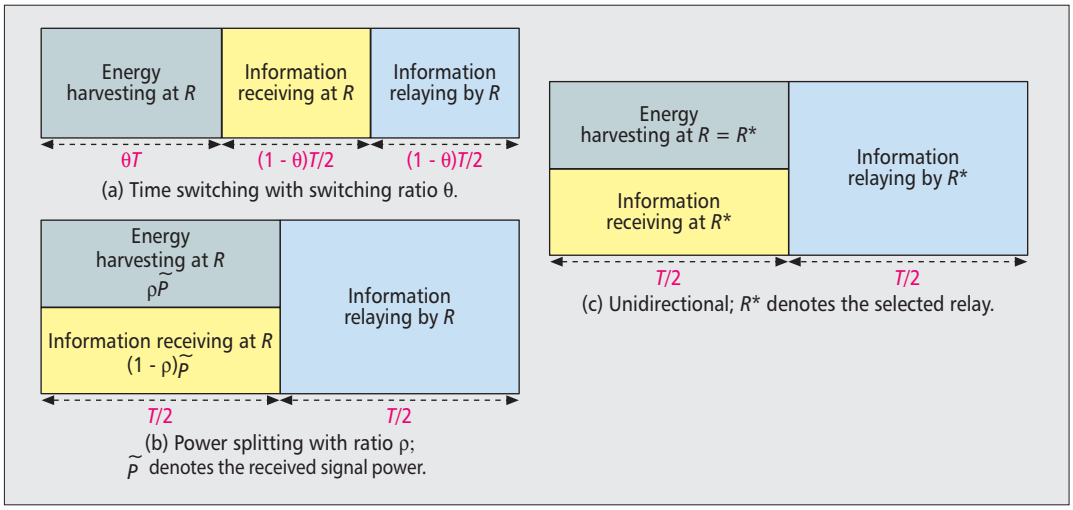


Figure 1. Cooperative relaying protocol based on different receiver architectures: a) time switching with switching ratio θ; b) power splitting with splitting ratio ρ; $\widetilde{P}$ denotes the received signal power; and c) unidirectional; R^* denotes the selected relay.

On the other hand, Fig. 1b depicts the timeline of cooperative relaying based on PS receiver architecture, where a proportion ρ of the received signal power $\widetilde{P}$ is used for EH, and the remaining power $(1-\rho)\widetilde{P}$ is reserved for the information decoder. Similar to the TS architecture, the power splitting ratio ρ is relevant to the performance of EH relays and presents a tradeoff. Increasing ρ is beneficial to information relaying as more energy will be available to relay transmission, but it jeopardizes information decoding because of the degraded signal strength. Compared with TS, PS allocates the same amount of time to EH and source transmission. However, only partial signal energy is accessible to the energy harvester and information decoder when using the PS receiver, while the TS counterpart can fully utilize the signal energy.

To quickly assess how these two receivers perform, let us evaluate their throughput, which is defined as the amount of correctly transmitted information per channel use. Denote $\mathcal{R}$ to be the source transmission rate in bits/sec/Hz and P_{out} the outage probability, which is the probability that the received signal-to-noise ratio (SNR) at the destination node falls below the required threshold $2^{\mathcal{R}} - 1$ for correct data detection. According to Fig. 1, the maximum throughput of TS and that of PS are equal to $(1-\theta)(1-P_{\text{out}})\mathcal{R}/2$ and $(1-P_{\text{out}})\mathcal{R}/2$, respectively [7]. The factor $(1-\theta)$ for $0 < \theta < 1$ that appeared in the throughput expression of TS suggests that TS performs worse than PS at high SNR (where P_{out} is eligible). Further performance comparisons of separated receivers with other receivers will be given at the end of this section.

Integrated

An integrated receiver architecture is proposed in [6] by combining information decoding and EH circuits such that a part of the information decoding functionality is taken care of by the rectifier. In this case, information must be modulated by different energy levels along with energy detection in the decoder. Such an integrated receiver is shown to greatly reduce the energy cost for information decoding, particularly when the circuit power consumption is relatively large.

Unidirectional

The above three receiver architectures focus on enabling simultaneous wireless information and power transfer (SWIPT) without specifically taking into account energy storage at relays. Hence, the harvested energy is consumed immediately within a transmission cycle. This often makes the transmission power of the relay node fairly limited due to severe propagation loss and low energy conversion efficiency of WPT. A feasible solution is to utilize the energy storage (e.g. a rechargeable battery or a super capacitor) such that the harvested energy can be accumulated to boost the transmission power level at relays. With the aid of energy storage, we propose a simple receiver architecture, referred to as the *unidirectional* receiver, where the energy flow in the relay either enters or leaves the energy storage without splitting in time or power. To enable SWIPT, a single relay is chosen to receive the source information, while at the same time the rest of the relays perform energy harvesting to eliminate time-domain and power-domain multiplexing used in TS and PS, respectively. As illustrated in Fig. 1c, in the first half of the transmission cycle, the source transmits information. In the meantime, the selected relay R^* receives the source information, and those nonselected relays harvest energy from the source signal. In the second half of the transmission cycle, R^* forwards the source information and the remaining relays may perform EH or enter the sleep mode to save energy. We defer the discussion of relay selection to the next section and compare different receiver architectures using numerical examples to conclude this section.

Performance Comparison

We evaluate the performance of the aforementioned three receivers, including TS, PS, and unidirectional. The ideal receiver and the integrated receiver are not considered mainly because of their practicality. Since relays are commonly

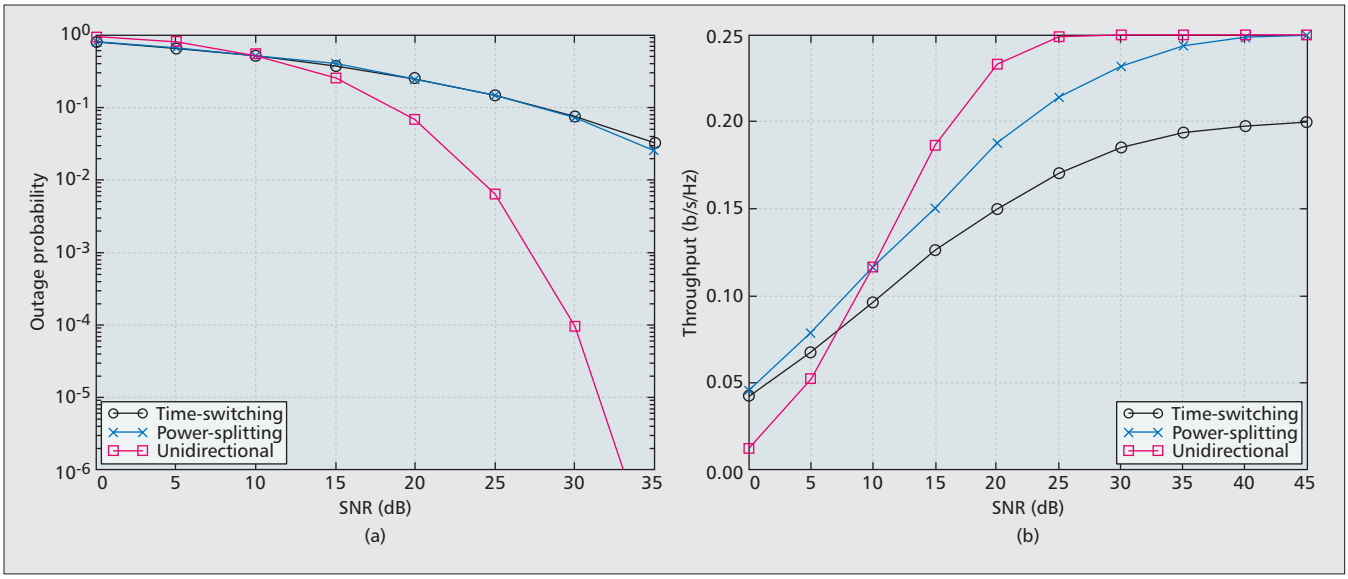


Figure 2. Performance comparisons of different receiver architectures: a) outage probability; and b) throughput.

deployed to explore spatial diversity, our evaluation takes into account the randomness of relay locations by distributing them according to a two-dimensional Poisson point process (PPP) with density λ in an area A of radius l_A. Within this area, the source and the destination nodes are placed such that the halfway between them coincides with the center of A. In simulations, we set l_A = 10 m, and the distance between the source and the destination is 7 m. This setting resembles low-power sensor networks where the transmission range of sensor nodes is small due to low transmission power. All the channels between two arbitrary nodes experience both large-scale fading due to distance with path-loss exponent α as well as small-scale Rayleigh fading with unit power. Unless specified, a constant transmit power P is employed for both the source and the relay nodes. At each receiver, the noise power N_0 is assumed to be unity. Thus the ratio P/N_0 is referred to as the system SNR. In addition, each relay is equipped with a finite battery with size B, which scales with P, i.e. $B = \beta P$ where $\beta > 0$ is referred to as the battery scaling factor. The performance metrics of interest are outage probability and throughput.

In Fig. 2 we set relay density $\lambda = 0.2$, path-loss exponent $\alpha = 4$, battery scaling factor $\beta = 2$, energy conversion efficiency $\kappa = 0.6$, and target rate $\mathcal{R} = 0.5$ bits/sec/Hz. For TS, the time switching ratio $\theta = 0.2$. For PS, the power splitting ratio $\rho = 0.6$. These values are chosen according to [7]. One can see from Fig. 2a that the outage probability curve of the unidirectional receiver sharply decreases with SNR, and the decaying rate is much higher than that of TS and PS receivers. This is because with unidirectional receivers, relays can fully (in both time and power domains) replenish energy when they are not designated as the cooperating relay, in contrast to the other two receivers. The throughput performance shown in Fig. 2a further indicates the superiority of the unidirectional receiver, which achieves the maximum throughput at SNR of 25 dB. In comparison, the TS and PS receivers incur about 20 dB SNR penalty in achieving the maximum throughput.

Relay Selection (RS)

Next we consider a multi-relay setup, where multiple EH relays are installed in the network to assist a single source-destination pair. Such a setup can be found in remote sensing applications where sensor nodes report their local sensing data to a data sink via intermediate relays. In principle, employing a single selected relay to forward the source information can not only fully exploit diversity gain but also minimize the multiplexing loss, known as the diversity-multiplexing tradeoff (DMT) [8]. Several diversity-achieving RS protocols have been proposed for DF and AF relays. In the following, we limit our discussion to AF relays and refer interested readers to [9] for DF relays.

Distance-Based: For some scenarios such as in the outdoors, the channel quality is dominated by the distance between the sender and the receiver rather than by multi-path fading. It is thus intuitive to select the relay that is closest to the source node [10].

Channel State Information (CSI)-Based: For conventional relays, CSI-based RS has been proven to be optimal in terms of DMT. The best relay in CSI-based RS is defined as the one with the superior end-to-end channel condition because this relay offers the highest achievable rate among the others. Some early work on EH relays also adopt CSI-based RS [11, 12].

Battery-Aware: The relay selected according to CSI only may not have enough power to transmit. This may result in severe performance loss of EH relays because the source information cannot be forwarded via the selected relay. A simple strategy to overcome this problem is to modify the RS rule as follows. Define forwarding set as the set of EH relays with sufficient energy to transmit. In the forwarding set, the relay that has the best end-to-end SNR is then chosen as the best relay.

This relay selection rule, referred to as bat-

tery-aware (BA) RS, can be implemented by requesting each relay to examine its battery status at the RS epoch. If the remaining battery power of a particular relay exceeds the prescribed transmission power, this relay declares itself as a candidate. Otherwise, it simply enters harvesting mode without participating in cooperative retransmission.

Performance Evaluation

We evaluate the performance of BA relay selection considering two different implementations of AF relays: variable-gain AF (VAF) and fixed-gain AF (FAF). For VAF relays, the relay adapts its amplification gain to the instantaneous channel condition between the source and itself. As to FAF, a constant amplification gain is employed, regardless of the channel condition. Comparing these two variants of AF relays, VAF relies on continuously monitoring the source-relay channel condition to determine the amplification gain and thus incurs a higher complexity than FAF, which does not require any CSI. For this reason, FAF is also known as blind AF.

In determining the transmission power, VAF relays vary their amplification gains to maintain a constant transmission power, which is set equal to the source transmission power P. As to FAF relays, their transmission powers are identical to the constant amplification power gain denoted as g.

Figure 3 depicts the outage probabilities of the three RS rules mentioned above for VAF relays using the unidirectional architecture under relay density $\lambda = 0.5$, pathloss exponent $\alpha = 4$, energy conversion efficiency $\kappa = 0.5$, and battery scaling factor $\beta = 10$. All the curves shown in the figure descend with SNR, except CSI-based RS, which shows an error floor at high SNR. As explained before, the relay selected according to CSI-based RS will be idle if its battery power is less than the required transmission power. This idle probability is constant to SNR and will dominate the outage probability when the batter size is not sufficiently large. The BA scheme significantly outperforms the CSI-based scheme and the distance-based scheme, which achieves diversity gain of one only.

The case of FAF relays is shown in Fig. 4 for $g = 10$. Again, the BA scheme is superior to the distance-based and the CSI-based schemes. Different from Fig. 3, here the outage probability of the CSI-based scheme decreases with SNR. Recall that the relay transmission power of FAF relays is set equal to constant g, which is relatively small compared to P at high SNR. As a result, the EH rate is higher than the consumption rate such that the idle probability becomes negligible as SNR increases.

Figure 5 demonstrates the outage probability of the BA scheme under different values of pathloss exponent α for $\beta = 5$ and SNR = 10 dB. Here we focus on FAF relays with $g = 10$ since the case of VAF relays reveals the similar trend and thus is omitted. Intuitively, a larger pathloss exponent implies more severe signal attenuation due to distance and thus the outage probability decreases with α. It is worth noting that the linear curve shown in the figure plotted with logarithmic scale corresponds to exponential decay of the outage probability with λ. This reveals that EH relays can still exploit diversity gain provided by spatially random relays even without permanent power supplies.

In Fig. 6 we investigate the impact of battery size on the outage probability by manipulating the battery scaling factor β for both VAF and FAF relays considering CSI-based and BA RS schemes. Here we set λ=0.2 and SNR = 15 dB. As expected, increasing β helps reduce the outage probability, but the performance improvement provided by a larger battery is rather limited. As shown in the figure, the outage performance is not sensitive to β when β>5. Overall, maintaining high relay density and operating them in high SNR regions are more important

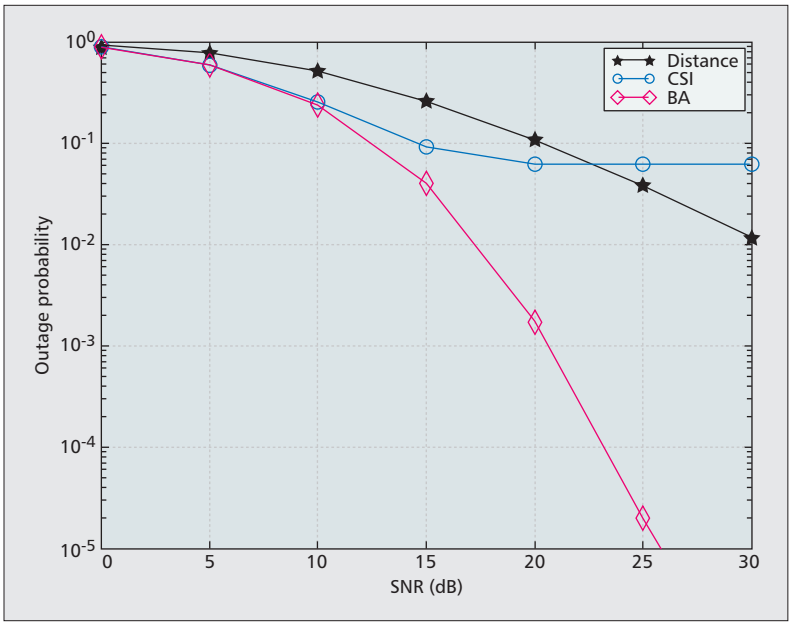


Figure 3. Comparison of relay selection schemes based on VAF relays for the battery scaling factor $\beta = 10$.

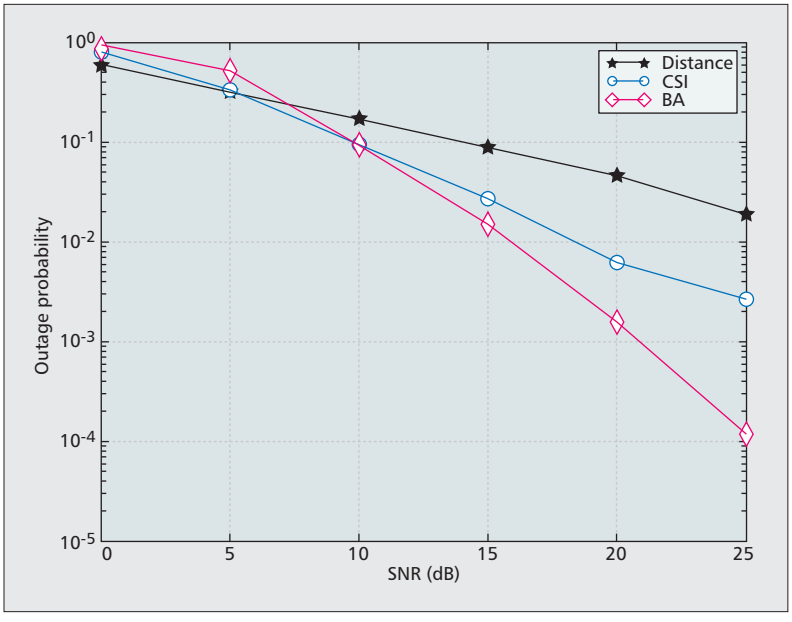


Figure 4. Comparison of relay selection schemes based on FAF relays for the battery scaling factor $\beta = 10$ and the fixed amplification gain $g = 10$.

to the performance of cooperative relaying based on EH relays than increasing the battery size.

Multi-User Scenario

So far our discussions have been concentrated on the single source-destination setting, the most straightforward application of cooperative relaying. Relays can also be deployed to assist multiple source nodes communicating to common or distinct destination nodes. Such a multi-user scenario requires relays to distribute their limited resources (e.g. transmission time and power) in order to maximize the overall welfare. In particular, transmission power is deemed as the most treasured resource for EH relays. Even with an equal power allocation, it has been shown that the outage probability decays at a rate of log SNR/SNR, which outperforms the decay rate 1/SNR without any power allocation performed [10]. Further improvement is possible, for example, by using a water-filling power allocation along with user scheduling. However, this strategy requires global CSI for relays to determine their transmission power and scheduling priority. To alleviate signal overhead, decentralized mechanisms should be desired and deserve further investigations. One representative work along this line is [10], which develops a distributed power allocation strategy based on game theory. This work focuses on a single-relay scenario while in the more practical multi-relay scenario, other design dimensions may be further explored. For instance, relay selection can be jointly designed with power allocation and scheduling in order to maximize cooperative diversity and multi-user diversity gains, subject to energy availability imposed by energy-constrained relays.

Challenges for Future Research

Diversity-Achieving Protocols

Early works have largely focused on exploring the potential gain of EH relays based on some ideal assumptions. One of the fundamental requirements to exploiting diversity benefit is to provide interference-free transmission. Otherwise, cooperative relaying may encounter a complete loss of diversity gain due to co-channel interference (CCI). Conventionally, orthogonal transmission is achieved via specially designed waveforms or exclusive resource allocations. As the emerging wireless networks are designed to meet high spectral efficiency, it becomes more challenging to achieve orthogonal transmission. Despite its adverse impact on achievable diversity performance, the presence of CCI may be regarded as an alternative energy source to EH relays. Since the wireless channel condition varies with time, a relay that suffers severe CCI during a certain period may be scheduled to perform EH, and then perform cooperative relaying when the CCI is diminished. The time-varying nature of wireless channels suggests that the relaying protocol should be jointly designed with interference management schemes such that CCI can be opportunistically utilized to charge the relay battery and avoid its catastrophic effect on cooperative relaying. In this case, a certain degree of CSI is necessary to instruct the relay operation to switch between EH and information relaying. The more power reserved for the acquisition of CSI, the better decision making for relay operation but the less energy left for information relaying. How to balance this rate-energy trade-off presents a new challenge to the design of diversity-achieving protocols using EH relays.

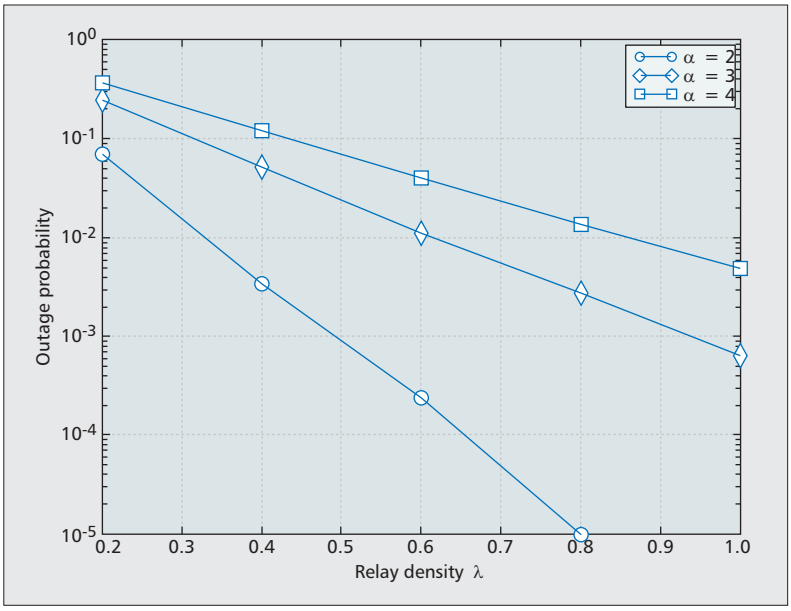


Figure 5. Outage probability of BA-FAF versus relay density λ.

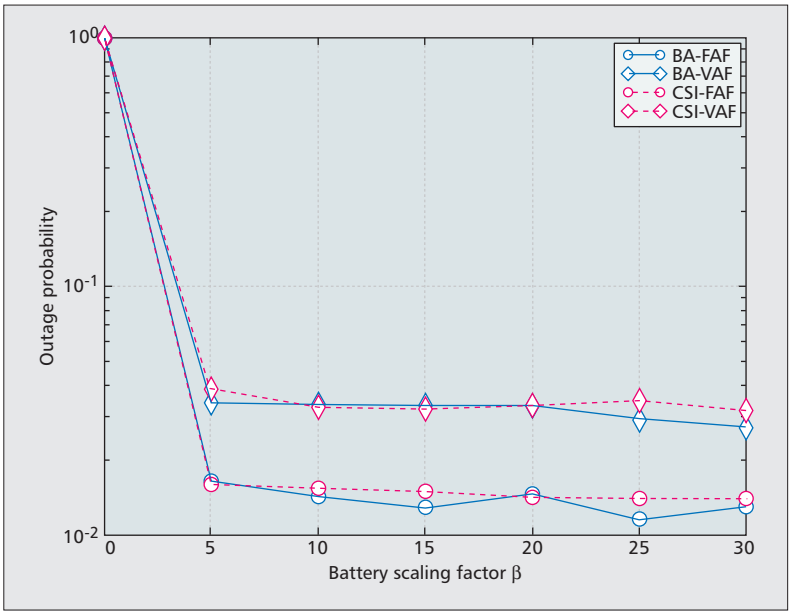


Figure 6. Impact of battery scaling factor β to the outage probability.

Multi-Antenna Systems

Advanced multi-antenna technology promises new design dimensions to improve cooperative transmissions based on EH relays. In conventional wireless networks, multiple antennas are often used to boost the information rate or mitigate CCI. For EH wireless networks, multiple transmit antennas can be employed to achieve highly efficient energy harvesting via so called

While existing results on EH devices have confirmed their promising gains, we show that simply reusing existing approaches may not satisfy the ultimate goal of self-sustainable relays. We thus point out future research directions that we believe are of vital importance to the success of self-sustainable relays.

energy beamforming [13]. In this context, the time resource needs to be properly divided for acquiring CSI and performing energy harvesting. To simplify the implementation, random beamforming is considered in [14] to enhance the rate-energy trade-off for EH nodes based on the TS receiver architecture. Without assuming any CSI at the transmitter, it is proven that using one single random beam can achieve the optimal rate-energy trade-off asymptotically. However, transmit beamforming has not been explored for EH relays and should be further investigated, considering that both the channel conditions and the available power at relays are relevant to the beam design.

Full-Duplex Relays

All the receiver types introduced in this article are half-duplex in nature, yet they mimic the full-duplex operation via different multiplexing approaches. Lately, full-duplex relays have attracted more research attention as full-duplex radios can transmit and receive at the same time without self interference, which doubles the transmission rate. However, using full-duplex radios does not imply that information transmission and EH can be performed simultaneously. Full-duplex relaying is treated in [15], considering a two-antenna configuration to enable full-duplex based on the TS receiver architecture. New enabling methods, including the receiver architecture, cooperative relaying protocols, and resource allocation strategies remain open and are of timely importance.

Concluding Remarks

Abundant research on cooperative relaying has proven its encouraging benefits to future wireless communication systems. Existing designs emphasize achieving high diversity gain but incur extra energy consumption that goes against the trend of green communications. The potential of self-sustainable relays discussed in this article opens a new paradigm toward green wireless communications system by using EH relays in place of conventional relays that rely on fixed power supplies. While existing results on EH devices have confirmed their promising gains, we show that simply reusing existing approaches may not satisfy the ultimate goal of self-sustainable relays. We thus point out future research directions that we believe are of vital importance to the success of self-sustainable relays.

Acknowledgement

Kuang-Hao Liu's work was supported partly by the Ministry of Science and Technology of Taiwan under Grant MOST 103-2221-E-006-081, and the Institute for Information Industry, ROC under Grant A3KG1422. Phone Lin's work was supported in part by the Institute for Information Industry R.O.C under Grant A3KG1422, by the Ministry of Science and Technology of Taiwan under Grants MOST 103-2221-E-002-152-MY3, MOST 103-2221-E-002-249-MY3, MOST 104-2923-E-002-005-MY3, 103-2622-E-009 -012, 103-2218-E-002-032, 103-2627-E-002-008, by Chunghwa Telecom, by ICL/ITRI, and by MoE ATU Plan.

References

[1] D. Gunduz *et al.*, "Designing Intelligent Energy Harvesting Communication Systems," *IEEE Commun. Mag.*, vol. 52, no. 1, Jan. 2014, pp. 210–16.
[2] S. Sudevalayam and P. Kulkarni, "Energy Harvesting Sensor Nodes: Survey and Implications," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 3, 2011, pp. 443–61.
[3] L. X. Cai *et al.*, "Dimensioning Network Deployment and Resource Management in Green Mesh Networks," *IEEE Wireless Commun.*, vol. 18, no. 5, Oct. 2011, pp. 58–65.
[4] T. Le, K. Mayaram, and T. Fiez, "Efficient Far-Field Radio Frequency Energy Harvesting for Passively Powered Sensor Networks," *IEEE J. Solid-State Circuits*, vol. 43, no. 5, May 2008, pp. 1287–302.
[5] P. Grover and A. Sahai, "Shannon Meets Tesla: Wireless Information and Power Transfer," *Proc. IEEE ISIT*, Austin, TX, June 13–18 2010.
[6] X. Zhou, R. Zhang, and C. K. Ho, "Wireless Information and Power Transfer: Architecture Design and Rate-Energy Tradeoff," *IEEE Trans. Commun.*, vol. 61, no. 11, Nov. 2013, pp. 4754–67.
[7] A. A. Nasir *et al.*, "Relaying Protocols for Wireless Energy Harvesting and Information Processing," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, July 2013, pp. 1536–1276.
[8] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge, U.K: Cambridge University Press, 2005.
[9] K.-H. Liu, "On the Performance of Time-Orthogonal Incremental Relaying based on Demodulate-and-Forward with Distributed Channel Access," *IEEE Trans. Vehic. Commun.*, vol. 61, no. 2, Feb. 2012, pp. 737–47.
[10] Z. Ding *et al.*, "Power Allocation Strategies in Energy Harvesting Wireless Cooperative Networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, Feb. 2014, pp. 846–60.
[11] I. Ahmad, *et al.*, "Joint Power Allocation and Relay Selection in Energy Harvesting AF Relay Systems," *IEEE Commun. Lett.*, vol. 2, no. 2, Apr. 2013, pp. 239–42.
[12] Z. Ding *et al.*, "Wireless Information and Power Transfer in Cooperative Networks with Spatially Random Relays," *IEEE Trans. Wireless Commun.*, to appear.
[13] G. Yang, C. K. Ho, and Y. L. Guan, "Dynamic Resource Allocation for Multiple-Antenna Wireless Power Transfer," *IEEE Trans. Signal Proc.*, vol. 62, no. 14, June 2014, pp. 3565–77.
[14] H. Ju and R. Zhang, "A Novel Mode Switching Scheme Utilizing Random Beamforming for Opportunistic Energy Harvesting," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, Apr. 2014, pp. 2150–62.
[15] C. Zhong *et al.*, "Wireless Information and Power Transfer with Full Duplex Relaying," *IEEE Trans. Commun.*, vol. 62, no. 10, Oct. 2014, pp. 3447–61.

Biographies

Kuang-Hao (Stanley) Liu [S'06, M'08] (khliu@mail.ncku.edu.tw) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Canada, in 2008. He is currently an associated professor with the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan. His recent research focuses on green wireless communications, dense small-cell networks, and spectrally-efficient multiple access.

Phone Lin (plin@csie.ntu.edu.tw) is a professor at National Taiwan University, holding a professorship in the Department of CSIE, Graduate Institute of Networking and Multimedia, Telecommunications Research Center, and the Optoelectronic Biomedicine Center. Lin is an IEEE Senior Member and ACM Senior Member. He received his BSCSIE and Ph.D. degrees from National Chiao Tung University Taiwan in 1996 and 2001, respectively. His major research is area is mobile communications networking.

Wireless-Powered Cellular Networks: Key Challenges and Solution Techniques

Hina Tabassum, Ekram Hossain, Adedayo Ogundipe, and Dong In Kim

ABSTRACT

Energy harvesting from ambient sources is emerging as a sustainable and environment-friendly technique to prolong the lifetime of wireless devices. However, harvesting energy from these sources may not be feasible for quality-of-service (QoS)-constrained wireless applications. As such, dedicated wireless-powered cellular networks (WPCNs) are currently being investigated to ensure the reliability as well as improved battery lifetime of the wireless devices. With emerging WPCNs, a true wireless network can be envisioned, which is free of connectors, cables, and battery access panels, and guarantees freedom of mobility. To illustrate and understand the design requirements of WPCNs, this article first points out the key challenges of designing energy-harvesting cellular networks. These include the doubly near-far problem, the signal-to-noise ratio (SNR) outage experienced by the energy-harvesting devices located toward the cell-edge, the spatial characterization of the SNR outage zone, the additional resource consumption at energy transmitting sources, and the problems related to designing fairness-constrained user scheduling schemes. A brief overview of the related research advancements in WPCNs and a summary of their limitations are then provided. Finally, we list a few suggestions and design considerations that can potentially tackle the major challenges in emerging WPCNs and open up new research directions.

Hina Tabassum, Ekram Hossain, and Adedayo Ogundipe are with the University of Manitoba.

Dong In Kim is with Sungkyunkwan University (SKKU).

This work was supported in part by a Collaborative Research and Development Grant (CRD PJ 461412 - 13) 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

Energy harvesting in wireless cellular networks is a cornerstone of emerging 5G cellular networks as it aims to "cut the last wires" of the existing wireless devices [1]. In particular, energy harvesting has significant potential to attract subscribers since it promotes mobility and connectivity *anywhere* and *anytime*, which is one of the key visions of emerging 5G networks.

Until very recently, energy harvesting for wireless communications systems mainly considered ambient energy sources (e.g., solar, motion and vibration, temperature, wind, thermoelectric effects, interference from radio frequency (RF) sources, etc.). As such, the efficiency of energy-harvesting systems exploiting those sources depends on available energy levels that may vary significantly over time, location, and weather conditions. Nevertheless, to satisfy the power demands of delay-constrained wireless applications, it is of utmost importance to ensure the availability of sufficient energy at the user terminals whenever required. This fact has motivated researchers toward the development of dedicated wireless-powered cellular networks (WPCNs) where a hybrid access point (HAP) takes care of both the energy and information transmission to/from the subscribers. Consequently, new research challenges and activities have surfaced on a global scale and on a broad range of topics. These topics include prototyping receiver architectures [2, 3], optimizing time allocation for wireless energy and information transfer considering single antenna or multi-antenna HAPs [4, 5], wireless energy transfer in heterogeneous smallcell networks [6], relay-assisted cooperative protocol designs [7, 8], etc. Besides, a few other dedicated techniques (such as mobile charging vehicles [9], dedicated power-beacons (PBs) [10]) are also under investigation. PBs are low-cost devices that can potentially charge wireless terminals by transferring energy in a directional or omni-directional manner. WPCNs have thus initiated a paradigm shift in designing wireless networks where the energy can be scheduled to users on their request, resulting in perpetual network connectivity.

In this article we highlight the primary challenges in the WPCNs mainly from the following perspectives: resource allocation, SNR outage performance of a user who is located arbitrarily in the signal-to-noise-ratio (SNR) outage zone, and the significance of the deployment of dedicated energy sources. In this context,

- We first characterize the SNR outage zones in the conventional cellular networks and in WPCNs theoretically as well as numerically. It is shown that the SNR outage zone of a WPCN is wider than that of a conventional cellular network.
- We then analyze the gains of deploying the PBs and distributed antenna element (DE)-assisted PBs over a centralized HAP in a WPCN. The performance measure is the SNR outage probability of a user in the SNR outage zone.
- Quantitative and comparative analysis of the symmetric and random deployment of

0163-6804/15/$25.00 © 2015 IEEE

Theoretically, RF signals can carry both energy and information simultaneously. Therefore, an ideal energy harvesting device could be one in which energy reception and information decoding can be performed concurrently from the same RF signal.

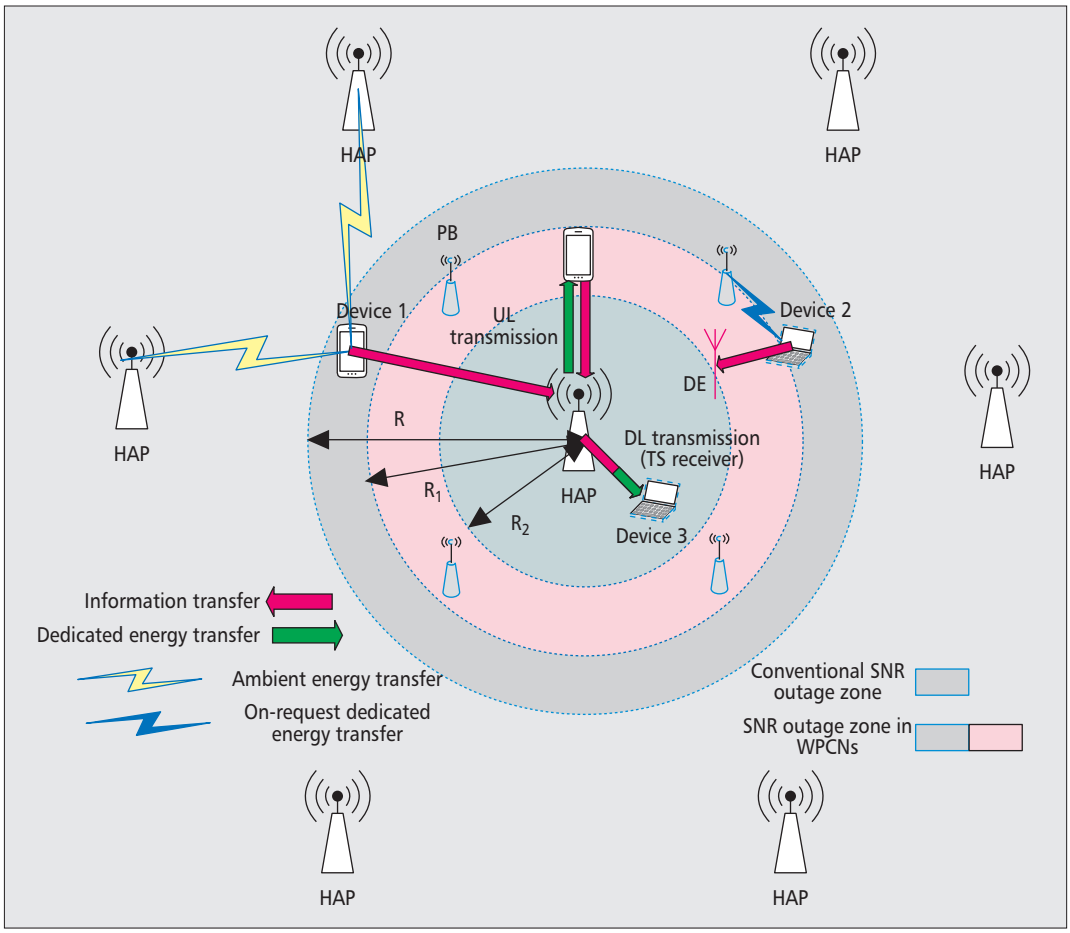


Figure 1. Graphical illustration of WPCN and characterization of the SNR outage zones in the uplink transmissions of conventional cellular networks and WPCNs. A demonstration of selecting the appropriate combination of dedicated (e.g., HAP, DEs, etc.), on-request dedicated (e.g., PBs), and ambient (e.g., interfering transmissions from nearby HAPs) RF energy sources for subscribers located geographically apart from each other.

the PBs and DE-assisted PBs is then provided.

A qualitative overview of recent developments and common assumptions in the design of WPCNs are then presented. Finally, we point out future research directions that can potentially tackle the primary challenges in a WPCN.

A graphical illustration of the emerging WPCNs is shown in Fig. 1, where network connectivity of all users is ensured by combining the capabilities of dedicated (e.g., HAPs, DEs), on-request dedicated[1] (e.g., PBs) and ambient RF energy sources (e.g., interfering transmissions from nearby HAPs). DEs are spatially separated antenna nodes that are connected to HAPs via a dedicated backhaul link (typically optical fiber).

Key Challenges in Designing WPCNs

In this section we discuss the fundamental challenges in designing WPCNs, e.g., design of receiver architecture to deal with the simultaneous information and power transfer from HAPs in the downlink, SNR outage regions in the uplink due to doubly near-far problem, additional resource consumption at the HAPs that includes energy transfer time and transmission power to charge wireless devices, etc.

Simultaneous Information and Power Transfer (SWIPT)-Enabled Receiver Architecture

Theoretically, RF signals can carry both energy and information simultaneously. Therefore, an ideal energy harvesting device could be one in which energy reception and information decoding can be performed concurrently from the same RF signal. This is commonly referred to as simultaneous wireless information and power transfer (SWIPT). Unfortunately, the existing receiver circuits in wireless devices are not capable of directly extracting energy and information from the same RF signal. As a result, receiver architecture design is a challenge for concurrent wireless energy transfer (WET) and wireless information transfer (WIT) in the downlink of WPCNs. One root cause of this problem is the considerably different power sensitivities of receivers, i.e., –10 dBm for energy harvesting and –60 dBm for information decoding. Thus, efficient receiver architecture and operating policies that can ensure quality-of-service (QoS) and quality of wireless charging service (QoCS) at the subscribers while handling

[1] On-request dedicated energy sources are supplementary devices that can provide energy to subscribers on request.

simultaneous energy and information transfer are of significant importance.

Doubly Near-Far Problem in Energy-Harvesting Cellular Networks

Because of distance-dependent signal attenuation in both the uplink and downlink, any user located closer to a HAP harvests more energy in the downlink and also requires less power to attain a given SNR at the HAP in the uplink (given a fixed transmission and energy harvesting time slot). On the other hand, distant users harvest low energy in the downlink but require higher transmission power to achieve the same SNR in the uplink. This problem is typically referred to as the "doubly near-far problem" in the literature [4]. This phenomenon may significantly reduce fairness among different users that are located spatially apart (e.g., cell-center and cell-edge users) from the HAP. However, this is an interference-free perspective of the doubly near-far problem that does not consider the interference experienced by cell-edge users. In practice, a macrocell is surrounded by several tiers of neighboring macrocells. Therefore, cell-edge users are typically exposed to higher interference levels and may accumulate higher energy compared to cell-center users. It is therefore important to study the impact of ambient RF interferers (energy transmitters) on the average accumulated energy of the cell-edge users in a large-scale network setting where both dedicated and ambient RF energy sources coexist.

Additional Resource Consumption at HAPs

HAPs are responsible for energy transfer in order to support both uplink and downlink transmissions. With this obligation, a HAP is typically assumed to allocate a specific portion of time, power, channels, or antennas for energy transfer. Note that the simultaneous use of a channel (or resource) for optimized downlink WIT in addition to WET for uplink transmission has not been considered in recent studies [4, 5]. This additional resource consumption at the HAP is particularly significant for uplink transmission scenarios. In conventional uplink cellular networks, the maximum power consumption of a wireless device is relatively low compared to that of a base station (BS). In contrast, WPCNs consume extra time/power resources of the HAP, especially for the uplink transmission of far-away users. This fact triggers an asymmetric additional power consumption that is not the case in traditional cellular networks. We refer to this phenomenon as *asymmetric additional power consumption in uplink*.

To perform WET in a resource-efficient manner, efficient channel or antenna, power, and time adaptation policies need to be exploited in WPCNs. Adaptive spectrum allocation will be required to optimize the channel allocations for both WET and WIT, taking into account the locations of energy harvesting devices. For instance, some channels for downlink WIT may be concurrently used for WET to charge the users for their uplink transmissions (i.e., satisfying the QoS of a downlink user and the QoCS of an uplink user).

Broadband Energy Harvesting

As has been mentioned above, the doubly near-far problem deteriorates the uplink transmissions in WPCNs, especially for devices located in the cell-edge areas. Although cell-edge users are far away from their dedicated HAP, they may be able to receive energy from RF transmissions of neighboring HAPs, small cells, radio or TV broadcast towers, etc. Thus, they may potentially exploit the benefits of their proximity to such ambient RF sources to satisfy their energy requirements. To implement this in a practical network, new receiver architectures need to be introduced with circuits that can harvest and accumulate energy from signals with a wide range of frequencies. This feature is referred to as "out-of-band" energy harvesting. Although ambient RF sources are unreliable, they may potentially support starving users in the cell-edge areas.

Multiuser Scheduling in WPCNs

Conventional uplink multiuser scheduling schemes (e.g., greedy/opportunistic scheduling, round-robin scheduling) do not consider the amount of harvested energy and/or energy requirements of the selected users, and therefore, when used in WPCNs, they can lead to

- *Energy (or transmit power) outages*, in which the harvested energy of a scheduled user drops below the minimum energy (or power) required for transmission, and
- *Energy overflows*, in which the harvested energy of a scheduled user exceeds its finite battery level (and hence energy is wasted at the HAP for wireless charging).

Energy outage and energy overflow events mainly occur, respectively, when cell-center and cell-edge users are scheduled for transmission. Note that traditional uplink scheduling methods such as the greedy scheduling method generally considers uplink channel state information (CSI) only while round-robin scheduling chooses any user with equal probability regardless of their channel conditions. Therefore, energy outage events may significantly affect fairness-constrained scheduling schemes (e.g., round-robin scheduling scheme) due to higher chances of scheduling of the cell-edge users when compared to opportunistic (i.e., channel state-aware) scheduling schemes. New scheduling methods will therefore be required that consider these events and optimize performance metrics such as the spectral efficiency of uplink transmission and spectral efficiency per unit of charging power.

In the following, we will provide a brief overview of existing solution techniques that address some of the aforementioned challenges.

To perform WET in a resource-efficient manner, efficient channel or antenna, power, and time adaptation policies need to be exploited in WPCNs. Adaptive spectrum allocation will be required to optimize the channel allocations for both WET and WIT, taking into account the locations of energy harvesting devices.

Overview of the Existing Designs of WPCN

Receiver Operating Strategies to Enable SWIPT

To resolve the challenges encountered in designing circuits that perform both energy harvesting and information decoding simultaneously, differ-

Compared to the TS mode, the PS mode achieves higher information rate and harvested energy level. However, its associated cost and hardware complexity make it less attractive to system designers. Specifically, the PS mode requires a power splitter, whereas the TS mode requires a simple switch.

ent receiver architectures have been proposed recently in the literature [2, 3] as described below.

Time Switching (TS) Receiver [2]: This architecture involves an antenna periodically switching between information decoding and energy harvesting circuits. Each transmission block is divided into two orthogonal time slots, one for harvesting energy and the other for transmitting data.

Power Splitting (PS) Receiver [2]: This architecture incorporates a power splitting device that separates the received signal into two streams for the information and energy receivers with the same/varying power levels. An optimal power splitting ratio is designed in [11]. The PS mode has two special cases:

- Uniform Power Splitting (UPS): All antennas at the receiver have same power splitting ratio.
- Dynamic Power Splitting (DPS) [3]: It splits the received signal into two streams with adjustable power levels for WET and WIT. The power allocation for WET and WIT adapts according to CSI which is assumed to be known at the receiver.

Antenna Switching (AS): For multiple antenna devices, the receiving antennas are divided into two groups, one group for information decoding and the other group for energy harvesting. AS reduces to TS for a single antenna device.

Compared to the TS mode, the PS mode achieves higher information rate and harvested energy level [2, 11]. However, its associated cost and hardware complexity make it less attractive to system designers. Specifically, the PS mode requires a power splitter, whereas the TS mode requires a simple switch. Because of this, the PS mode allows only in-band WET, whereas the TS mode can also support out-of-band WET. The TS mode can accommodate both the sensitivity difference of energy/information receivers and the channel/ interference power dynamics to optimize its switching operation. For instance, the receiver can be switched to harvest energy when the channel (or interference) is strong, or decode information when the channel (or interference) is relatively weaker [11]. This adaptability of the TS mode may potentially assist in overcoming the doubly near-far problem (e.g., cell-edge users may switch to harvest mode when they experience strong interference from nearby HAPs).

Existing Techniques to Mitigate the Doubly Near-Far Problem

The major techniques to mitigate the doubly near-far problem in cellular networks are described below.

Common Throughput Maximization in the Harvest-then-Transmit (HTT) Protocol [4]: This algorithm proposes to adapt and optimize the time allocation for WIT according to the channel states of different users such that the common throughput (equal throughput of all users) is maximized. In particular, this is achieved by assigning shorter and longer time periods to the near and far-away users, respectively. This is similar to the concept of traditional SNR balancing of users in cellular networks.

User Cooperation with the Harvest-then-Cooperate (HTC) Protocol ([12, 14] and references therein): A user cooperative protocol is proposed by [12] to mitigate the doubly near-far problem. A two-user system is considered in which a user with better uplink and downlink channel gains assists another user with poor channel gains in transmitting information to the HAP. In particular, the HAP broadcasts RF signals for WET in the downlink and both the users harvest energy during this time. The near user then uses part of its allocated transmission time and harvested energy to relay the information of the far-away user, and the remaining time and energy to transmit its own information. Reference [13] derives the outage probability expressions for users in a cooperative network with an energy-harvesting relay. The cooperation among users has been modeled as a canonical coalition game and the grand coalition has been shown to be stable.

Relay Cooperation with the HTC Protocol ([7, 8] and references therein): Recently, relay-based cooperation has been shown to be useful in mitigating the doubly near-far problem of a WPCN consisting of a HAP, a source node, and a relay node. For instance, [8] considers an energy-constrained relay node that harvests energy broadcast by a source node and uses this energy to forward the source signal to a destination node. In this regard, two relaying protocols, i.e., the time switching-based relaying protocol and the power splitting-based relaying protocol, are proposed to separate information processing and energy harvesting at the relay node. Similarly, in [7] the source and relay nodes harvest energy from the HAP in the downlink and work cooperatively in the uplink to transmit the source's information using a TDMA protocol. The throughput of the proposed protocol is also analyzed in a multi-relay scenario with opportunistic and partial relay selection schemes.

Multi-Antenna Transmission with Energy Beamforming ([14] and references therein): In contrast to the aforementioned TDMA-based techniques, [14] presents an approach that considers the use of energy beamforming at multi-antenna HAPs to mitigate the doubly near-far problem. Energy beamforming is a technique that allows multiple antennas to transmit RF energy in the direction of an intended energy harvester for better transmission gains. In particular, the HAP first performs WET to all users via energy beamforming in the downlink, and then the users perform WIT based on space division multiple access (SDMA) simultaneously in the uplink using their harvested energy. The problems of time allocation in the uplink and downlink, energy beamforming, and uplink transmit power allocation with receive beamforming are investigated.

Compared to the HTT protocol, the HTC protocol has been shown to outperform in all considered scenarios at the cost of additional system complexity and overheads (e.g., time synchronization, CSI of cascaded channels with shadowing and fading, etc.). The performance gains can be enhanced further with the use of energy beamforming at multi-antenna HAPs.

	System model	CSI	Energy storage	Benefits	Limitations	Objective/ constraint	Co-channel interference
HTC [7]	Three node WPCN (HAP, source, relay), TDMA, WET in DL, WIT in UL	Perfect UL and DL CSI at relay and HAP	Unlimited storage at source and relay; harvest-use method[1]	Overcome doubly near-far problem	Synchronization complexity; not useful if source is in energy outage	Optimized rate for different relay selection schemes with partial (one hop) and complete CSI (two hops)	no
HTC [12]	Two user WPCN with an HAP, TDMA, WET in DL, WIT in UL	Perfect CSI at HAP and at relaying users	Harvest-use method; unlimited storage at users	Overcome doubly near-far problem	Synchronization complexity; perfect CSI of other users at relaying user	Jointly optimize time and power to compute users' weighted sum-rate with total time constraint	no
SWIPT [11]	Point-to-point (SISO), TS mode, DL WET and WIT	Known CSI of channel and interference at RX; both known and unknown CSI at TX	Harvest-use method; unlimited storage at RX	Interference-based harvesting; optimal mode switching rule	Joint WET and WIT not possible; low performance compared to PS mode	Maximize capacity of WIT and harvested energy in WET	yes
SWIPT [2]	Three-node WPCN (1 TX and 2 RX), MIMO	Perfect CSI at TX	Harvest-use method; unlimited battery size	TS and PS receiver modes	Joint WET and WIT from the same RF signal not possible	Optimal transmission and trade-off between rate and energy transfer	no
Full-duplex-WPCN [5]	Dual-antenna HAP; multiple single antenna users; simultaneous WET in DL and WIT in UL	Perfect UL and DL CSI at HAP	Harvest-use; unlimited storage at harvesting device	Flexible energy harvesting time with perfect and imperfect SIC	Self-interference; additional resource consumption at HAP	Maximize weighted sum-rate to optimize the time and transmit power allocation with total time; average and peak transmit power constraints at HAP	yes

[1] The network node utilizes all harvested energy for transmission purposes.

Table 1. Qualitative overview of existing energy harvesting protocols in WPCNs.

Full-Duplex Transmission for Enhanced Flexibility in Energy Harvesting Time

Full-duplex operations are typically supported in WPCNs with the use of dual-antenna HAPs [5]. One of those antennas performs downlink WET to users and the other is used for WIT in the uplink. This is different from conventional full-duplex communications systems in which simultaneous information transmission and reception takes place in the same frequency band and at the same antenna. Self-interference is an issue for full-duplex HAPs since part of the downlink energy and uplink information signals interfere with each other. However, several self-interference cancellation (SIC) schemes are currently under investigation [5]. Full-duplex transmission offers considerable time flexibility for energy harvesting compared to half-duplex transmission since a user can continue to accumulate energy during the transmission of other users. It also ensures judicious use of available bandwidth as proposed in [5], where the same frequency is utilized for WET and WIT with the SIC schemes in place.

Most of the aforementioned studies assume simplistic system models, i.e., two or three node-network model, infinite energy storage at receivers, perfect SIC, ignore channel diversity of multiple harvesting devices that may be located at cell-center or cell-edge, assume a co-channel interference-free environment, and a single source for energy harvesting (i.e., HAP), etc. A qualitative overview of some of the aforementioned research works is summarized in Table 1.

The performance and design of existing WPCNs need to be enhanced by considering the following:

- Co-channel interference degrades the communication link quality during WIT. However, it is beneficial from a WET perspective. Thus, the trade-offs involved in mitigating the interference need to be critically characterized to maximize information as well as energy transfer.
- Energy cooperation needs to be exercised where multiple users (or HAPs) can share a portion of their harvested energy (or battery energy) with the other users (or HAPs) via wireless transfer [15].
- Most of the existing works assume that harvesting is the only source of energy. However, both regular and harvesting-based energy sources are highly likely to co-exist in practice.
- To make energy harvesting appealing to all users regardless of their locations, deployment of low-cost dedicated energy sources is crucial [10]. The deployment of such sources needs to be limited and planned in

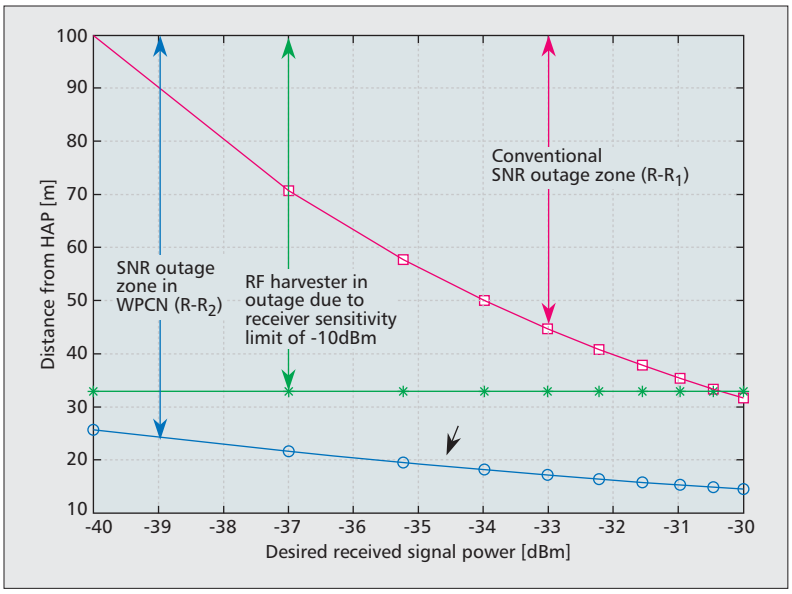


Figure 2. SNR outage zones for both conventional cellular networks and WPCNs as a function of P_0, $\beta = 2$, $P_{max} = 1$ mW, $T/\tau = 0.5$, $P_{HAP}G_t = 40$ W, $\eta = 0.5$, $G_r = 3.98$, $\lambda = 0.328$.

an optimal manner to minimize deployment, operation, and maintenance costs while enhancing the QoCS and QoS of cell-edge users. This could reduce the burden of energy transfer to far-away users from a centralized HAP.

In the following section we provide a characterization and comparison of the SNR outage zones in conventional cellular networks and WPCNs. To improve the performance of users located at cell-edges, several techniques are outlined and quantitatively analyzed.

Design Guidelines to Overcome the Limitations of WPCNs

Characterization of SNR Outage Zones

The root cause of SNR outage of the cell-edge users is the distance between HAPs and harvesting devices, finite battery power, and minimum energy requirements of the harvesting devices for successful transmission and reception of signals.

Conventional Uplink Cellular Networks: In conventional uplink cellular networks, a wireless device adapts its transmit power to achieve a target received signal power P_0 at macrocell BS [16]. This power adaptation compensates the long-term path-loss variations. The distant devices from BSs are required to transmit with higher power in order to combat path-loss, i.e., $P^{tx} = P_0 d^{\beta}$, $P_0 d^{\beta} < P_{max}$, where d denotes the distance of the wireless device, β denotes the path-loss exponent, and P_{max} is the maximum transmit power of the wireless device. For a clear exposition, we do not consider the shadowing and fading effects in the propagation model. Due to the maximum transmit power constraint, all users beyond a certain distance R_1 from the serving macrocell BS will not be able to achieve P_0. This boundary distance R_1 can be calculated by equating $P_{max} = P_0 R_1^{\beta}$ which gives

$$R_1 = \left(\frac{P_{max}}{P_0}\right)^{\frac{1}{\beta}}.$$

We refer to the region beyond R_1 as the SNR outage region for conventional cellular networks.

Wireless-Powered Cellular Networks: To characterize the SNR outage zone in WPCNs, we consider a HAP located in the center of a circular macrocell. The amount of power harvested by an arbitrary user at a distance d from the HAP is given using a simplified form of the Friis Equation provided in the Powercast Wireless Power Calculator ver. 1.5 available at www.powercastco.com. According to the Friis Equation, RXPower $= P_{HAP} G_t G_r (\lambda/4\pi)^{\beta} d^{-\beta}$, where P_{HAP} is the transmit power of the HAP in the downlink, G_t and G_r are the transmit and receive antenna gains, λ is the wavelength of the RF carrier, and d is the distance of the user from the HAP. This is the power received at the energy harvester before performing RF-to-DC conversion. The total harvested energy after conversion can then be given as $E = T\eta$RXPower, where T is the energy harvesting duration in the downlink and η is the harvesting efficiency of the receiver which depends on the efficiency of the antenna, the accuracy of the impedance matching between the antenna and the voltage multiplier, and the power efficiency of the voltage multiplier that converts the received RF signals to DC voltage [17]. However, in most of the research studies, the overall efficiency η of the RF energy harvester is taken in between 50–80 percent [4, 14, 15]. Therefore, we consider $\eta = 50$ percent for our numerical results.

When all users target a received signal quality level of P_0 at the HAP, the two primary conditions, i.e., $E \geq P_0 d^{\beta}\tau$ and $P_0 d^{\beta} \leq P_{max}$, need to be satisfied, where τ is the length of the transmission time slot. The threshold distance R_2 can then be characterized using the first condition as

$$R_2 = \left(\frac{T_\eta P_{HAP} G_t G_r \left(\frac{\lambda}{4\pi}\right)^{\beta}}{\tau P_0}\right)^{\frac{1}{2\beta}}.$$

Note that R_1 denotes the threshold distance beyond which all users experience 100 percent SNR outage in conventional cellular networks. On the other hand, R_2 denotes the boundary beyond which all users experience 100 percent SNR outage in a WPCN. As such, $R - R_1$ and $R - R_2$ are the SNR outage regions of conventional and wireless-powered cellular networks, respectively.

The receiver sensitivity of the RF energy harvester is reported to be –10 dBm in the literature [17, Sec. III.E]. With this value of receiver sensitivity, we calculate a boundary distance from the HAP beyond which no RF energy harvester can sense and harvest energy. This distance d_{th} is derived based on the simplified form of the Friis Equation detailed above. To achieve

an RXPower of at least –10 dBm (before conversion), d_{th} can then be calculated as

$$d_{\text{th}} = \left(\frac{0.0001}{P_{HAP} G_t G_r \left(\frac{\lambda}{4\pi} \right)^{\beta}} \right)^{-\frac{1}{\beta}}.$$

This distance is also demonstrated in Fig. 2.

Figure 2 demonstrates the additional SNR outage zone in the uplink due to energy harvesting from the HAP as a function of P_0. The two main observations are as follows:

- At lower values of P_0, the gap between the two SNR outage zones is large. This calls for efficient energy transfer mechanisms at the HAP, optimizing the HAP parameters, and/or the deployment of supplementary energy sources.
- Since the nearby users can harvest a reasonable amount of energy, the gap between the two SNR outage zones reduces with increasing P_0. However, since only nearby users are able to achieve their target signal quality levels, a large number of users suffer from both the information transfer and energy harvesting perspectives. This fact reveals the importance of using spatially apart information receivers (e.g., DEs) along with supplementary energy sources (e.g., PBs).

Deployment of Dedicated Energy Sources

The deployment of PBs has been shown to be efficient in improving the average harvested energy per subscriber [10]. Although there are no backhaul requirements, PBs incur additional operation, maintenance, and deployment costs. Therefore, compared to a large number of randomly placed PBs, a small number of geographically planned PBs are more appealing.

Figure 4 demonstrates the effects of different harvesting strategies on the users located beyond R_2, namely,

- Harvesting from HAPs only.
- Harvesting from HAPs and randomly deployed PBs.
- Harvesting from HAPs and symmetrically deployed PBs (Fig. 3 for a graphical illustration).

The symmetric deployment refers to a deployment in which all PBs are placed in a circular fashion around the HAP at a fixed distance. On the other hand, the random deployment refers to an arbitrary placement of the PBs. First, it can be observed that the users located beyond R_2 suffer from 100 percent SNR outage probability when harvesting energy from the HAP. Second, the circular deployment of the PBs at optimal distance from the HAP can potentially reduce the SNR outage probability compared to the randomly deployed PBs. Moreover, the optimal location of PBs in symmetric deployment is a function of the number of PBs. For a limited number of PBs, the gains of symmetric deployment over random deployment are dominant over a wider range. On the other hand, with a larger number of PBs, the gains of the circular deployment become comparable to random deployment. However, the SNR outage tends to reduce significantly with the increase in the number of PBs. Nonetheless, continuously increasing the number of PBs (though unlikely) ultimately causes the random deployment to outperform the symmetric deployment of PBs.

The SNR outage can be minimized even further if the supplementary PBs are combined with the DEs. In a cellular network infrastructure with already deployed DEs, additional gains can therefore be achieved in a cost-efficient manner. A user in the SNR outage region can harvest energy from all PBs and then transmit the information through the nearest antenna element (which may not necessarily be a HAP). With this procedure, the SNR outage can be minimized significantly, as illustrated in Fig. 5, compared to the case in which a user transmits to the HAP only. With this DE-assisted PB setup, the useful-

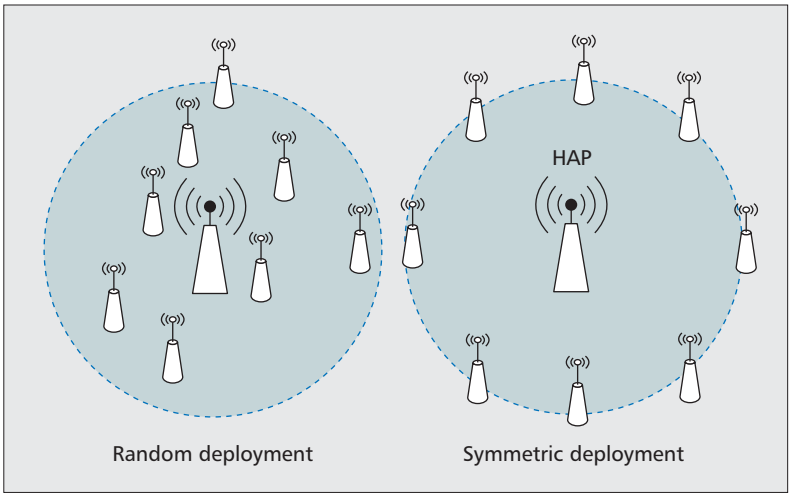


Figure 3. Graphical illustration of random and circular deployments of dedicated energy sources such as PBs or DE-assisted PBs.

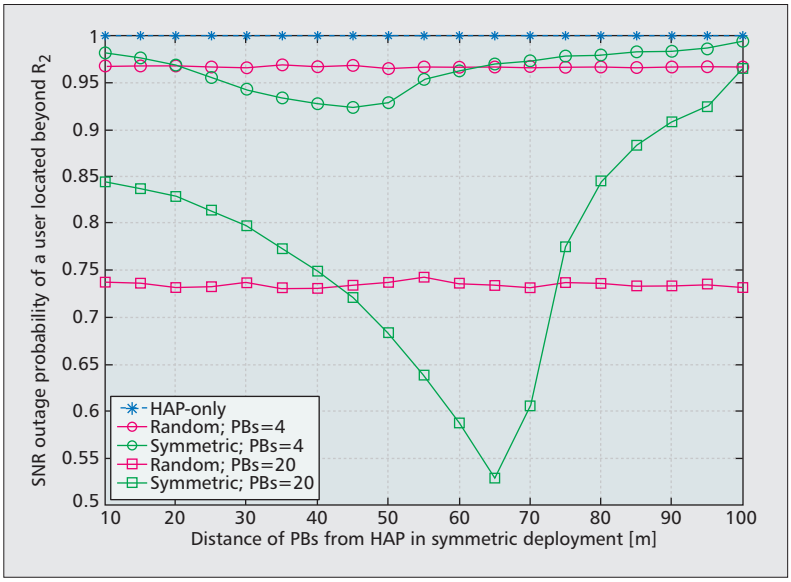


Figure 4. SNR outage probability of a user located beyond R_2 as a function of the distance of PBs from the HAP considering both the random and cicular deployments of PBs, for R = 100 m, β = 2, $P_{\max}$ = 1 mW, T/τ = 0.5, P_{HAP} = 40 W, η = 0.5.

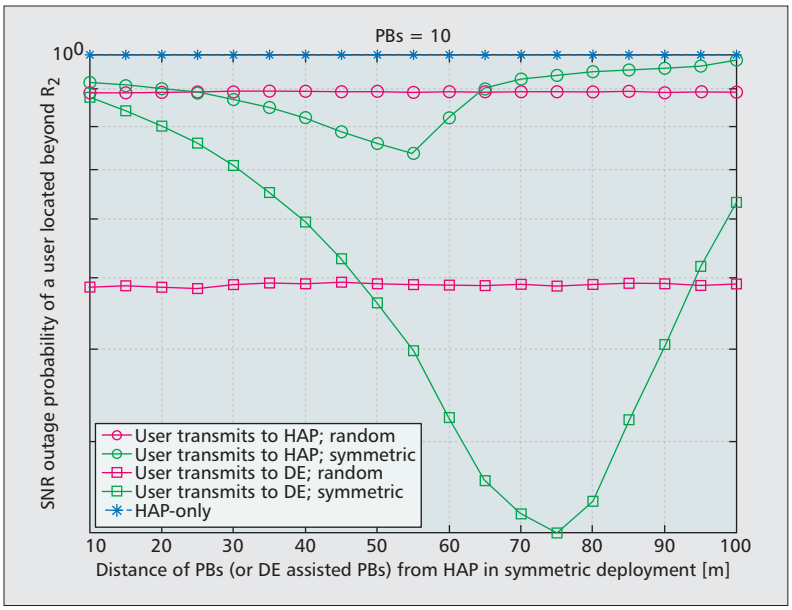


Figure 5. SNR outage probability of a user located beyond R_2 as a function of the distance of DE-assisted PBs from HAP considering both the random and circular deployments, for $R = 100$ m, $\beta = 2$, $P_{max} = 1$ mW, $T/\tau = 0.5$, $P_{HAP} = 40$ W, $\eta = 0.5$.

ness of the circular deployment of PBs (co-located with DEs) at optimal locations becomes more evident compared to the random deployment.

Other Design Considerations for WPCNs

On Request Energy Transfer/Cease Protocols: WPCNs rely mainly on a centralized entity (such as a HAP) that coordinates both WET and WIT. Depending on the objectives of network operators (e.g., throughput maximization [4]), a specific duration is reserved for the users for energy harvesting. Nevertheless, these users may accumulate sufficient energy from ambient RF sources as well. Therefore, it will be crucial to design lightweight protocols that allow communication between the HAP and harvesting devices (e.g., to request more energy from the HAP in the case of impending SNR outage or inform the HAP in the case of accumulating sufficient energy beforehand). This will reduce unnecessary resource consumption at the HAP that may lead to energy overflow at the harvesting devices. In addition, these protocols may also enable the subscribers to request additional energy from on-request dedicated power resources (e.g., PBs, as shown in Fig. 1), if available.

Coordination Among HAPs on Uplink Channels: Coordination among HAPs is a desirable feature of emerging WPCNs. The definition of coordination among HAPs (especially for uplink transmissions) needs to be rethought and modified. Note that if all HAPs coordinate and synchronize their energy transfer phases, it will help cell-edge users to accumulate higher energy levels. On the other hand, their coordination during the uplink transmission phase mitigates strong interference received at a HAP. This fact calls for optimizing the energy and information transfer time in which all HAPs can coordinate to maximize their mutual benefits on uplink transmission channels (i.e., accumulation of energy for cell-edge users in the downlink and interference mitigation at HAP in the uplink). Note that adapt- ing the energy harvesting duration (according to the channel conditions of the users) at all HAPs in a distributed fashion (e.g., [4] and other follow-up studies) does not allow coordination among HAPs. Thus, the cell-edge users may lose the benefits of interference from other HAPs during WET and a HAP will be exposed to higher interference (i.e., interference from nearby HAPs) during the uplink WIT. As such, to promote synchronization and coordination among HAPs, the time duration of WET and WIT need to be optimized jointly by all HAPs.

Harvesting-Constrained Scheduling: As mentioned earlier, traditional multiuser scheduling schemes need to be modified. In particular, the knowledge of downlink CSI can be exploited to estimate the energy that can be harvested by a user who is scheduled to transmit on the uplink channel. With this estimation, new scheduling schemes can be developed that can select a specific set of users who are expected to fulfill the minimum energy requirements (or minimum target rate constraints). On one hand, it helps to reduce unnecessary energy consumption of the energy sources. On the other hand, it results in zero energy outage probability at the scheduled user.

Acquisition of Downlink Channel State Information (CSI): The knowledge of downlink CSI plays a crucial role in improving the performance of multi-user scheduling schemes for uplink transmission in WPCNs. With the knowledge of downlink CSI, a HAP may adapt and optimize either its transmit power or its energy harvesting duration according to the channel conditions of the selected users and perform directional WET toward selected users. Thus, efficient CSI acquisition techniques are required at HAPs in order to utilize resources efficiently.

Offloading Users to Open Access Small Cells: Dense deployment of small cells is one of the most important features of emerging 5G cellular networks. However, the role of small cells as energy transmitters is not appealing yet due to their low transmit powers. Although small cells may not be very useful in transferring sufficient energy to cell-edge users who are associated to HAPs, they can assist the users that are offloaded to them. Efficient user-offloading mechanisms are thus crucial to identify and assist users in SNR outage zones. In this case, each small cell can serve as a potential HAP for a user who is located in its close vicinity. Thus, offloading the users in the SNR outage zone to their nearby small cells can significantly reduce resource consumption at HAPs while maintaining the desired SNR of the suffering users with reduced energy requirements.

Concluding Remarks

This article has highlighted the primary challenges of RF-based energy harvesting in a cellular network where several types of energy sources can coexist, namely dedicated (e.g., HAPs, distributed antenna elements, etc.), on-request dedicated (e.g., power beacons), and

ambient (e.g., interfering transmission from nearby HAPs) energy sources. The usefulness of different harvesting and transmission configurations have been analyzed and compared quantitatively from the perspective of users in the SNR outage zone. Numerical results have demonstrated the significance of a planned deployment of dedicated energy sources and information receivers (e.g., power beacons and distributed antenna elements) compared to an unplanned deployment. Finally, some possible solution techniques have been provided to improve the energy harvesting and information transfer efficiency of users in WPCNs, especially those located in the SNR outage zone.

References

[1] E. Hossain *et al.*, "Evolution Towards 5G Multi-Tier Cellular Wireless Networks: An Interference Management Perspective," *IEEE Wireless Commun.*, vol. 21, no. 3, June 2014, pp. 118–27.

[2] R. Zhang and C. K. Ho, "MIMO Broadcasting for Simultaneous Wireless Information and Power Transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, May 2013, pp. 1989–2001.

[3] X. Zhou, R. Zhang, and C. K. Ho, "Wireless Information and Power Transfer: Architecture Design and Rate-Energy Tradeoff," *IEEE Trans. Commun.*, vol. 61, no. 11, Nov. 2013, pp. 4754–67.

[4] H. Ju and R. Zhang, "Throughput Maximization for Wireless Powered Communication Networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, Jan. 2014, pp. 418–28.

[5] —, "Optimal Resource Allocation in Full-Duplex Wireless-Powered Communication Network," *IEEE Trans. Commun.*, vol. 62, no. 10, Sept. 2014, pp. 3528–40.

[6] M. Erol-Kantarci and H. Mouftah, "Radio-Frequency-based Wireless Energy Transfer in LTE-A Heterogenous Networks," *Proc. IEEE Symp. Computers and Communication (ISCC)*, June 2014, pp. 1–6.

[7] H. Chen *et al.*, "Harvest-Then-Cooperate: Wireless-Powered Cooperative Communications," arXiv preprint arXiv:1404.4120, 2014.

[8] A. A. Nasir *et al.*, "Relaying Protocols for Wireless Energy Harvesting and Information Processing," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, 2013, pp. 3622–36.

[9] Y. Shi *et al.*, "On Renewable Sensor Networks with Wireless Energy Transfer," *Proc. IEEE Intl. Conf. Comp. Commun. (INFOCOM)*, Apr. 2011, pp. 1350–58.

[10] K. Huang and V. K. Lau, "Enabling Wireless Power Transfer in Cellular Networks: Architecture, Modeling and Deployment," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, Feb. 2014, pp. 902–12.

[11] L. Liu, R. Zhang, and K.-C. Chua, "Wireless Information Transfer with Opportunistic Energy Harvesting," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, Jan. 2013, pp. 288–300.

[12] H. Ju and R. Zhang, "User Cooperation in Wireless Powered Communication Networks," arXiv preprint arXiv:1403.7123, 2014.

[13] Z. Ding and H. Poor, "Cooperative Energy Harvesting Networks with Spatially Random Users," *IEEE Signal Proc. Lett.*, vol. 20, no. 12, Oct. 2013, pp. 1211–14.

[14] L. Liu, R. Zhang, and K.-C. Chua, "Multi-Antenna Wireless Powered Communication with Energy Beamforming," *IEEE Trans. Commun.*, early access, Nov. 2014.

[15] B. Gurakan *et al.*, "Energy Cooperation in Energy Harvesting Communications," *IEEE Trans. Commun.*, vol. 61, no. 12, 2013.

[16] H. Tabassum *et al.*, "A Statistical Model of Uplink Inter-Cell Interference with Slow and Fast Power Control Mechanisms," *IEEE Trans. Commun.*, vol. 61, no. 9, July 2013, pp. 3953–66.

[17] X. Lu *et al.*, "Wireless Networks with RF Energy Harvesting: A Contemporary Survey," arXiv preprint arXiv:1406.6470, 2014.

Biographies

HINA TABASSUM (hina.tabassum@umanitoba.ca) received the B.E. degree in electronic engineering from the NED University of Engineering and Technology (NEDUET), Karachi, Pakistan, in 2004. During her undergraduate 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. degrees 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] (ekram.hossain@umanitoba.ca) 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. His current research interests include the 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 the *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. He 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 Fellow of the IEEE. 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.

ADEDAYO OGUNDIPE (ogundipa@cc.umanitoba.ca) obtained his B.Sc. in electronic and electrical engineering from the Ladoke Akintola University of Technology, Nigeria, in 2006. From February 2009 to August 2012 he worked for LM Ericsson Nigeria Ltd. as a service engineer before he started his M.Sc. in the Department of Electrical and Computer Engineering at the University of Manitoba, Canada, in May 2014. His research interest lies in resource management in energy-harvesting wireless systems.

DONG IN KIM [S'89, M'91, SM'02] (dikim@skku.ac.kr) 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 the *IEEE Transactions on Wireless Communications* from 2002 to 2011. From 2008 to 2011 he served as the co-editor-in-chief of the *Journal of Communications and Networks*. He is currently the founding Editor-in-Chief of *IEEE Wireless Communications Letters*, and has been serving as an editor of spread spectrum transmission and access for *IEEE Transactions on Communications* since 2001. He was the recipient of the Engineering Research Center (ERC) for Wireless Energy Harvesting Communications Award.

Although small cells may not be very useful in transferring sufficient energy to cell-edge users who are associated to HAPs, they can assist the users that are offloaded to them. Efficient user-offloading mechanisms are thus crucial to identify and assist users in SNR outage zones.

Wireless Energy Harvesting in Interference Alignment Networks

Nan Zhao, F. Richard Yu, and Victor C.M. Leung

ABSTRACT

Wireless energy harvesting (WEH) is becoming one of the key techniques in energy harvesting in wireless networks. On the other hand, interference alignment (IA) is a promising solution for interference management in wireless networks. Although plenty of effort has been conducted on WEH and IA, these two important areas have been addressed separately in most of the existing literature. In this article we provide an overview of WEH in IA networks, and present a unified framework to jointly study WEH and IA. To simultaneously optimize both information transmission (IT) and WEH performance in IA networks, we propose a power splitting optimization (PSO) algorithm. In addition, we study the power allocation problem in the proposed PSO algorithm. Simulation results are presented to compare the performance of the proposed schemes for WEH in IA networks. Some interesting research challenges are also presented for the WEH in IA networks.

INTRODUCTION

In wireless networks, energy consumption is becoming one of the key issues, due to the growing demand in wireless applications, a shortage in energy supply, and the need for environmental protection. Green communications have attracted much interest from both academia and industry. Recently, energy harvesting (EH) has emerged as an important method to achieve green wireless communications [1]. In EH, the energy captured from the ambient environment can be collected to provide a power supply for green self-sufficient wireless nodes [2]. As radio-frequency (RF) signals carry energy, they can be a new source for energy harvesting. Indeed, wireless energy harvesting (WEH) is becoming an important aspect of EH. Since RF signals can be used as a vehicle for both transmitting information and transferring energy in wireless networks, simultaneous wireless information and power transfer (SWIPT) has attracted much attention [3–10].

Some pioneering works on SWIPT have been done in [3, 4], in which the optimal information transmission (IT) versus WEH performance in the single-input single-output (SISO) channel is analyzed. In [5], Zhang and Ho study IT versus WEH performance in a multi-input multi-output (MIMO) broadcast network with two receivers, one for WEH and the other for IT. In [6], the optimal power splitting schemes are designed in the SISO and single-input multiple-output (SIMO) systems to achieve various trade-offs between the performance of WEH and IT. In [7], Chen *et al.* consider a multi-antenna system where the receiver can harvest wireless energy to support its wireless information transmission, and the tradeoff of wireless energy and information transfer is studied. There are some excellent works on SWIPT in interference channels [8–10]. In [8], Timotheou *et al.* focus on the SWIPT in MISO interference channels based on zero-forcing or maximum-ratio-transmission schemes. In [9], the authors investigate SWIPT in a two-user MIMO interference channel with time-switching receivers. A novel SWIPT scheme in a K-user SISO interference channel is proposed in [10], which first considers the special case of a two-user SWIPT system, and then extends to the K-user system.

On the other hand, interference is one of the most fundamental and challenging aspects of wireless communications. Consequently, interference management is another key issue in next generation wireless networks with hyper-dense heterogeneous cells. Recently, interference alignment (IA) has emerged as a promising solution to interference management in wireless networks thanks to its ability to approach the sum capacity of the interference channel [11–14]. In IA wireless networks, the transmitted signals of all the users are cooperatively designed through precoding matrices to constrain all the interferences into certain subspaces at the unintended receivers, and the desired signal can be achieved by the decoding matrix in the remaining interference-free subspace at each receiver. Due to its promising performance, IA has been successfully applied in various networks, e.g., heterogeneous networks, cognitive radios, multi-cell OFDMA networks, etc. [12].

Although there have been some excellent works on WEH and IA, these two important areas are usually studied separately in the literature. For example, in the existing WEH studies, recent advances in IA are largely ignored. On the other hand, in conventional IA networks, the interferences are usually leveraged to separate

Nan Zhao is with Dalian University of Technology.

F. Richard Yu is with Carleton University.

Victor C. M. Leung is with the University of British Columbia.

This research was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61201224 and 61372089, and the Fundamental Research Funds for the Central Universities under DUT14QY44.

0163-6804/15/$25.00 © 2015 IEEE

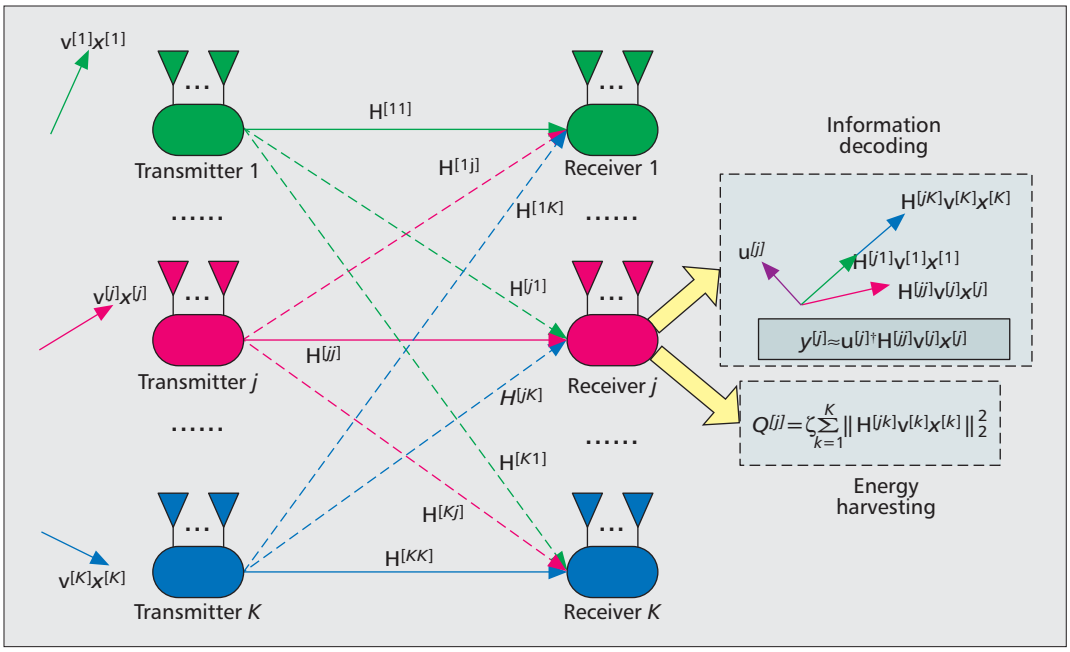


Figure 1. A linear IA wireless network with K MIMO users. (An example is shown for the information decoding (ID) and wireless energy harvesting performed at receiver j).

In practical multi-user networks, some users may want to transmit information at a high rate, while the other users may aspire to harvest wireless energy when their batteries are running out or their rate requirements are low at certain time slots. Thus it is reasonable for some of the IA receivers to harvest wireless energy from the transmitters during certain time slots.

out the desired signal and then discarded, instead of re-utilizing the interferences, which is a great waste of energy in wireless networks.

In this article we provide an overview of WEH in IA networks, and present a unified framework to jointly study WEH and IA. The main benefit of the proposed SWIPT scheme based on IA compared to the existing works on the SWIPT in interference channels lies in that the interferences among the users can be completely eliminated with the help of IA, and the transceivers are much easier to design by using the iterative IA algorithms. To simultaneously optimize both IT and WEH performance in IA networks, we propose a power splitting optimization (PSO) algorithm. In addition, we study the power allocation problem in the proposed PSO algorithm. Simulation results are presented to compare the performance of the proposed schemes for WEH in IA networks. Furthermore, we present some interesting research challenges for WEH in IA networks.

The rest of this article is organized as follows. First we describe the system model. Then a user selection scheme for WEH in IA networks is presented. We propose a power-splitting optimization algorithm for WEH with power allocation in IA networks, and the performance of the proposed schemes is presented. Some research challenges are presented. Finally, we conclude the article.

Wireless Energy Harvesting in Linear Interference Alignment Networks

IA can be achieved in time, frequency, or spatial dimensions [12]. In this article, linear IA in the spatial dimension is considered, i.e., IA in multi-user MIMO systems. A K-user IA network with one data stream[1] of each user is demonstrated in Fig. 1, where an example is shown for the information decoding (ID) and WEH performed at receiver j.

In conventional IA networks, only information transmission is performed. For an arbitrary user j, its signal $x^{[j]}$ is precoded by a unitary precoding vector $\mathbf{v}^{[j]}$. The coded signal $\mathbf{v}^{[j]}x^{[j]}$ is transmitted by the antennas from transmitter j to receiver j, which will also cause interference at other unintended receivers. At receiver j, the interferences from other transmitters, $\mathbf{H}^{[jk]}\mathbf{v}^{[k]}x^{[k]}$, $\forall k \neq j$ are constrained into the same subspace that is orthogonal to vector $\mathbf{u}^{[j]}$, through the cooperation of the precoding vectors of the other users. $\mathbf{H}^{[jk]}$ is the channel matrix from transmitter K to receiver j. Thus the desired signal of user j can be recovered as $\mathbf{u}^{[j]}\mathbf{H}^{[jj]}\mathbf{v}^{[j]}x^{[j]}$ at receiver j with all the interferences perfectly eliminated. Nevertheless, the interferences from the unintended transmitters in the IA network is aligned and discarded, which is a waste of resources.

In practical multi-user networks, some users may want to transmit information at a high rate, while the other users may aspire to harvest wireless energy when their batteries are running out or their rate requirements are low at certain time slots. Thus it is reasonable for some of the IA receivers to harvest wireless energy from the transmitters during certain time slots. If WEH is performed at receiver j as shown in Fig. 1, the received signals will not be processed by vector $\mathbf{u}^{[j]}$; instead, they are collected to recharge the battery. The power harvested at receiver K when it is dedicated to WEH, $Q^{[j]}$ can be expressed as Eq. (8) in [15], where $\zeta \in (0, 1)$ is a constant representing the energy conversion efficiency in the transducer for converting the harvested energy to electrical energy [5].

[1] This article mainly focuses on WEH in IA, instead of multiple data streams in IA. The case of WEH with more data streams of each user in IA can be easily extended.

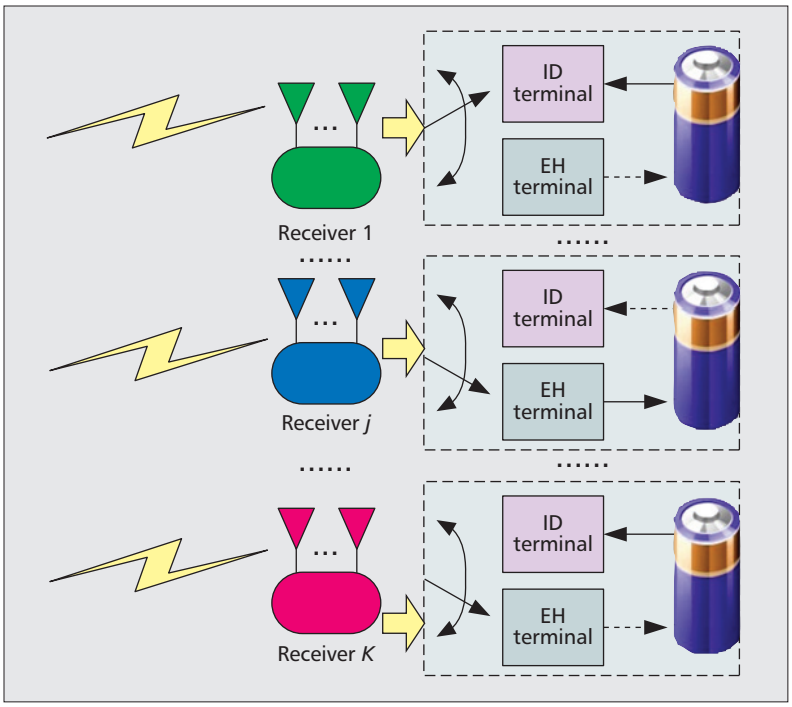


Figure 2. WEH-user selection in a K-user IA wireless network with both ID and EH terminals at each receiver.

User Selection Scheme for WEH in IA Networks

Assume that both EH and ID terminals are equipped at each receiver in the IA network. In each time slot, all the transmitters are active, and every receiver can be used as either an EH terminal or an ID terminal. Thus SWIPT can be achieved in the network as shown in Fig. 2. In practical systems, the receivers should not be all dedicated to WEH, because information transmission needs to be performed; on the other hand, we should not take all the receivers as ID terminals either, as some receivers may need to replenish energy to support their transmission and prolong the battery's life time. Thus only some of the receivers in the IA network should be selected as EH terminals at each time slot, and the others perform as ID terminals [15]. An example of user selection for WEH in the IA network is depicted in Fig. 2, where receiver 1 acts as an ID terminal, while receiver K is dedicated to WEH to recharge its battery.

A simple and natural idea of user selection for WEH in IA networks is to select the EH terminals in a round-robin principle, i.e., to assign WEH and IT users of the IA network during successive time slots in a circular order without any priority, which can be denoted as the round-robin selection (RRS) algorithm. The RRS algorithm is simple to implement; however, the performance of WEH and IT is not optimized, and its performance can be further improved.

In order to further improve the performance of the RRS algorithm for WEH in IA networks, a parameter called power-to-rate ratio (PRR) is defined to compare the instantaneous WEH and IT capability of an IA user. The PRR of the jth user, $\eta^{[j]}$, is the ratio of its harvested power $Q^{[j]}$ when adopted as an EH terminal to its transmission rate $R^{[j]}$ when used as an ID terminal. When $\eta^{[j]}$ is large, it means that it is better for user j to harvest energy than to transmit information at the time slot and vice versa. According to the PRR, we propose a PRR-based selection (PRRS) algorithm for WEH user selection in IA networks, in which the L users with largest PRRs are selected as WEH users at a certain time slot, $L \le K$, and the remaining $K - L$ users are dedicated to IT. In the PRRS algorithm, WEH can be performed by the users with higher WEH capability and lower IT capability, and both WEH and IT performance of the IA network can be improved simultaneously.

Comparing the RRS and PRRS algorithms for WEH in IA networks, the following observations can be obtained.

- The RRS algorithm is much easier to implement, due to the simple round-robin principle for selection it used, while the PRRS algorithm is more complex, because the PRR parameters of all the users in the IA network should be calculated based on the instantaneous WEH and IT capability, and the WEH users are selected accordingly.
- Both WEH and IT performance of the PRRS algorithm is better than that of the RRS algorithm with the same number of WEH receivers L, due to the selection according to the PRR parameters in the PRRS algorithm, which reflects the WEH versus IT performance of these IA users.
- In both the RRS and PRRS algorithms, only user selection is performed to achieve WEH in IA networks, and a certain receiver may be devoted to either WEH or IT. Thus the WEH and IT performance may not be continuously optimized according to the specific requirements of rate and energy. Actually, an IA receiver can be used as EH and ID terminals simultaneously by power splitting, which will be introduced in the next section.

Power-Splitting Optimization Algorithm for WEH with Power Allocation in IA Networks

In this section a power-splitting optimization algorithm with power allocation is proposed to perform WEH and IT simultaneously at each receiver in IA networks. The performance of the user-selection and PSO algorithms for WEH in IA networks is also analyzed.

Power-Splitting Optimization for WEH in IA Networks

In the RRS and PRRS algorithms described above, the specific users' requirements are not considered. For example, a user may want to transmit information and harvest energy simultaneously at a time slot; however, it may be selected to perform WEH according to the user selection algorithms, and the rate requirement of this user cannot be fully satisfied. When power splitters can be equipped at the receivers, the special require-

ments of each user can be satisfied at each time slot, by splitting the received power into two parts, i.e., ID and EH. In this subsection we propose a power-splitting optimization algorithm for WEH in IA networks as shown in Fig. 3. In addition, power allocation for the PSO algorithm is studied.

Assume that both EH and ID terminals are equipped at each receiver in the IA network shown in Fig. 3. A power splitter is equipped at each receiver, which can be used to induce the received power from the antennas to the ID or EH terminals according to the requirements of the system [6]. $\rho^{[j]} \in [0, 1]$ is the portion of signal power that is split to the ID terminal at receiver j, and correspondingly $1 - \rho^{[k]}$ is the portion of power used for WEH. In the RRS and PRRS algorithms of the previous section, $\rho^{[j]}$ can be either set to 0 or 1, which means user j is dedicated to WEH or IT through user selection, respectively. By contrast, the PSO algorithm can optimize the WEH and IT performance over ρ continuously, which means WEH and ID can be performed at each receiver simultaneously according to the specific requirements of the system. Therefore, the PSO algorithm is more flexible to use and can well satisfy the needs of all the users in IA networks.

We should also define two parameters to be used in the PSO algorithm, i.e., $\alpha^{[j]}$ and $\beta^{[j]}$, the sum of which is equal to 1. $\alpha^{[j]}$ and $\beta^{[j]}$ can denote the weights for the specific requirements of the needed rate and energy of user j, respectively. When $\alpha^{[j]}$ becomes larger, it means that the IT requirement of user j is relatively high or its battery power at receiver j is sufficient; when $\alpha^{[j]}$ is set smaller, it means its battery is running out or its rate requirement is relatively low. Thus $\alpha^{[j]}$ should be set carefully according to the requirements of user j at each time slot.

Assume that the transmitted power of each user in the IA network is P_t. We can define the objective function $\mathcal{F}_1$ of the PSO algorithm as

$$\sum_{j=1}^{K}\left(\alpha^{[j]}R_{\rho^{[j]}}^{[j]}+\beta^{[j]}\left(1-\rho^{[j]}\right)Q^{[j]}\right),$$

where $R_{\rho^{[j]}}^{[j]}$ is the transmission rate of user j with the portion of power split to the jth ID terminal equal to $\rho^{[j]}$, and $Q^{[j]}$ is the harvested power of receiver j when it is only served as an EH terminal.

In the optimization problem of the PSO algorithm, the objective function $\mathcal{F}_1$ should be maximized over the variables $\rho^{[1]}, \rho^{[2]}, \ldots, \rho^{[K]}$, and the optimal solution $\rho^{*[j]}$ can be obtained, $j = 1, 2, \ldots, K$. Besides, the optimization problem of the PSO algorithm is convex, and its optimal solution can be easily calculated. Furthermore, we find that the closed-form optimal solution $\rho^{*[j]}$ of user j is not affected by the α and β parameters of other users, i.e., the specific WEH and IT requirements in the network will not interact among users.

Power Allocation for the PSO Algorithm in IA Networks

In the above discussions, equal power P_t is assumed at each transmitter, and power allocation is not involved. In practical systems, the wireless channel is usually not symmetric, and the IT performance may be varied dramatically according to the instantaneous status of channel fading. Therefore, power allocation should be considered to guarantee the performance of WEH and IT of IA networks.

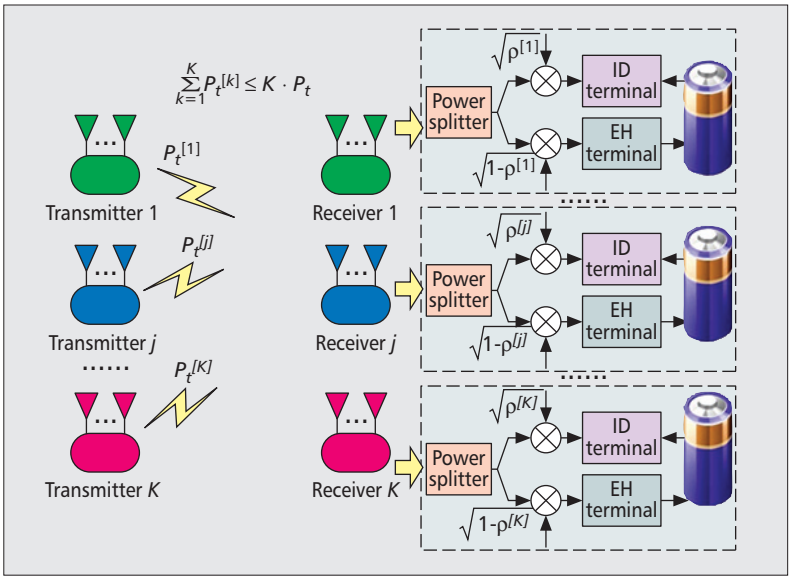


Figure 3. PSO scheme for WEH in the K-user IA network with power allocation.

Assume that the receivers are self-sufficient by wireless energy harvesting, and the transmitters are supplied by the power grid. The sum of the transmitted power of all the transmitters is subject to a constraint, which is equal to $K \cdot P_t$ as in Fig. 3. The objective function of optimization in the PSO algorithm should be updated as $\mathcal{F}_2$, in which the $R_{\rho^{[j]}}^{[j]}$ and $Q^{[j]}$ in $\mathcal{F}_1$ are changed into $\hat{R}_{\rho^{[j]}}^{[j]}$ and $\hat{Q}^{[j]}$, respectively, when the transmitted power of the IA users are $P_t^{[1]}, P_t^{[2]}, \ldots, P_t^{[K]}$. In the optimization problem of the PSO algorithm with power allocation, the objective function $\mathcal{F}_2$ should be maximized over the variables $\rho^{[1]}, \rho^{[2]}, \ldots, \rho^{[K]}$ and $P_t^{[1]}, P_t^{[2]}, \ldots, P_t^{[K]}$, and the optimal solutions $\rho^{*[j]}$ and $P_t^{*[j]}$ can be obtained, $j = 1, 2, \ldots, K$.

When power allocation is considered, the optimization problem of the PSO algorithm is not convex due to the product of $\rho^{[j]}$ and $P_t^{[j]}$, and its closed-form solutions are difficult to obtain. Fortunately, there are many simple but effective methods (e.g., the interior-point algorithm) for solving the continuous optimization problems.

Two extremes of the power allocation problem in the PSO algorithm are interesting and worth noting, i.e., $\forall j$, $\alpha^{[j]} = 1$ and $\alpha^{[j]} = 0$. When $\forall j$, $\alpha^{[j]} = 1$, $\rho^{[j]}$ will all converge to 1. Only the ID terminals are active at all the receivers, and the optimization problem becomes a conventional power allocation problem in IA networks, which can be solved by the famous "water-filling" power allocation strategy. When $\forall j$, $\alpha^{[j]} = 0$, $\rho^{[j]}$ will all turn to 0. Only WEH is performed at all the receivers, and this can happen when the batteries at the receivers are all at low levels and need to be recharged.

In practical systems, these two extremes can hardly happen. A common situation is that some

of the receivers with low-level batteries will replenish more energy with a low transmission rate, while the others may have sufficient power supply of their batteries, and they wish to transmit more information with less harvested energy. The PSO algorithm with power allocation can satisfy the specific requirements of all the IA users through power splitting to perform WEH and IT at each receiver simultaneously.

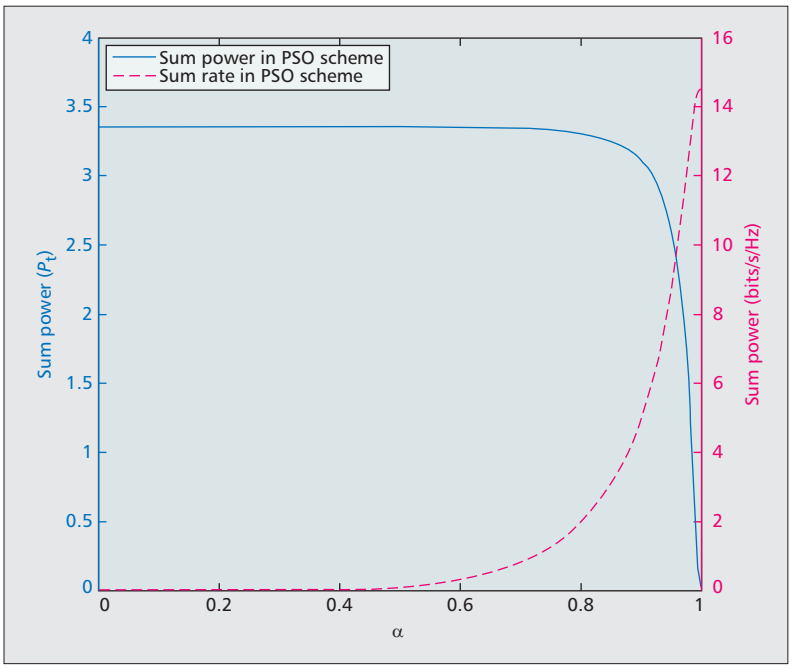


Figure 4. Comparison of sum harvested power and sum rate of the PSO algorithm with different values of α in five-user IA network, when the average received SNR is 10 dB.

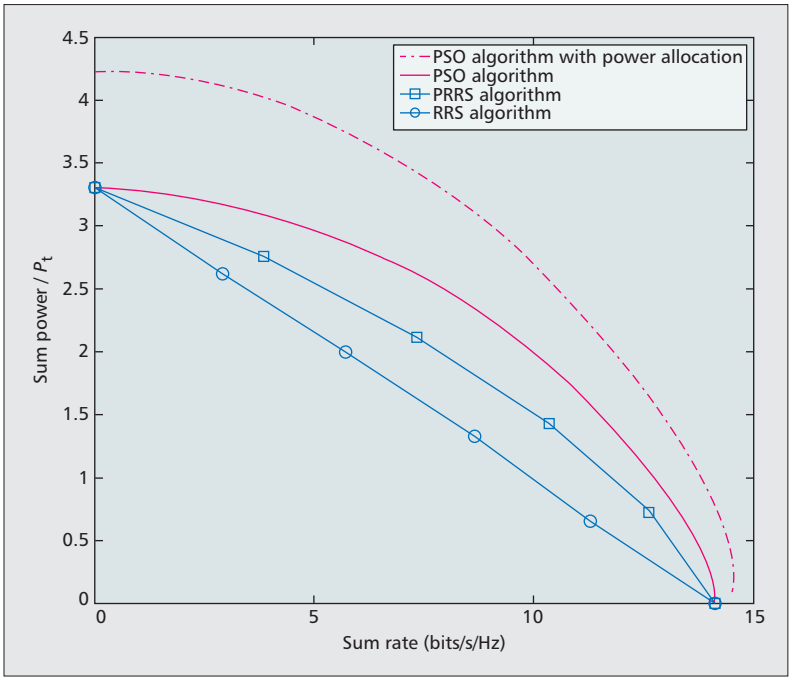


Figure 5. Power-rate tradeoffs of the PSO algorithms with and without power allocation, and the PRRS and RRS algorithms, for WEH in the five-user IA network. Average received SNR is 10dB.

Simulation Results and Discussions

A five-user IA network with one data stream for each user is considered. Three antennas are equipped at each transceiver according to the feasibility of IA [12]. In the simulation, Rayleigh block channel fading is used, and perfect channel state information (CSI) is assumed to be available at each node. The path loss is set to 0.1, i.e., each element in the channel coefficient matrix $\mathbf{H}^{[jk]}$ follows $\mathcal{CN}(0, 0.1)$. ζ is equal to 0.5. The conventional iterative algorithm is utilized to obtain the solutions of IA transceivers [12]. In the PSO algorithm, the values of α of all the users is set to the same value, i.e., $\alpha^{[1]} = \alpha^{[2]} = \ldots = \alpha^{[5]}$. The average received SNR of all the IA users is 10 dB.

Figure 4 shows the sum harvested power and sum rate in a five-user IA network with different values of α, when the PSO algorithm is performed without power allocation. From the results, we can observe that when α becomes larger in the PSO algorithm, the sum rate of the network increases, and the sum harvested power of the network decreases. Also, we can see that when $\alpha^{[k]} = \alpha$, $\forall k \in \{1, 2, \ldots, K\}$, is set below 0.4, the sum rate of the network will be 0, and all the receivers are adopted as EH terminals. Thus, $\alpha^{[k]} = \alpha$, $\forall k \in \{1, 2, \ldots, K\}$, can be set in a reduced domain of [0.4 , 1].

The power-rate trade-offs of the PSO algorithm with and without power allocation, the PRRS and RRS algorithms, are shown in Fig. 5. From the results, we can observe that the power-rate performance of the PSO algorithm is better than that of the PRRS and RRS algorithms, because WEH and IT can be performed simultaneously at each receiver and the power split to the EH and ID terminals is optimized. When power allocation is applied to the PSO algorithm, its performance can be significantly improved due to optimization of the allocated power among the transmitters with the sum transmitted power constrained. The performance of the PRRS algorithm is better than that of the RRS algorithm when $0 < L < K$, due to the user selection method based on the parameters of η in the PRRS algorithm. When $L = 0$ or $L = K$, the performance of the PRRS and RRS algorithms is the same, because no selection is performed in these two situations.

Although the performance of the PSO algorithm is much better than that of the PRRS and RRS algorithms, its computational complexity is much higher. Therefore, these algorithms should be adopted in different scenarios according to the requirements of the systems.

Figure 6 shows the average power harvested, transmission rate, transmitted power allocated, and corresponding parameter ρ of the users in the power-allocation PSO algorithm with different values of α. The values of α of the five users are set to 0.05, 0.2, 0.35, 0.5, and 0.65, respectively. From the results we can see that the average power allocated to each user and the parameter ρ of each user can be adjusted according to the values of α set by each user, and the expected ID and EH performance can be achieved. For example, $\alpha^{[1]}$ is very small ($\alpha^{[1]} =$

0.05), which means that the battery of receiver 1 is running out, and it wants to harvest more energy to recharge its battery instead of transmitting information. Thus its transmitted power $P_t^{[1]}$ and portion of power splitting $\rho^{[1]}$ is low, and it can collect more energy with a low transmission rate. In contrast, the power supply of receiver 5 is sufficient, and it wants to transfer more information than energy, with $\alpha^{[5]}$ set to 0.65. Consequently, its transmitted power $P_t^{[5]}$ and the portion of power splitting $\rho^{[5]}$ is high, and its transmission rate is high with almost no energy harvested.

Research Challenges

Despite the potential vision of WEH in IA wireless networks, several significant research challenges remain to be addressed. In this section we present some research challenges on WEH in IA wireless networks.

Proper Setting of Parameter α: $\alpha^{[j]}$ is an important parameter in the optimization of the proposed algorithms to make a tradeoff between IT and EH for user j. Although $\alpha^{[j]}$ can be set according to the specific requirements of user j at each time slot, how to determine the accurate value of $\alpha^{[j]}$ according to the status of the batteries and rate requirements is not discussed in this article; it will be studied in our future work.

Joint Optimization of Precoding Matrices and Power Splitting: In the proposed PSO algorithm, the precoding and decoding matrices are first obtained through the iterative IA algorithm, and then the power splitting is optimized, which is easy to implement. If we do not pursue perfect alignment of interferences, and optimize the precoding matrices and power splitting jointly, the performance of SWIPT can be further improved. Nevertheless, jointly optimizing precoding matrices and power splitting is a non-convex problem, and it is challenging to solve it in practical systems with lower complexity.

Topology Management of IA Networks for WEH: Most of the existing works on IA are focusing on symmetric networks. To meet the requirements of practical applications, asymmetric IA networks should be considered based on various path losses. Topology management of IA networks for WEH should be studied. When a receiver is located close to the center of the network, it may be more suitable for energy harvesting; otherwise, it may be more suitable to transmit information. Therefore, it is essential to design a topology management of IA networks for WEH, which can make this technique more suitable for practical systems.

Opportunistic IA networks for WEH: A user selection scheme is proposed for WEH in IA networks; however, only the users that already exist in the IA network are selected as EH or IT users. When opportunistic IA is adopted in WEH, some of the users are dedicated to form the IA network, while the others can harvest energy in each time slot. Thus, a large number of candidates would result in better WEH performance. It is interesting to study WEH in the framework of opportunistic IA networks.

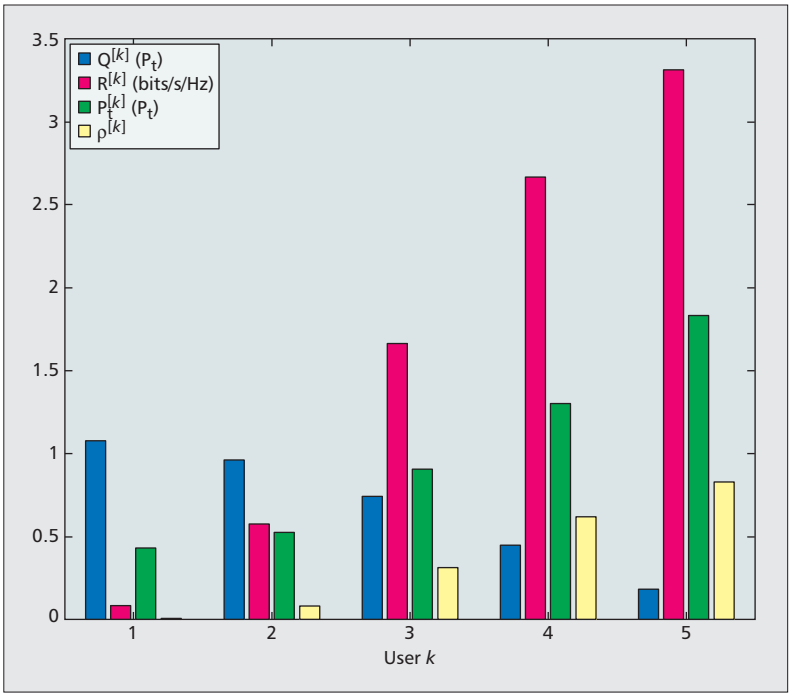


Figure 6. The average power harvested, transmission rate, transmitted power allocated, and corresponding parameter ρ of users in the power-allocation PSO algorithm with different values of α, when the average received SNR is 10 dB. $\alpha^{[1]} = 0.05$, $\alpha^{[2]} = 0.2$, $\alpha^{[3]} = 0.35$, $\alpha^{[4]} = 0.5$, $\alpha^{[5]} = 0.65$.

Conclusions

In this article we presented an overview of two emerging technologies in wireless networks: wireless energy harvesting (WEH) and interference alignment (IA). A unified framework was proposed to jointly study WEH and IA. The PRRS and RRS algorithms were introduced to select some of the IA users to perform WEH, and the other users to decode information. To optimize WEH and IT performance of IA networks, a PSO algorithm was proposed to optimize the power split to the EH and ID terminals of each receiver, and power allocation of the PSO algorithm was studied. Simulation results were presented to show the effectiveness of the proposed algorithms for WEH in IA networks. Finally, we presented some research challenges for WEH in IA networks.

Acknowledgment

We thank the editor and reviewers for their detailed reviews and constructive comments, which have helped to improve the quality of this article.

References

[1] Z. G. Wan, Y. K. Tan, and C. Yuen, "Review on Energy Harvesting and Energy Management for Sustainable Wireless Sensor Networks," *Proc. IEEE ICCT'11*, Jinan, China, Sept. 2011 (invited) , pp. 362–67.

[2] R. V. Prasad *et al.*, "Reincarnation in the Ambiance: Devices and Networks with Energy Harvesting," *IEEE Commun. Surv. Tutor.*, vol. 16, no. 1, Feb. 2014, pp. 195–213.

[3] L. R. Varshney, "Transporting Information and Energy Simultaneously," *Proc. IEEE ISIT'08*, Toronto, ON, July 2008, pp. 1612–16.

[4] P. Grover and A. Sahai, "Shannon Meets Tesla: Wireless Information and Power Transfer," *Proc. IEEE ISIT'10*, Austin, TX, June 2010, pp. 2363–67.
[5] R. Zhang and C. K. Ho, "MIMO Broadcasting for Simultaneous Wireless Information and Power Transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, May 2013, pp. 1989–2001.
[6] L. Liu, R. Zhang, and K.-C. Chua, "Wireless Information and Power Transfer: A Dynamic Power Splitting Approach," *IEEE Trans. Commun.*, vol. 61, no. 9, Sept. 2013, pp. 3990–4001.
[7] X. Chen, C. Yuen, and Z. Zhang, "Wireless Energy and Information Transfer Tradeoff for Limited-Feedback Multiantenna Systems with Energy Beamforming," *IEEE Trans. Vehic. Tech.*, vol. 63, no. 1, Jan. 2014, pp. 407–12.
[8] S. Timotheou, I. Krikidis, and B. Ottersten, "MISO Interference Channel with QoS and RF Energy Harvesting Constraints," *Proc. IEEE ICC'13*, Budapest, Hungary, Sept. 2013, pp. 4191–96.
[9] J. Park and B. Clerckx, "Joint Wireless Information and Energy Transfer in a Two-User MIMO Interference Channel," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, Aug. 2013, pp. 4210–21.
[10] S. Lee, L. Liu, and R. Zhang, "Collaborative Wireless Energy and Information Transfer in Interference Channel," *IEEE Trans. Wireless Commun.*, to appear. 17
[11] V. R. Cadambe and S. A. Jafar, "Interference Alignment and Degrees of Freedom of the K-User Interference Channel," *IEEE Trans. Info. Theory*, vol. 54, no. 8, Aug. 2008, pp. 3425–41.
[12] S. A. Jafar, "Interference Alignment — A New Look At Signal Dimemsions in a Communication Network," *Found. Trends Commun. Info. Theory*, vol. 7, no. 1, 2010, pp. 1–130.
[13] N. Zhao *et al.*, "A Novel Interference Alignment Scheme based on Sequential Antenna Switching in Wireless Networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, Oct. 2013, pp. 5008–21.
[14] X. Chen and C. Yuen, "Performance Analysis and Optimization for Interference Alignment over MIMO Interference Channels with Limited Feedback," *IEEE Trans. Signal Proc.*, vol. 62, no. 7, Apr. 2014, pp. 1785–95.
[15] N. Zhao, F. R. Yu, and Victor. C. M. Leung, "Simultaneous Wireless Information and Power Transfer in Interference Alignment Networks," *Proc. IWCMC'14*, Nicosia, Cyprus, Aug. 2014 (invited), pp. 1–5.

BIOGRAPHIES

NAN ZHAO [M] (zhaonan@dlut.edu.cn) is an associate professor in the School of Information and Communication Engineering at Dalian University of Technology, China. He received the B.S. degree in electronics and information engineering in 2005, the M.E. degree in signal and information processing in 2007, and the Ph.D. degree in information and communication engineering in 2011, from Harbin Institute of Technology, Harbin, China. From June 2011 to June 2013 he did postdoctoral research in Dalian University of Technology, Dalian, China. His recent research interests include interference alignment, cognitive radio, wireless power transfer, and optical communications. He has published more than 50 papers in refereed journals and international conferences. He serves as an area editor of the *AEU-International Journal of Electronics and Communications*. Additionally he served as a technical program committee (TPC) member for many conferences, e.g., Globecom, VTC, WCSP.

F. RICHARD YU [SM] (richard.yu@carleton.ca) is an associate professor at Carleton University, Canada. He received the IEEE Outstanding Leadership Award in 2013, the Carleton Research Achievement Award in 2012, the Ontario Early Researcher Award (formerly the Premier's Research Excellence Award) in 2011, the Excellent Contribution Award at IEEE/IFIP TrustCom 2010, the Leadership Opportunity Fund Award from Canada Foundation of Innovation in 2009, and the Best Paper Awards at IEEE ICC 2014, Globecom 2012, IEEE/IFIP TrustCom 2009, and the Int'l Conference on Networking 2005. His research interests include cross-layer design, security, green IT and QoS provisioning in wireless networks. He serves on the editorial boards of several journals, including *IEEE Transactions on Vehicular Technology* and *IEEE Communications Surveys and Tutorials*. He has served on the Technical Program Committee (TPC) of numerous conferences, as the TPC co-chair of IEEE INFOCOM-MCV'15, Globecom'14, INFOCOM-MCC'14, Globecom'13, GreenCom'13, CCNC'13, INFOCOM-CCSES'12, ICC-GCN'12, VTC'12S, Globecom'11, INFOCOM-GCN'11, INFOCOM-CWCN'10, IEEE IWCMC'09, VTC'08F and WiNITS'07, as the publication chair of ICST QShine'10, and the co-chair of ICUMT-CWCN'09.

VICTOR C. M. LEUNG [F] (vleung@ece.ubc.ca) is a professor of electrical and computer engineering and holder of the TELUS Mobility Research Chair at the University of British Columbia (UBC). His research is in the areas of wireless networks and mobile systems, in which he has coauthored more than 800 technical papers in archival journals and refereed conference proceedings, several of which have won best paper awards. Dr. Leung is a Fellow of IEEE, a Fellow of the Royal Society of Canada, a Fellow of the Canadian Academy of Engineering and a Fellow of the Engineering Institute of Canada. He is serving or has served on the editorial boards of JCN, IEEE JSAC, *Transactions on Computers*, *Wireless Communications*, and *Vehicular Technology*, *Wireless Communications Letters*, and several other journals. He has provided leadership to the technical program committees and organizing committees of numerous international conferences. Dr. Leung was the recipient of the 1977 APEBC Gold Medal, NSERC Postgraduate Scholarships from 1977-1981, a 2012 UBC Killam Research Prize, and an IEEE Vancouver Section Centennial Award.

A Survey of Energy Harvesting Communications: Models and Offline Optimal Policies

Yejun He, Xudong Cheng, Wei Peng, and Gordon L. Stüber

ABSTRACT

As people pay more attention to environmental protection and energy conservation issues, energy consumption in communications have become a hot research field. In wireless communications networks such as wireless sensor networks, traditional battery-operated devices or nodes have a short lifetime and die after the batteries are depleted, and replacing the batteries may be very costly and sometimes will be impossible. Therefore, energy harvesting (EH) communications have become a good means to solve this problem. EH communications mean the nodes can continue working by harvesting ambient energy. EH communications are different from the traditional battery-operated communications, so we need new models and optimal transmission policies to maximize the throughput. In this article we review different methods of harvesting the ambient energy in EH communications and the models of EH communications. We focus on offline optimal policies, then compare different policies and classify them into certain types. Finally, we propose several open research challenges and directions for future work.

INTRODUCTION

With the widespread deployment of wireless networks and devices, energy consumption management in wireless devices has become a recent topic of interest. Some wireless network devices, such as cellular phones, can maintain operation by simply charging or changing their batteries, a process that may be very difficult for other types of wireless network devices. For example, with wireless sensor network deployments, the sensor area may be large and the sensors randomly located, so the replacement of sensor batteries after the batteries have been depleted is expensive or impossible. Energy harvesting (EH) approaches have been proposed for such cases, whereby the lifetime of wireless devices is extended by harvesting ambient energy.

Energy harvesting devices can harvest ambient energy from sources such as the sun, radio waves, and vibration [1], and turn these sources into electricity for usage or storage. With the development of integrated circuits and other low-power electronic devices, energy harvesting technologies are entirely feasible. A new Energy harvesting-Communication networks: OPtimization and demonStration (E-CROPS) project began its work in February 2013 [2]. The project is funded by European coordinated research on long-term CHallenges in Information and Communication Sciences and Technologies-European Research Area-net (CHIST-ERA), whose purpose is to use energy harvesting and smart energy management technologies in communication and mobile devices to achieve an optimal balance between the quality of service (QoS), performance, and efficient use of energy.

Energy harvesting technologies can be used in a wide range of applications, including wireless sensor networks, building automation networks, machine to machine communications, and the smart grid [3]. Such wireless networks can become self-sustaining and maintenance-free by using EH technology to prolong the lifetime of the network devices. EH devices operate by harvesting ambient energy, which is fundamentally different from the traditional battery-operated devices. Battery-operated devices have a fixed amount of reliable energy, whereas EH devices harvest a random and uncertain amount of energy. Therefore, it is critical to optimize the transmission policy for EH devices. Optimal transmission policies are required for efficient usage of harvested energy, to maximize the amount of transmitted data by a given deadline or to minimize the transmission completion time by making full use of the energy.

For a single-hop model, Yang and Ulukus [4] derived the optimal packet scheduling policy to minimize the transmission completion time and gave some important lemmas for designing the optimal policy. Ozel *et al.* [5] suggested a *directional water-filling* algorithm that takes into account both the channel condition and energy capacity to maximize the throughput. Orhan *et al.* [6] proposed a *directional glue pouring* algorithm to compute the optimal policy with processing energy cost for communication on a fading channel. Maria and Miquel [7] suggested

Yejun He and Xudong Cheng are with Shenzhen University.

Wei Peng is with Huazhong University of Science and Technology.

Gordon L. Stüber is with Georgia Institute of Technology.

0163-6804/15/$25.00

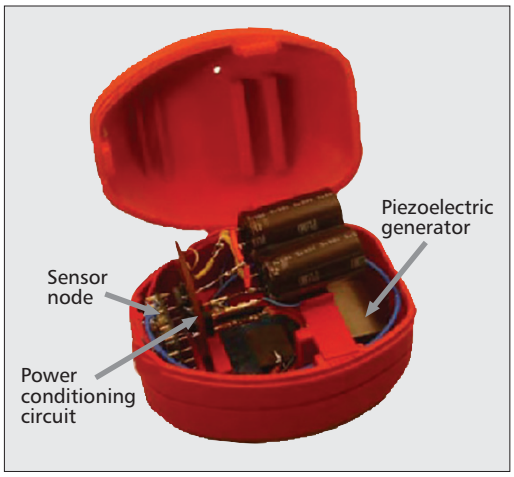


Figure 1. A real EH node [10].

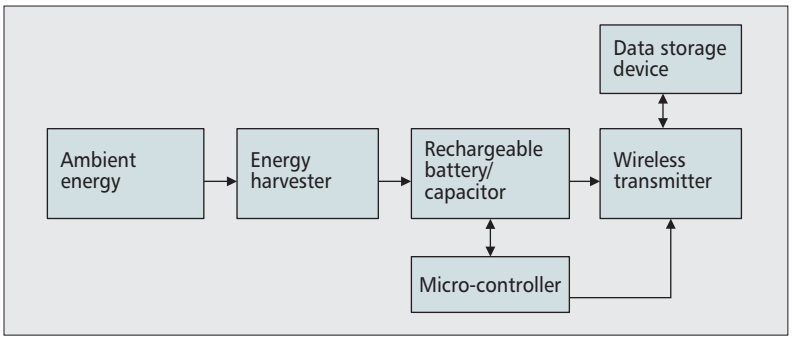


Figure 2. Practical model diagram of EH systems.

using a QoS constraint to prevent battery overflows.

For a two-hop EH communications model, Yaming Luo *et al.* [8] proposed an improved *directional water-filling* algorithm that maximizes throughput. Deniz and Bertrand [9] considered both full-duplex and half-duplex communication with an EH relay node in the two-hop model.

This article presents an overview of the optimal transmission policies in EH communications. Different methods of harvesting ambient energy and the models of EH communications are reviewed, with a focus on off-line optimal policies. Finally, we discuss several open problems and propose future research challenges and directions.

Methods of Energy Harvesting

Many sources of ambient energy that can be harvested and used, for example, piezoelectric harvesting devices, can convert mechanical energy into electrical energy, and we can see a real EH node in Fig. 1. Different types of harvesting devices can scavenge for different kinds of energy such as solar, radio-frequency, thermal energy, etc. However, the amount of available energy to be harvested varies over space and time. Solar energy is available in the daytime while it vanishes at night, and radio frequency energy may be available in urban areas, while wind energy is available in open areas. Hence, it is very important to choose the appropriate energy harvesting method based on the network's energy harvesting environment. Table 1 lists various energy harvesting methods and their power generation capability [1].

A practical model diagram for EH systems is shown in Fig. 2. The energy harvester converts ambient energy into electrical energy, which is stored in the rechargeable battery or capacitor, which is called the energy buffer. The rechargeable battery or capacitor in turn provides power for the micro-controller and transmit module. The micro-controller can manage the entire node, including power supply, information to transmit or receive. Usually, there is a data storage device that is called the data buffer to store the data that have been harvested but not yet transmitted.

System Model

There are two different approaches for designing optimal transmission policies: online and offline. With online approaches, the nodes only have statistical knowledge of the energy harvesting process, while offline approaches assume that the node has full knowledge of the amount and arrival time of the harvesting energy. Offline approaches are an idealistic situation, but can

Energy source	Power density	Advantages	Disadvantages
Solar	15 mW/cm^3	Sufficient energy in the daytime, high output voltage	Disappear at night
Vibration (piezoelectric)	200 uW/cm^3	Without voltage source	Brittle materials
Thermoelectric	40 uW/cm^2	Long life, reliable with low maintenance	Low energy conversion efficiency
Acoustic noise	960 nW/cm^3	High energy conversion efficiency	Rare environments with high acoustic noise levels
Airflow	1 mW/cm^2	Sufficient in certain place and time	Big size
Radio frequency	1 uW/cm^2	Sufficient in urban areas	Few in suburbs

Table 1. Comparison of different harvesting methods [1].

provide analytical and heuristic solutions for designing the optimal transmission strategy [7]. Many studies have been done to analyze point-to-point offline optimal transmission policies. The entire collection of EH communications models may be classified into single-hop models, two-hop models, and multi-hop models. A single-hop EH communications model is shown in Fig. 3a. The transmitter is an EH node having a data queue and an energy queue, where both the data and the energy are packetized, such that the EH communications process is modeled as a packet arrival and transmit process.

We define E_i as the amount of energy from the ith energy harvesting and B_i as the number of bits in the ith data packet arrival. The total energy consumed by the transmitter up to this time t is $E(t)$, while the total transmitted data is $B(t)$. Then $E_{\max}$ denotes the energy buffer capacity and $B_{\max}$ denotes the data buffer capacity, which are the red lines in Fig. 3a, where the energy and data that exceed the red lines must be discarded. Let the rate-power function $r(p)$ be the transmission rate at a transmission power $p(t)$; we also define $h(t)$ as the channel state information (CSI), and $r(p)$ is a non-negative, monotonically increasing and strictly concave function as shown in [3, 5, 6].

The single-hop model is a simplified situation that admits easier analysis. However, in many scenarios, the channel condition from the source (transmitter) to the destination (receiver) is such that the source node cannot transmit data directly to the destination node. In this case, a relay node is needed for data storage and forwarding, resulting in a two-hop or multi-hop transmission. Figure 3b shows the two-hop EH communications model, where E_i^s is the amount of energy from the ith energy harvesting at the source node, and E_i^r denotes the amount of energy from the ith energy harvesting at the relay node. The multi-hop model in Fig. 3c is more realistic and complicated, and we can divide it into simplified single-hops for analysis.

In energy harvesting communications transmission policies, there are two causality constraints for all the models: the energy may not be used before it is harvested, and the data packet cannot be delivered before it has arrived [4]. The total consumed energy cannot be more than all the harvested energy, and the total transmitted data cannot be more than all the arrived data, which is very different from battery-operated systems. If we consider the energy buffer capacity and data buffer capacity as finite, we must guarantee that the instantaneous energy and data cannot be more than the capacity; otherwise, some energy or data may be lost, resulting in suboptimal policies.

In the case of battery-operated devices, there is an initial amount of energy in the battery, and no energy is harvested. It can be proven by using Jensen's inequality that transmitting at a constant power will maximize the total transmitted data B by the deadline T [3]. An important factor that determines the performance of an EH system is the EH profile, which models the variation of the harvested energy with time [11]. The EH profile is depicted in Fig. 4. In Fig. 4a, if we have total $4E$ energy at time $t = 0$, the red dashed line is

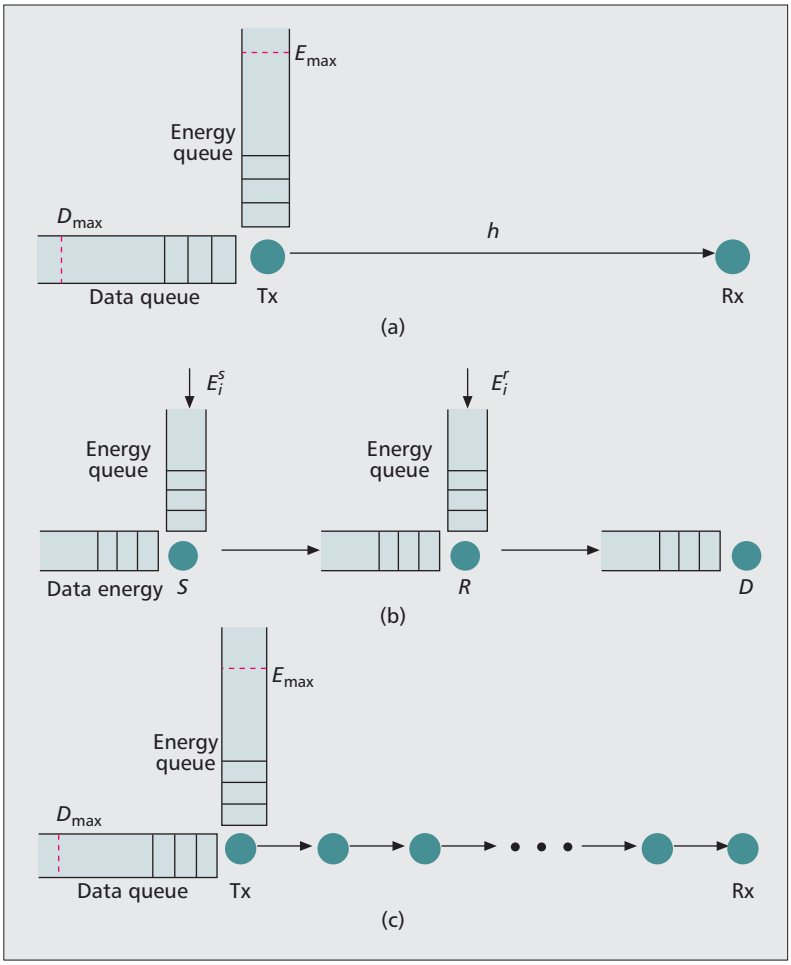


Figure 3. EH communications models: a) single-hop model; b) two-hop model; c) multi-hop model.

the best policy because it transmits at a constant power and finally uses up all the energy and the slope of the line corresponds to the transmit power. However, the red dashed line is impossible for EH communications systems, because the red dashed line has an intersection with the energy harvested line, which violates the energy causality constraint. So the feasible policies for an EH system fall under the black thick solid lines, such as the green thin solid lines.

Offline Optimal Policies

In this section we survey point-to-point EH communications optimal offline policies, which include the single-hop and two-hop models. In a single-hop model the transmitter sends data to the receiver directly over a wireless channel; in a two-hop model the transmitter uses a relay node to forward its data to the receiver.

Single-Hop Model

For a single-hop model, the goal is to minimize the transmission completion time T by when all packets are delivered to the receiver. Yang and Ulukus [4] discussed the optimal packet scheduling policy for minimizing the time T, and it has two scenarios. One scenario is when there are a total of B bits available at time $t = 0$ and no

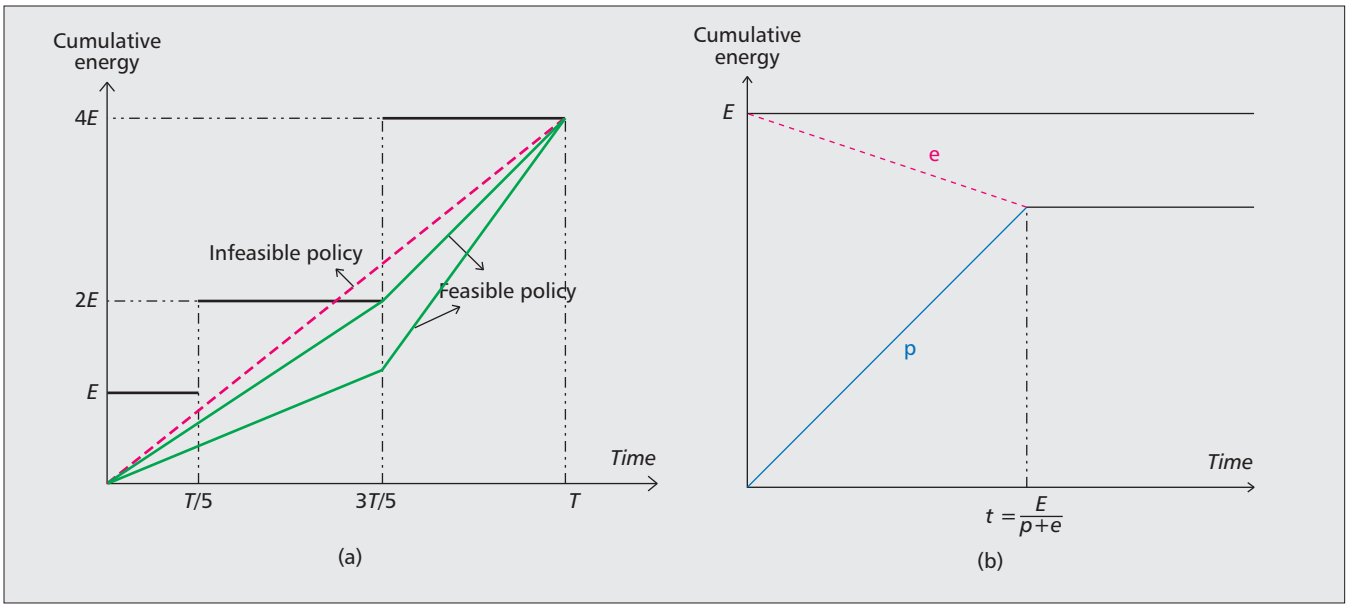


Figure 4. (a) EH profile [2] [11]; (b) EH profile with battery leakage [14].

packets arrive during the transmissions; the other scenario is when packets arrive during the transmissions. Yang and Ulukus [4] assume the energy and data buffer capacity as infinite, the CSI is perfect, and all the energy is used for transmission. There are three very important lemmas in their article. The first is that under the optimal policy, the transmit power/rate increase monotonically. The second is that the transmission power/rate remains constant between two event epochs, where an epoch is the time between two events, such as the time between the arrival of two energy packets. The third is that the energy consumed up to a specified instant is equal to the energy harvested up to that instant, meaning that the energy is used up at time T and no energy is left. As shown in Fig. 4a, the energy consumption curve must touch the energy harvesting curve at that energy harvesting instant. These three lemmas have been proved in [4] and they can help us design the optimal policy. Also, [4] gives two algorithms to minimize the transmission completion time corresponding to the two scenarios.

The model in [4] is a very optimistic single-hop model, which is not feasible in reality. Ozel and Ulukus [12] studied the optimal problem with random energy arrivals for a classical additive white Gaussian noise (AWGN) channel condition. They provide two schemes, save-and- transmit and best-effort-transmit, to achieve the optimal power management for maximizing average throughput. Meanwhile, Ozel and Tutuncuoglu [5] have proposed an algorithm called the *directional water-filling* algorithm that takes both the CSI and energy buffer capacity into account, meaning that the transmitter has a limited energy buffer capacity and communication is over a wireless fading channel. They consider two optimization problems. One is to maximize the data transmitted B by a deadline T; the other is to minimize the transmission completion time T by which the data B is completed. The *water-filling* algorithm adapts the allocation of transmission power according to the channel condition, where more power is allocated to the better channels, and poorer channels get less power in order to maximize the transmission rate. The transmitter must know the CSI to implement the water-filling algorithm. The *directional water-filling* algorithm in EH communications system is somewhat different from the water-filling algorithm because of the energy causality constraint. The power that is allocated can only transfer from left to right, which means energy transfer from past to future, because the energy we harvest in the future can not be used in the past. This is called a *right permeable tap* [5].

Although Ozel and Tutuncuoglu [5] considered CSI, they assumed the transmitter has perfect CSI, which is not practical in reality. Luo *et al.* [11] considered the optimal pilot symbol placement and power for EH communications systems. The training period and training power were optimized to obtain accurate CSI, resulting in very different solutions compared to the non-EH systems.

The solutions to maximize the data transmitted B by a deadline T and to minimize the time T by which the transmission of data B is completed are closely related, which means the two optimization problems yield identical power allocation policies, as was proven in [13]. Kaya and Aylin [13] also considered the optimal transmission policies with battery limitations, and put forward *Throughput Maximizing* and *Transmission Completion Time Minimization* algorithms to solve the two problems.

As mentioned in [2], with microelectronic systems becoming smaller and less energy demanding, the transmission energy dominates energy consumption with small transceivers. Many studies assume that all the harvested energy is only used for transmission, and the energy required for processing is not considered. Actually, the wireless systems also have processing energy cost and Orhan *et al.* [6] considered the energy consumed by both the data transmission and the processing circuitry. A *directional glue pouring* algorithm is described to compute the optimal

Model	Parameter	Scenario	Purpose	Contribution/conclusion
Yang & Ulukus [4]	Single-hop, infinite energy capacity, perfect CSI	One scenario is that all data arrived before transmission and no packets arrived during transmission, the other is that data packets may arrive during transmission	Minimize the transmission completion time	Three algorithms to minimize the transmission completion time
Ozel *et al.* [5]	Single-hop, finite energy capacity, fading wireless channel	Channel condition may change during the transmission interval	Maximize the transmitted data by a deadline and minimize the transmission completion time	A *directional water-filling* algorithm is proposed to solve the two problems
Tutuncuoglu & Yener [13]	Single-hop, finite energy capacity, AWGN channel	Sufficient amount of data is available at the beginning of data transmission	Maximize the transmitted data by a deadline and minimize the transmission completion time	*Max* algorithm and *Mini* algorithm. Maximizing the transmitted data is equivalent to minimizing the completion time
Devillers & Gunduz [6]	Single-hop, finite energy capacity, battery leakage	Single energy packet and *N* energy packet	Maximize the transmitted data by a deadline	An algorithm to maximize the transmitted data under a battery leakage condition
Orhan, Gunduz & Erkip [6]	Single-hop, finite energy capacity, fading wireless channel, processing cost	Sufficient amount of data is available at the beginning of data transmission	Maximize the transmitted data by a deadline	A *directional glue pouring* algorithm is proposed to get the optimal policy
Luo, Zhang & Letaief [8]	Two-hop, non-EH relay node, fading wireless channel	Half-duplex, sufficient amount of data is available at the source	Maximize the transmitted data by a deadline	Modified the *directional water-filling* algorithm to maximize the throughput
Gunduz & Devillers [9]	Two-hop, EH relay node, AWGN channel	Both half-duplex and full-duplex, sufficient amount of data at the source	Maximize the transmitted data by a deadline	Divide a two-hop process into two single-hop processes and use the *Max* algorithm

Table 2. Comparison of different optimal policies.

policy with processing energy cost for communication over a fading channel.

In Bertrand and Deniz [14], an EH communications system with battery leakage was proposed. The energy stored in the battery is assumed to leak at a constant finite rate, and no energy leaks when the battery is empty. The model is shown in Fig. 4b, where the red thin dashed line is the energy leaked from the battery at a constant power e, and the blue thin solid line is the energy used for transmission at a constant power p. Bertrand and Deniz [14] provided the optimal EH transmission algorithm with battery leakage.

As mentioned above, energy overflows from the battery may result in a suboptimal policy. Maria and Miquel [7] considered this problem, and a QoS constraint was put forward. A minimum data departure $B_{qos}(t)$ was defined as the smallest amount of data that must be transmitted at time T to satisfy the QoS constraint, and it can prevent battery overflows when no data is waiting for transmission.

Two-Hop Model

Many scenarios for traditional communications require two-hop or multi-hop transmission. For example, in wireless sensor networks, many nodes must use multiple hops to transmit their data to a sink node. Similarly, in EH communication systems, we require two-hop or multi-hop transmission. Due to the differences between EH systems and traditional communication systems, we need to restudy the optimal policy. Recently, many studies have focused on two-hop EH communication systems, i.e. there is a relay node between the source and destination nodes, where the source node is an EH node and the relay node may be either an EH node or battery-operated equipment.

Yaming *et al.* [8] considered an EH source and a non-EH half-duplex relay node, which means the relay node cannot harvest energy from the ambient environment and the relay node cannot receive and forward the data at the same time. The directional water-filling algorithm for single-hop EH communications system mentioned above was applied in designing the throughput maximization policy for two-hop EH communications systems. The optimal solution was obtained according to a temporary solution, and an improved directional water-filling algorithm was proposed in [8].

Orhan and Erkip [15] also considered the half-duplex optimization problem but with an

The standardization of energy harvesting communications is necessary. Either EH devices or EH communications protocols need to be standardized to assure the compatibility of different EH devices from different vendors, and for the convenience of network management.

EH relay node, and the optimal policies were studied to maximize data transmitted to the destination by a deadline T. In Deniz and Bertrand [9], both full-duplex and half-duplex communication with an EH relay node were considered. They divided a two-hop communication process into two single-hop communication processes. That is to say, with the optimal transmission schedule, the source transmits an amount of data using all its energy first, and then the relay forwards all the data received from the source to the destination by using some optimal single-hop policies as mentioned above, which have been proved in [9].

Guideline for Optimal Policy

Although both the single-hop and two-hop models are simplified scenarios, they admit analytical solutions and guidelines for designing more complicated EH communications systems. Here we conclude with some guidelines from the literature that can provide useful design insights.

For all the EH communication models:

- The optimal power management policy maintains a constant transmit power in each event epoch unless there is a new energy arrival or the channel state changes. As shown in Fig. 4a, the feasible line can only change at the corner of the EH profile.
- All energy is used up by the deadline T, i.e. the total consumed energy is equal to total harvested energy by the deadline T; otherwise it may result in suboptimal policy.
- In an optimal transmission policy, the transmit power/rate should increase monotonically in time.
- The optimal policies should follow the two causality constraints, and a battery overflow may only occur when there is no data to be transmitted.

For two-hop EH communication models:

- The source node and the relay node batteries cannot be empty simultaneously at a given time; also either source node or relay node transmits at a given time, which means they can never be silent at the same time.
- The source node transmits first and then the relay node forwards in the rest of the time, and the source and relay nodes transmit the same amount of data by the deadline in order to prevent data loss.

Table 2 compares the different optimal policies.

Research Challenges and Directions in Energy Harvesting Communications

EH communications still has some callenges to reach, and we believe that the following research directions require more attention.

Realistic Systems

The models we have studied for both offline and online energy management are very optimistic. The capacities of the battery and the data buffer capacity are assumed infinite, and the channel condition is either perfect or there is perfect CSI at the transmitter. Moreover, the energy is only used for data transmission, and other energy consuming processes in the devices are ignored, which is unrealistic in practice. When the battery capacity is much larger than the harvested energy, we can assume the battery capacity is sufficient and regard it as infinite. However, in reality wireless devices may not have large-capacity batteries due to cost constraints, and we can harvest more energy with the development of energy harvesting technology so that we have to consider the problem of the battery capacity, which is the same as the data capacity. Future work should consider battery and data capacity, processing cost, and imperfect CSI at the transmitter to model realistic EH communications environments.

Network Architecture for Energy Harvesting Communications

Existing studies have considered single-hop and two-hop EH communications systems. A realistic wireless communication network may consist of many nodes. Future studies may consider the multi-hop model shown in Fig. 3c, i.e. EH network architectures and optimization problems. Moreover, in addition to the wireless network nodes, the base stations may also use EH equipment. An interesting direction of research is to build an optimal or sustainable energy harvesting network architecture.

Algorithms and Protocols

There are a variety of protocols and routing algorithms in wireless sensor networks, and many of them consider the energy management problem. The optimal policies reviewed in this article are simple, optimistic, point-to-point algorithms. Hence, algorithms and protocols for EH communications are needed that are chosen according to the type of energy harvesting network architecture, such as a broadcast protocol for point to multi-point, and optimal routing algorithms based on multi-hop architectures. Moreover, the security of communications with EH may also be considered, which may be more complex and require more careful attention than their non-EH counterparts.

Standardization

Most technologies and solutions for EH communications are still not mature. Thus, the standardization of energy harvesting communications is necessary. Both the EH devices and the EH communications protocols need to be standardized to assure the compatibility of different EH devices from different vendors, and for the convenience of network management.

Conclusions

This article has surveyed existing EH communications technologies and theories. Our focus has been on offline optimal policies to provide for the best management of EH power, and we compared different kinds of optimization schemes. Some guidelines have been given for designing an optimal EH policy. Finally, we discussed the challenges and directions for future energy harvesting communications, which can help us design better EH communications systems.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 61372077), the Fundamental Research Program of Shenzhen City (No. JC201005250067A and No. JCYJ20120817163755061), and the Technical Research and Development Program of Shenzhen City (No. CXZZ20120615155144842).

REFERENCES

[1] R. V. Prasad *et al.*, "Reincarnation in the Ambiance: Devices and Networks with Energy Harvesting," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 1, 1st Quarter, 2014, pp. 195–213.
[2] E. Gelenbe *et al.*, "Energy Harvesting Communication Networks: Optimization and Demonstration (The E-CROPS Project)," *Green ICT (TIWDC), 2013 24th Tyrrhenian Int'l. Wksp. Digital Commun.*, Sept. 2013, pp. 1–6.
[3] D. Gunduz *et al.*, "Designing Intelligent Energy Harvesting Communication Systems," *IEEE Commun. Mag.*, vol. 52, no. 1, Jan. 2014, pp. 210–16.
[4] J. Yang and S. Ulukus, "Optimal Packet Scheduling in an Energy Harvesting Communication System," *IEEE Trans. Commun.*, vol. 60, no. 1, Jan. 2012, pp. 220–30.
[5] O. Ozel *et al.*, "Transmission with Energy Harvesting Nodes in Fading Wireless Channels: Optimal Policies," *IEEE JSAC*, vol. 29, no. 8, Sept. 2011, pp. 1732–43.
[6] O. Orhan, D. Gunduz, and E. Erkip, "Throughput Maximization for an Energy Harvesting Communication System with Processing Cost," *2012 IEEE Info. Theory Wksp. (ITW)*, Sept. 2012, pp. 84–88.
[7] M. Gregori and M. Payaro, "Energy-Efficient Transmission for Wireless Energy Harvesting Nodes," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, Mar. 2013, pp. 1244–54.
[8] Y. Luo, J. Zhang, and K. B. Letaief, "Throughput Maximization for Two-Hop Energy Harvesting Communication Systems," *2013 IEEE Int'l. Conf. Commun. (ICC)*, June 2013, pp. 4180–84.
[9] D. Gunduz and B. Devillers, "Two-Hop Communication with Energy Harvesting," *2011 4th IEEE Int'l. Wksp. Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Dec. 2011, pp. 201–04.
[10] E. S. Leland, E. M. Lai, and P. K. Wright, "A Self-Powered Wireless Sensor for Indoor Environmental Monitoring," *WNCG Conf.*, 2004.
[11] Y. Luo, J. Zhang and K. B. Letaief, "Training Optimization for Energy Harvesting Communication Systems," *2012 IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2012, pp. 3365–70.
[12] O. Ozel and S. Ulukus, "Information-Theoretic Analysis of an Energy Harvesting Communication System," *2010 IEEE 21st Int'l. Symp. Personal, Indoor and Mobile Radio Communications Wksps. (PIMRC Workshops)*, Sept. 2010, pp. 330–35.
[13] K. Tutuncuoglu and A. Yener, "Optimum Transmission Policies for Battery Limited Energy Harvesting Nodes," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, Mar. 2012, pp. 1180–89.
[14] B. Devillers and D. Gunduz, "A General Framework for the Optimization of Energy Harvesting Communication Systems with Battery Imperfections," *J. Commun. and Networks*, vol. 14, no. 2, Apr. 2012, pp. 130–39.
[15] O. Orhan and E. Erkip, "Energy Harvesting Two-Hop Networks: Optimal Policies for the Multi-Energy Arrival Case," *2012 35th IEEE Sarnoff Symposium (SARNOFF)*, May 2012, pp. 1–6.

BIOGRAPHIES

YEJUN HE [SM'09] (heyejun@ieee.org) received a Ph.D. degree in information and communication engineering from Huazhong University of Science and Technology in 2005. He has been a professor at Shenzhen University since 2011. He was a visiting professor at the University of Waterloo and Georgia Institute of Technology. His research interests include channel coding and modulation, MIMO-OFDM wireless communication, space-time processing, energy harvesting communications, and smart antennas.

XUDONG CHENG (cxd199181@126.com) received the B.S. degree from the College of Information Engineering at Shenzhen University in 2013. He is currently pursuing a M.S. degree at the College of Information Engineering at Shenzhen University. His research interests include channel modeling, especially polarized MIMO channel modeling, energy harvesting communications, smart antennas, and signal processing.

WEI PENG [M'07, SM'11] (pengwei@hust.edu.cn) received a Ph.D. degree from the University of Hong Kong in 2007. She is an associate professor at Huazhong University of Science and Technology. She was a postdoctoral fellow and an assistant professor at Tohoku University. Her research interests include transceiver design and parameter estimation for wireless communication systems with high-dimensional antenna arrays, signal processing, and compressive sensing.

GORDON L. STÜBER [F'99] (stuber@ece.gatech.edu) received the B.A.Sc. and Ph.D. degrees in electrical engineering from the University of Waterloo, Ontario, Canada, in 1982 and 1986, respectively. In 1986 he joined the School of Electrical and Computer Engineering, Georgia Institute of Technology, where he is currently a professor and holds the Joseph M. Pettit Chair in Communications.

Different types of harvesting devices can scavenge for different kinds of energy such as solar, radio-frequency, thermal energy and so on. However, the amount of available energy to be harvested varies over space and time. So it is very important to choose the appropriate energy harvesting method according to the network's energy harvesting environment.

Cutting the Last Wires for Mobile Communications by Microwave Power Transfer

Kaibin Huang and Xiangyun Zhou

ABSTRACT

The advancements in microwave power transfer (MPT) over recent decades have enabled wireless power transfer over long distances. The latest breakthroughs in wireless communication — massive MIMO, small cells, and millimeter-wave communication — make wireless networks suitable platforms for implementing MPT. This can lead to the elimination of the "last wires" connecting mobile devices to the grid for recharging, thereby tackling a huge long-standing ICT challenge. Furthermore, the seamless integration between MPT and wireless communication opens up a new area called wirelessly powered communications (WPC) where many new research directions arise, such as simultaneous information and power transfer, WPC network architectures, and techniques for safe and efficient WPC. This article provides an introduction to WPC by describing the key features of WPC, shedding light on a set of frequently asked questions, and identifying the key design issues and discussing possible solutions.

INTRODUCTION

Recent decades have seen the explosive growth of wireless communications. A sequence of breakthroughs such as multiple-input multiple-output (MIMO), capacity achieving codes, millimeter-wave communications, and small cell networks have achieved gigabit speeds for wireless access. As wireless and wired access speeds are becoming comparable, mobile devices, including smartphones, tablets, and laptop computers, have replaced desktop computers as the dominant platforms for Internet access. In contrast, advancements in battery technologies have been much slower. The resultant short battery lives require mobile devices to be periodically tethered to the grid for battery recharging. The cables for recharging are the last barrier for the devices to attain true mobility and thus are called the "last wires" in this article. As mobile services have penetrated different fields of modern society, such as banking, health care, and civil defense, interruption due to dead batteries can cause issues far more severe than mere inconvenience, such as financial loss, and threats to health and public safety. Moreover, the production of billions of non-recyclable chargers per year poses a serious environmental issue. The urgency of addressing these issues and the existence of many market opportunities have recently motivated both the industry and academia to direct huge efforts and funding toward developing technologies for wireless power transfer. Breakthroughs in such technologies will solve the grand information and communications technologies (ICT) challenge of cutting the last wires.

The idea of wireless power transfer using radio waves was first conceived and experimented by Nicola Tesla in 1899. However, the area did not pick up until the 1960s when microwave technologies rapidly advanced, opening an active research field called microwave power transfer (MPT) [1]. In particular, the availability of large-scale antenna arrays and high-power microwave generators enables beaming high power in a desired direction. Moreover, the invention of rectifying antennas (rectennas) renders energy conversion loss to a practically negligible level. Many advanced MPT systems have been designed, such as wirelessly powered airborne vehicles that require no refueling and solar powered satellites [1]. However, the enormous antenna arrays (e.g., arrays with diameters of hundreds of meters) that are instrumental for efficient MPT in such systems are impractical for everyday life applications. This issue, together with safety concerns, has delayed the commercialization of MPT. On the contrary, without such issues, non-radiative technologies for wireless power transfer, inductive coupling and resonant coupling between two coils, have been standardized and widely commercialized in mobile devices, home appliances, and electric vehicles (see, e.g., [2]). However, such technologies have the drawbacks of extremely short transfer distances (typically no more than a meter) and lack of support for mobility.

With long propagation ranges and support of mobility and multicasting, MPT appears to be a promising candidate technology for cutting the "last wires" if the two main challenges, *high propagation loss* and *safety concerns*, can be overcome. Both relying on microwaves as transmission vehicles, MPT and wireless communication

Kaibin Huang is with the University of Hong Kong.

Xiangyun Zhou is with the Australian National University.

0163-6804/15/$25.00 © 2015 IEEE

have interwound R&D histories, yielding many common techniques and theories (e.g., beamforming and propagation). The similarity allows communication techniques and network designs to be applied to tackle the MPT challenges. Specifically, sophisticated signal processing techniques for channel estimation, power control, and adaptive beamforming can be adopted to ensure safety in MPT. Moreover, the latest breakthroughs in wireless communications — small cells, transmission using large-scale antenna arrays, and millimeter-wave communications — will dramatically reduce transmission distances and enable sharp beamforming, which will suppress propagation loss and achieve high power transfer efficiencies. The availability of enabling technologies suggests that the time has come for cutting the last wires, opening up the new area of wirelessly powered communications (WPC) with many exciting new research opportunities and applications. This article introduces WPC by discussing its key features, answering a set of frequently asked questions, and identifying the key design challenges.

Key Features of Wirelessly Powered Communications

In this section, we introduce several key features of WPC, including MPT, mobile architecture supporting energy harvesting. and simultaneous wireless information and power transfer (SWIPT).

Microwave Power Transfer

Power Beamforming: Efficient MPT hinges on concentrating radiated power in the direction of a target mobile by forming a microwave beam. Such beamforming for the sole purpose of power transfer is called *power beamforming*. A beam can be formed using an antenna array (or aperture antenna). The elements of the array are arranged with separation no larger than a half wavelength so as to avoid "grating lobes" (multiple beams). Under this constraint, the beam sharpness increases with the array size (or equivalently the number of elements). Sharp beamforming and short propagation distances are the key conditions for efficient MPT. They can be achieved by the two latest corresponding wireless communication technologies: *large-scale antenna arrays* (with hundreds to thousands of antenna elements) and *small cells*, currently under extensive development and expected to be deployed in next-generation wireless networks [3].

With the array size fixed, the beam sharpness can be increased by scaling up the carrier frequency and correspondingly packing more antennas into the array. Traditional MPT without dedicated spectrum uses the carrier frequency of either 2.4 GHz or 5.8 GHz in the industrial, scientific, and medical (ISM) band [1]. However, with the rapid advancement of millimeter-wave communication, the MPT embedded in WPC can be operated in the 60 GHz bandwidth in the near future. Such high frequencies enable ultra-sharp beamforming even when the array size is small, leading to dramatically improved power transfer efficiencies.

Power Transfer Channel and Beam Efficiency: Rich scattering is typical in a wireless communication channel, and can be combined with transmit and receive antenna arrays to support multiple parallel data streams without requiring additional bandwidth. In contrast, free-space propagation is essential for power beamforming, since a scatterer can disperse a power beam and cause the transfer efficiency to drop dramatically. Thus, a power transfer channel refers to one over free space. The propagation distance ranges from the *near field*, where the distance is comparable with the transmit array dimensions, to the *far field*. The factors determining the propagation loss include:

- The apertures of the transmit and receive arrays, denoted as A_t and A_r, respectively
- The wavelength λ
- The propagation distances d as elaborated below

The end-to-end MPT efficiency is equal to the product of three efficiencies:

- DC-to-RF power conversion efficiency
- *Beam efficiency* defined as the ratio between the received and radiated powers
- RF-to-DC power conversion efficiency

The state-of-the-art microwave generators and rectennas can achieve close-to-one values (e.g., 80 percent) for the first and third efficiencies, respectively [1]. Therefore, the beam efficiency is the bottleneck for efficient MPT over long distances.

Consider a pair of transmit/receive circular aperture antennas (which can be replaced by arrays with the same apertures) facing each other over a power transfer channel. For this scenario, the beam efficiency can be accurately approximated as [1]

$$\text{Beam Efficiency} = 1 - e^{-\beta} \quad (1)$$

where β is given as

$$\beta = \frac{A_t A_r}{(\lambda d)^2}. \quad (2)$$

Note that Eq. 2 is equivalent to the Friis equation for far-field transmission. The propagation as described in Eq. 1 covers both the near and far fields. For the far field where β is small (large d), the propagation loss is approximately equal to β, thus following the Friis transmission equation. For the near field where d is small and hence β is large, the beam efficiency is close to one. The beam efficiency is plotted in Fig. 1 as a function of the ratio between the transfer distance and the receiver antenna radius, where the carrier frequency is 2.4 GHz. Next, Eq. 1 suggests the trade-offs between the MPT parameters A_r, A_t, λ, and d. In particular, for a given beam efficiency, doubling the transmit-array radius or the carrier frequency doubles the transfer distance or supports recharging of smaller (half-size) mobiles. For instance, scaling up the frequency from 2.4 GHz to 60 GHz in the millimeter band increases the power transfer distance by 25 times. This, however, requires the numbers of transmit/receive antennas to grow by the square of this factor if the aperture antennas are replaced by antenna arrays.

With the rapid advancement of millimeter-wave communication, the MPT embedded in WPC can be operated in the 60 GHz bandwidth in the near future. Such high frequencies enable ultra-sharp beamforming even when the array size is small, leading to dramatically improved power transfer efficiencies

Mobile Architecture for WPC

A traditional mobile device consists of an information transceiver powered by a rechargeable battery. For WPC, an RF energy harvester is included in the mobile device for harvesting energy from the incident microwave signal to power the transceiver as shown in Fig. 2. The design of an energy harvester is rather simple and consists of a rectifying circuit for converting the RF signal at the antenna output to DC power that is stored using a rechargeable battery or a super capacitor. The most popular and efficient design of an RF energy harvester uses a rectenna that integrates a single antenna and a rectifying circuit.

The functionality of the antennas used by the information transceiver and energy harvester is different. The array attached to the transceiver enables multi-antenna communication and array processing (e.g., receive beamforming and interference nulling). Therefore, it is desirable to have as many (small) antenna elements as possible. On the other hand, the rectenna requires capturing as much incident power as possible; hence, the rectenna design aims for the largest possible antenna aperture. This is an important trade-off for designing a WPC receiver under a form-factor constraint.

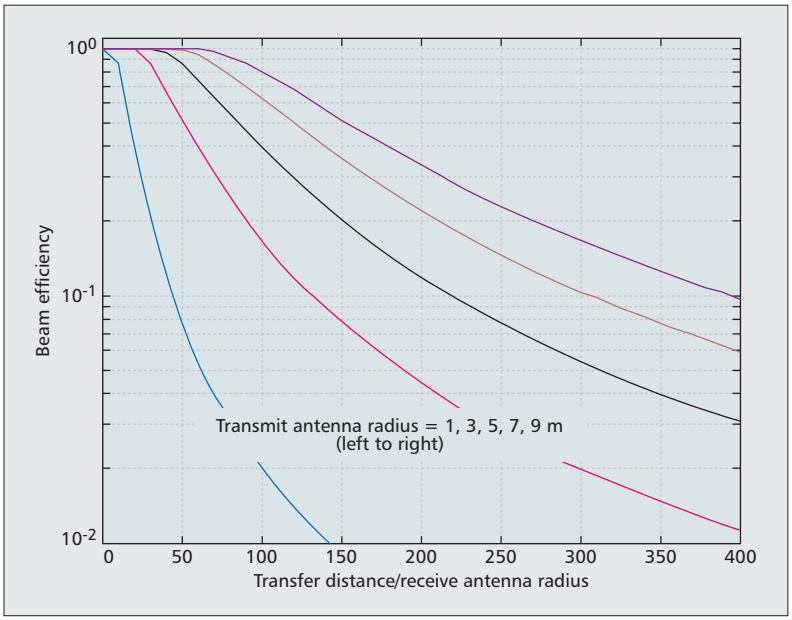


Figure 1. Beam efficiency for MPT vs. the ratio between the transfer distance and the receiver antenna radius for a carrier frequency of 2.4 GHz.

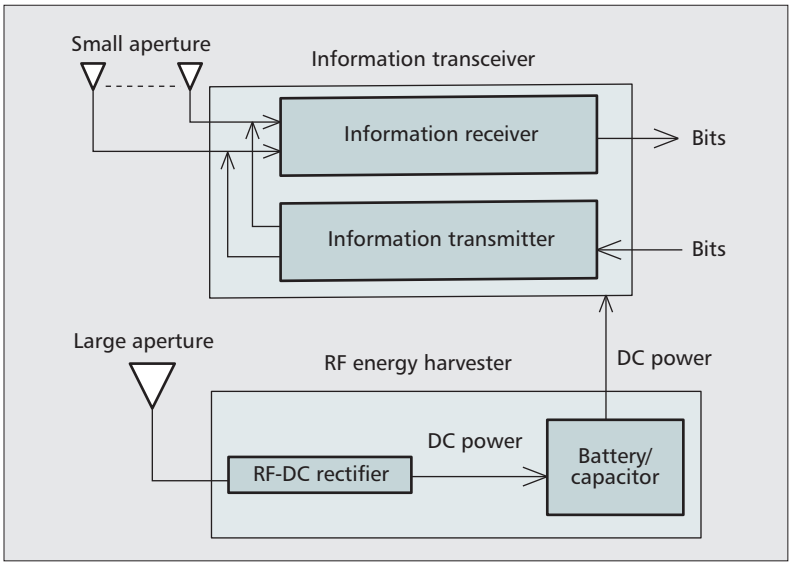


Figure 2. Mobile architecture for WPC.

Simultaneous Wireless Information and Power Transfer

Since the information transceiver and energy harvester have separate antennas and circuits, the mobile architecture supports SWIPT [4]. As illustrated in Fig. 3, there are three designs of a SWIPT system: *integrated SWIPT*, *closed-loop SWIPT*, and *decoupled SWIPT*, which are described as follows. Integrated SWIPT in Fig. 3a is the simplest design where power and information are extracted by the mobile from the same modulated microwave transmitted by a base station (BS). For this design, information *transfer* (IT) and *power transfer* (PT) distances are constrained to be equal. Closed-loop SWIPT in Fig. 3b consists of *downlink PT* and *uplink IT*. The signal power received at the BS originates from the BS radiated power, and its closed-loop propagation (downlink+uplink) incurs double attenuation [5]. Thus, closed-loop SWIPT only supports very short ranges and is unsuitable for cell edge mobiles. Last, a decoupled-SWIPT system in Fig. 3c builds on the traditional communication system to include an additional special station, called a *power beacon* (PB), dedicated for MPT to mobiles [6]. PT and IT are orthogonalized by using different frequency bands or time slots to avoid interference, which is given the name decoupled SWIPT. Unlike BSs, PBs require no backhaul links, and the resultant low cost allows dense deployment to enable efficient MPT.

Wirelessly Powered Communications: Frequently Asked Questions

Current research focuses on developing the WPC theory based on abstracted system models. Surprisingly, many frequently asked questions (FAQs) on, say, the practicality and safety of WPC, remain unanswered. In this section, an attempt is made to shed light on some FAQs.

How Far Can a Mobile Device Be Wirelessly Powered?

BSs can support communication ranges up to tens of kilometers. This can lead to the misconception that BSs/PBs can also power mobiles at comparable distances since both IT and PT use microwaves as vehicles. It is important to understand that the efficiency of PT depends on the received signal power while the reliability of IT is determined by the receive signal-to-noise ratio (SNR). Since the noise power is extremely low (e.g., –120 dBm), the received signal power for

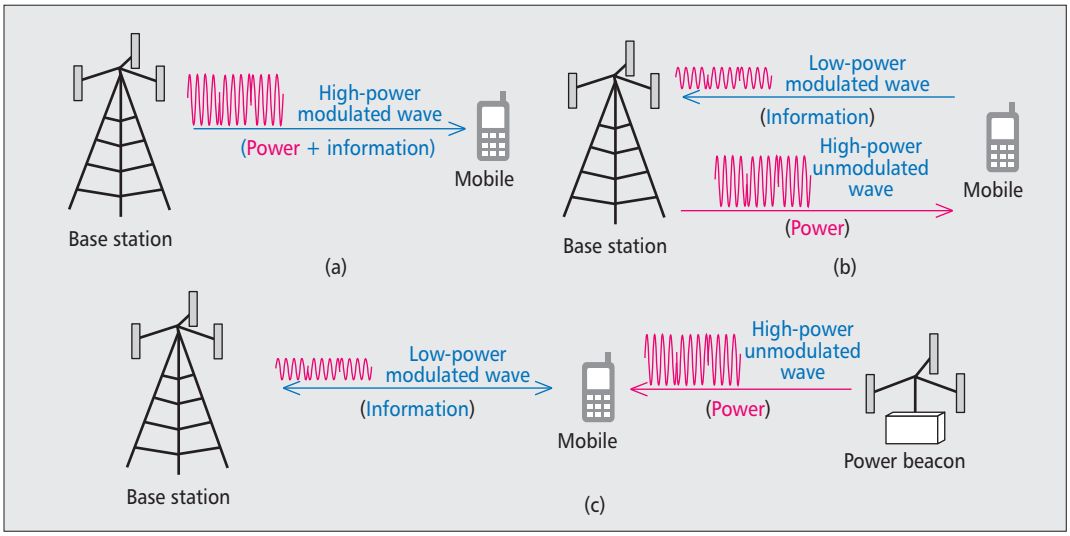


Figure 3. Three system configurations for SWIPT: a) integrated SWIPT; b) closed-loop SWIPT; c) decoupled SWIPT.

IT falls in the range of –100 dBm to –50 dBm, which is many orders of magnitude lower than the power consumption of mobile devices. Thus, one should expect drastically shorter ranges for PT than those for IT. For comparison, the typical values for received signal power and the power consumption of popular mobile devices are listed as follows.

- **Wireless signals:** –120 to –50 dBm
- **ZigBee devices or sensors:** 1 to 100 mW
- **Smartphones:** 19 mW to 1.3 W
- **Tablet computers:** 1 W to 11 W
- **Laptop computers**: 19 W to 52 W

One can see that the typical power consumption of mobile devices ranges from milliwatts for sensors or ZigBee devices to tens of watts for laptop computers, which are about 50–100 dB higher than the range of wireless signal power. In addition, the sensitivity level of a typical energy harvester is on the order of –10 dBm, and below this level little energy can be harvested.

To get a concrete answer to the current question, the PT distances can be computed using practical settings. Consider the scenario where a PB wirelessly powers a mobile device where transmit and receive antenna apertures are assumed to be circular. The power transfer ranges are computed numerically based on the beam efficiency equation, Eq. 1, and the results are plotted in Fig. 4 for different transmitted powers. The specifications for the numerical computation are summarized in the caption of Fig. 4. It is surprising that the PT ranges for small to medium devices (ZigBee/sensors, smartphones, tablets) are very similar. The reason is that a larger device can harvest more power (with a larger antenna aperture) that compensates for the increase in power consumption. One can observe that a PB transmitting tens of Watts can power sensors, smartphones, and tablets at a distance of around 10 m. Interestingly, this distance matches the smartphone recharging range of the MPT-based charging station developed by a new startup called COTA [7]. The relative short PT distances in Fig. 4 suggest that PBs should be equipped with large-scale antenna arrays and installed in an indoor environment or with a high density.

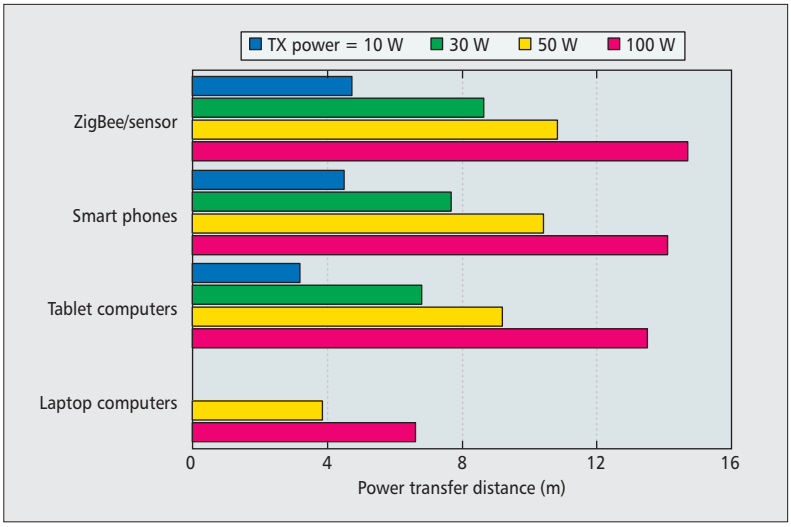


Figure 4. Power transfer ranges for different types of mobile devices. The transmitter antenna array has about 260 elements separated by half wavelength and arranged on a disk, having an aperture of 3 m in radius. The carrier frequency is 2.5 GHz. The RF-to-DC conversion efficiency is 70 percent. The power consumption and antenna radii for different types of mobile devices are assumed to be (50 mW, 1 cm) for ZigBee devices, (0.5 W, 3 cm) for smart phones, (5 W, 9 cm) for tablet computers, and (25 W, 11 cm) for laptop computers.

To summarize, under practical constraints, the MPT distance is in the range of 3–15 m for typical mobile devices depending on the radiated power. The MPT ranges for ZigBee devices, sensors, and tablets are similar and about twice those of laptop computers. Such distances are still substantially shorter than cell radii of fifth generation (5G) small cell networks, which are 50–100 m [3].

Is WPC Safe?

By sharp power beamforming, the power density can be high along the path from a PB to a target mobile. This can potentially cause health hazards

A relatively practical solution for providing network coverage for SWIPT is its implementation based on decoupled SWIPT that exploits low-cost PBs for dense deployment.

to a human body accidentally intercepting the path. According to international safety standards set by authorities such as the Federal Communications Commission (FCC) and International Commision on Non-Ionizing Radiation Protection (ICNIRP) [8, Table 3], a person should not be exposed to microwave radiation with an average power density higher than 10 W/m^2 over a half-hour time window. Note that the wavefront area of a beam grows with the propagation distance, and thus, the power density of the beam decreases accordingly. We can define the unsafe beam-interception distance (UBID) as the maximum propagation distance where the beam power density exceeds the mentioned exposure limit. Assuming that a beam contains 90 percent of the total radiated power, the beam efficiency in Eq. 1 allows us to compute the UBIDs for different practical configurations as follows:

- UBDI = **0.63 m** for radiated power P = 50 W if the transmission is **omnidirectional**.
- UBDI = **14.4 m** for P = 10 W and beamed transmission with an antenna aperture = 3 m^2.
- UBDI = **32 m** for P = 50 W and beamed transmission with an antenna aperture = 3 m^2.

For omnidirectional transmission, the UBDI is found to be less than a meter, and hence there is no safety concern in practice. On the contrary, for power beamforming, one can see from the numbers that if a person is within 15 m of the PB, he/she should be careful not to stand between the PB and the mobile for too long. The unsafe distance is higher for larger radiated power or a sharper beam (a larger antenna aperture). The PT distances in Fig. 4 are observed to be smaller than the above UBDI values, suggesting potential safety issues for WPC. However, the limits on microwave exposure as set by different authorities are average values over a long time window (0.5 h). This is more than sufficient for an intelligent WPC system to adapt its transmission power (within, e.g., milliseconds) to ensure safety.

In summary, although it is unsafe for a person to be illuminated by a power beam within the PT range for too long, intelligent beam control techniques can be designed to ensure safety in WPC (details are provided later).

Is SWIPT Practical?

A main motivation for implementing integrated and closed-loop SWIPT (Fig. 3) is that SWIPT can be realized by upgrading existing BSs (including WiFi access points) without changing the network architecture. However, constrained by the short PT ranges, BSs can provide SWIPT only to a small fraction of nearby mobiles. Supporting full network coverage for SWIPT requires the deployment of much denser BSs with extreme cell radii of 10–15 m, which can result in enormous cost and is thus impractical. On the other hand, the decoupled SWIPT design (Fig. 3) provides a more practical solution for network-wise SWIPT coverage due to the low deployment cost of backhaul-less PBs.

To provide an answer for the current question, a relatively practical solution of providing network coverage for SWIPT is its implementation based on decoupled SWIPT that exploits low-cost PBs for dense deployment.

Does MPT Interfere with Wireless Communications?

MPT can interfere with wireless communications in several direct or indirect ways. First, for decoupled SWIPT, the information-carrying signal can be deeply buried in the power-carrying signal if they are simultaneously received by a mobile due to their orders of magnitude difference in power, as mentioned earlier. Therefore, IT and PT must be sufficiently separated in frequency (e.g., a bandwidth should be reserved for the sole purpose of MPT). Second, even given sufficient separation in frequency, the power-carrying signal must be suppressed right at the antenna outputs of the mobile's information receiver rather than in the digital domain. Otherwise, the extremely strong signal can saturate the amplifier and ADC, thereby causing distortion and excessive quantization noise to the simultaneous information-carrying signal. Last, the circuit nonlinearity of microwave generators can cause strong harmonics of the carrier that interfere with wireless communications.

In summary, MPT can interfere with wireless communications in practice unless countermeasures are implemented in the system design.

Is It Possible to Power Mobile Devices by RF Energy Scavenging?

Cellular BSs and WiFi access points are ubiquitous in the urban environment. Their transmissions result in the existence of RF energy in the ambient environment. Their high frequencies (e.g., 2.4 GHz for WiFi) require resonant antennas of relatively small sizes (10–50 cm^2) for RF energy harvesting. Scavenging such energy for powering mobile devices is a green approach since it does not require installation of additional power sources. The amount of power that can be generated by energy scavenging depends on the power density. Some available measurement results are summarized in Table 1 [9, 10]. One can see that the maximum power density is on the order of 1 mW/m^2. Therefore, a mobile device (e.g., a smartphone) of a typical size smaller than 100 cm^2 can harvest peak power of tens of microwatts with the average on the order of 1 μW. This gives the following answer. RF energy scavenging is sufficient only for powering small sensors with sporadic activities. Wirelessly powering larger devices has to rely on dedicated PBs.

Wirelessly Powered Communications: Designs and Challenges

In this section, we discuss the key design considerations and challenges for WPC.

Efficient and Safe WPC

Pilot Signal Design for Retrodirective Beam Control: Power beamforming for MPT typically uses a phase array and a technique called

retrodirective beam control, which automatically steers a beam in the reverse direction of the incident pilot signal sent by the mobile by exploiting channel reciprocity. The simple procedure for retrodirective beam control is as follows:

1. The receiver transmits a pilot signal.
2. The transmitter computes the phase shift of the output of each antenna by comparing it to a local reference signal.
3. The phase shift for each antenna is conjugated and applied to the phase shifter for transmission.

An important aspect of implementing retrodirective beam control for WPC is the design of pilot sequences. They should be designed to initiate MPT for multiple mobiles at the same time. However, mobiles in different cells may transmit non-orthogonal or identical pilot sequences, resulting in pilot contamination, which lays a fundamental limit for the performance of a cellular network equipped with large-scale antenna arrays [11]. In WPC networks, pilot contamination not only degrades the IT performance but also decreases PT efficiency. To be specific, receiving multiple pilot signals can cause the retrodirective beamformer for MPT to autoreflect multiple beams toward both the intended and unintended mobiles, which reduces the beam efficiency to the former and, more importantly, causes safety threats to people in unintended directions. Tackling pilot contamination continues to be a key challenge in designing WPC networks.

In addition, the power of pilot signals and duty cycle of pilot transmission should be designed to optimize the trade-offs between PT efficiency, training overhead, and mobile energy consumption.

Safety Measures: Creating measures to ensure safe PT is a unique and important aspect of designing WPC systems. The retrodirective beamformer has a safety feature in that it can automatically de-phase a beam when it is intercepted by an object such as a human body. However, this measure does not protect people who are near a target mobile but do not intercept the beam. Thus, additional safety measures are required. For example, guard zones can be created around not only the PBs but also the mobiles using technologies such as microwave life detection [12]. Other technologies such as surveillance cameras, radar tracking, and network localization can be deployed for accurate human detection, thereby further enhancing the safety in WPC.

Efficient and Safe Power Transfer Using Multiple Coordinated Power Beacons: In a WPC network deploying dense PBs, a single mobile can be powered by multiple coordinated PBs based on the idea proposed in [13]. The PBs surrounding the target mobile form multiple incoming power beams from all directions that are coherently combined at the mobile location due to beacon coordination. PT using coordinated beacons is safe for two reasons. First, many incoming beams from different directions enable the detection of human presence at practically any arbitrary location near the target mobile, thereby overcoming the drawback of standalone PBs. Next, incident power beams from all directions have the combined effect of concentrating transmitted power at the mobile and very low power density at other locations. Moreover, multiple beams improve the chance of finding lines of sight for efficient MPT. As the combined result, MPT using coordinated beacons also improves the PT efficiency.

Spectrum	Environment	Power density (mW per m^2)
GSM (935–960 MHz)	Inner city, outdoor, on ground	$10^{-3} – 10^{-1}$
	Inner city, indoor, close to window	$10^{-2} – 10^{-1}$
GSM (1805–1880 MHz)	50 m base stations	$5 \times 10^{-3} – 5$
	200 m from base stations	$10^{-3} – 0.5$
	500 m from base stations	$5 \times 10^{-4} – 5 \times 10^{-2}$
WiFi	Within 8 m from access points	$10^{-3} – 5 \times 10^{-2}$
	12 m from access points	$10^{-4} – 5 \times 10^{-4}$

Table 1. Measured power densities of RF signals in the ambient environment.

WPC Network Architecture

A WPC network is designed to deliver two types of services, high-speed wireless access and MPT, to mobile devices, and its performance is measured by the coverage of both services. As illustrated in Fig. 5, the WPC network comprises BSs, PBs, and mobile devices. The main role of BSs is to provide network-wise wireless access coverage while also supporting SWIPT to nearby mobiles. Full MPT coverage is achieved by deploying dense PBs for supporting MPT to mobiles. Mobiles can be separated into receiving and transmitting mobiles. It is more challenging to wirelessly power transmitting mobiles since they require additional power for transmission, while both types of mobiles consume circuit power.

For designing a WPC network, one of the first challenges is to understand the required densities of BSs and PBs for providing network coverage for both wireless access and wireless power. Recently, the trade-off between these densities was quantified in [6] by modeling the WPC network using stochastic geometry and under reliability constraints on the network services. This tractable approach can be extended to design WPC networks with more complex architectures such as heterogeneous BSs/PBs. Apart from the fixed deployment of PBs, mobile PBs can be also deployed to support wider coverage with fewer PBs or more efficient MPT by shortening the transfer distances. One challenge there is to design the optimal routing for each mobile PB.

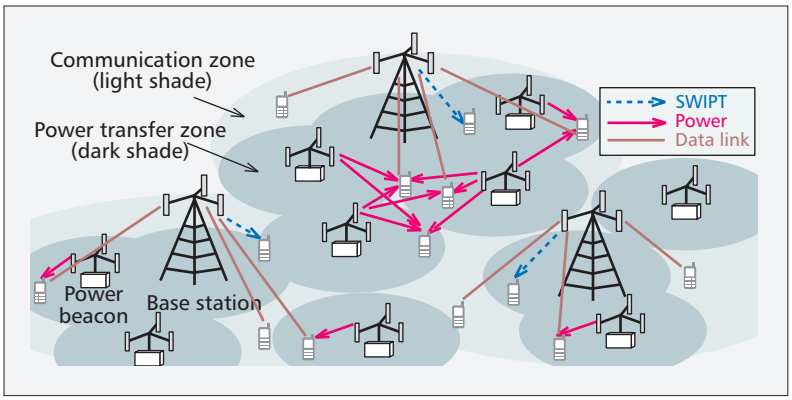


Figure 5. WPC network architecture.

WPC Protocols and Techniques

Compared to traditional wireless networks, the addition of PBs and the interaction between IT and PT enrich the WPC network architecture and the operational modes of network nodes. As a result, traditional communication protocols and techniques must be thoroughly redesigned to enable efficient WPC. Several key challenges are identified as follows, which point to promising research directions:

Cognitive WPC: The principle of cognitive radios can be applied to design cognitive WPC systems to enable seamless integration of PT and IT, and accommodate passive secondary nodes (e.g., [14]). In particular, a cognitive PB can sense the spectrum and choose a proper subset of frequency sub-channels for MPT to avoid interfering with IT and at the same time reduce power spectrum density to meet the safety requirements set by authorities.

Cooperative PB/BS clustering: Grouping PBs/BSs for cooperation enhances the PT efficiency (as discussed earlier) besides mitigating interference in IT. However, PB/BS clustering is much more complex than that for traditional multi-cell cooperation due to many new factors for consideration, including multi-user beam efficiencies, BS modes (SWIPT or IT only), and wireless signaling overhead between backhaul-less PBs.

Relay-assisted WPC: In WPC systems with relatively sparse BSs/PBs, mobiles far away from them receive double penalties: lower PT efficiencies but larger power required for uplink transmission. Thus, it is critical to address the issue of fairness in designing such systems. An alternative cost-effective approach, apart from deploying dense PBs, is to motivate mobiles to cooperate by relaying information/power for each other or deploying dedicated passive relay stations (e.g., [15]). This opens many new research issues on relay-assisted WPC ranging from signal processing methods, scheduling, and medium access control protocols to network performance.

Joint scheduling and resource allocation of PT and IT: For WPC networks of low-complexity and low-power devices, the communication protocols are often simple and predetermined. In contrast, for scenarios where the mobiles are able to handle complex algorithms, the optimal solution is to design and deploy intelligent communication and resource allocation algorithms for mobiles, PBs, and BSs that are adapted to the dynamic states of mobile energy storage, data queues, channels, and beam efficiencies.

Toward Truly Mobile Communications

Cutting the last wires of mobile devices will endow them with long desired immortality, which will bring users convenience, strengthen the reliability of widespread mobile services, and create a huge range of market opportunities. This task is far more than straightforward implementation of the MPT technology, but requires seamless integration between information and power transfers. As a result, many new research challenges arise, including designing network architectures for enabling SWIPT, achieving highly efficient and safe MPT to mobile devices, and revamping traditional communication techniques, such as cooperation, cognitive radios, and adaptive transceivers, to integrate power transfer into communication networks. This leads to a newly emerging area called wirelessly powered communications. It is through advancements in this area and relevant areas such as energy scavenging, batteries, and low-power electronics that the tens of billions of devices to be deployed in the coming decade will be free of the last wires and attain true mobility.

Acknowledgment

We thank Drs. Rahul Vaze and Salman Durrani for their comments, which have significantly improved the presentation of this article.

References

[1] N. Shinohara, *Wireless Power Transfer Via Radiowaves*, Wiley, 2014.
[2] S. Y. Hui, "Planar Wireless Charging Technology for Portable Electronic Products and Qi," *Proc. IEEE*, vol. 101, no. 6, 2013, pp. 1290–1301.
[3] J. G. Andrews *et al.*, "What will 5G Be?," *IEEE JSAC*, vol. 32, no. 6, 2014, pp. 1065–82.
[4] R. Zhang, and C. Ho, "MIMO Broadcasting for Simultaneous Wireless Information and Power Transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, May 2013, pp. 1989–2001.
[5] H. Ju and R. Zhang, "Throughput Maximization in Wireless Powered Communication Networks," *IEEE Trans. Commun.*, vol. 13, no. 1, 2014, pp. 418–28.
[6] K. Huang, and V. K. N. Lau, "Enabling Wireless Power Transfer in Cellular Networks: Architecture, Modelling and Deployment," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, Feb. 2014, pp. 902–12.
[7] B. Bodson, "COTA System Transmits Power Wirelessly at up to 30 Feet," Gizmag, Sept. 29, 2013, http://www.gizmag.com.
[8] K. R. Foster, "A World Awash with Wireless Devices: Radio-Frequency Exposure Issues," *IEEE Microwave Mag.*, vol. 14, no. 2, 2013, pp. 73–84.
[9] U. Bergqvist *et al.*, "Mobile Telecommunication Base Stations: Exposure to Electromagnetic Fields," rep. of a short-term mission within COST-244bis, 2000.
[10] H. J. Visser, A. C. Reniers, and J. A. Theeuwes, "Ambient RF Energy Scavenging: GSM and WLAN Power Density Measurements," *Proc. Euro. Microwave Conf.*, Oct. 27–31, 2008.
[11] F. Rusek *et al.*, "Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays," *IEEE Signal Proc. Mag.*, vol. 30, no. 1, 2013, pp. 40–60.
[12] Y. Chen *et al.*, "Microwave Life-Detection Systems for Searching Human Subjects Under Earthquake Rubble or Behind Barrier," *IEEE Trans. Biomedical Eng.*, vol. 47, no. 1, 2000, pp. 105–14.

[13] H. Zhai, H. K. Pan, and M. Lu, "A Practical Wireless Charging System Based on Ultra-Wideband Retro-Reflective Beamforming," *Proc. IEEE Antennas and Propagation Soc. Int'l. Symp.*, July 11–17, 2010.
[14] S. Lee, R. Zhang, and K. Huang, "Opportunistic Wireless Energy Harvesting in Cognitive Radio Networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, Sept. 2013, pp. 4788–99.
[15] A. A. Nasir *et al.*, "Relaying Protocols for Wireless Energy Harvesting and Information Processing," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, 2013, pp. 3622–36.

BIOGRAPHIES

KAIBIN HUANG [S'05, M'08, SM'13] (huangkb@eee.hku.hk) received his Ph.D. degree from the University of Texas at Austin in electrical engineering. Since January 2014, he has been an assistant professor in the Department of EEE at the University of Hong Kong. He is a Guest Editor for *IEEE JSAC*, and an Editor for *IEEE Transactions on Wireless Communications* and *IEEE Wireless Communications Letters*. He received a Best Paper Award from IEEE GLOBECOM 2006. His research interests focus on the analysis and design of wireless networks using stochastic geometry and multi-antenna techniques.

XIANGYUN ZHOU (xiangyun.zhou@anu.edu.au) is a senior lecturer at the Australian National University (ANU). He received his Ph.D. degree in telecommunications engineering from ANU in 2010. His research interests are in the fields of communication theory and wireless networks. He serves on the Editorial Boards of *IEEE Transactions on Wireless Communications* and *IEEE Communications Letters*. He was a recipient of the Best Paper Award at IEEE ICC 2011.

Energy Harvesting Small Cell Networks: Feasibility, Deployment, and Operation

Yuyi Mao, Yaming Luo, Jun Zhang, and Khaled B. Letaief

ABSTRACT

Small cell networks have attracted a great deal of attention in recent years due to their potential to meet the exponential growth of mobile data traffic, and the increasing demand for better quality of service and user experience in mobile applications. Nevertheless, wide deployment of small cell networks has not happened yet because of the complexity in the network planning and optimization, as well as the high expenditure involved in deployment and operation. In particular, it is difficult to provide grid power supply to all the small cell base stations in a cost-effective way. Moreover, a dense deployment of small cell base stations, which is needed to meet the capacity and coverage of next generation wireless networks, will increase operators' electricity bills and lead to significant carbon emission. Thus, it is crucial to exploit off-grid and green energy sources to power small cell networks, for which energy harvesting technology is a viable solution. In this article, we conduct a comprehensive study of energy harvesting small cell networks, and investigate important aspects, including a feasibility analysis, network deployment, and network operation issues. The advantages as well as unique challenges of energy harvesting small cell networks are highlighted, together with potential solutions and effective design methodologies.

INTRODUCTION

The proliferation of mobile devices, such as smartphones and tablets, is boosting the data traffic explosion in wireless ecosystems. In this context, cellular networks are faced with the challenges of providing enormous network capacity, achieving superior cellular coverage, and improving users' quality of experience. The small cell network (SCN) is a cost-effective and energy-efficient network paradigm to tackle these challenges. In SCNs, densely deployed small cell base stations (SCBSs), including micro, pico, and femtocells, bring the spatial reuse of radio resources to a new level, which will then help improve the area spectral efficiency and user experience. Besides this, low-cost and low-power SCBSs can easily be installed without costly cell site acquisition, and their self-organization manner further helps save operating expenditures [1, 2].

However, as SCBSs are densely and irregularly located, some of them may be inaccessible to the power grid. Moreover, the network power consumption of the SCNs will be high despite the small power consumption of a single SCBS, which will produce a significant amount of carbon emissions. As a result, it is desirable to exploit off-grid and green energy sources to power the SCNs. Energy harvesting (EH) technology is a viable and promising solution, which can harvest ambient renewable energy (e.g., solar and wind energy) to power SCBSs [3]. It is estimated that applying EH techniques to SCNs can achieve a 20 percent CO_2 reduction in the information and communication technology (ICT) industry [4].

Communication networks with EH capability have been extensively studied in recent years, from point-to-point systems [5, 6], two-hop systems [7], multi-user systems [8], to EH heterogeneous networks [9]. However, so far, there has been no systematic study on how to effectively utilize the EH techniques in SCNs, that is, how to power SCBSs by EH, how to deploy EH-SCBSs, and how to optimize network operations for EH-SCNs. The goal of this article is to provide a comprehensive study for EH-SCNs. Specifically, a feasibility analysis of EH-SCNs is conducted first, and then network deployment issues are addressed from the basic trade-offs to practical deployment considerations. Over deployed EH-SCNs, the challenges and design methodologies for network operation are elaborated.

POWERING SMALL CELL NETWORKS BY ENERGY HARVESTING: A FEASIBILITY ANALYSIS

In this section, we investigate the feasibility of powering SCNs by renewable energy sources. We first highlight the main differences between the energy consumption models for the macro base station (BS) and SCBSs, following which the potential of different EH techniques to power SCBSs is discussed. In particular, it is revealed that a hybrid solar-wind energy harvester is an ideal candidate to enable EH-SCNs.

The authors are with Hong Kong University of Science and Technology.

This work is supported by the Hong Kong Research Grant Council under Grant No. 610212.

0163-6804/15/$25.00 © 2015 IEEE

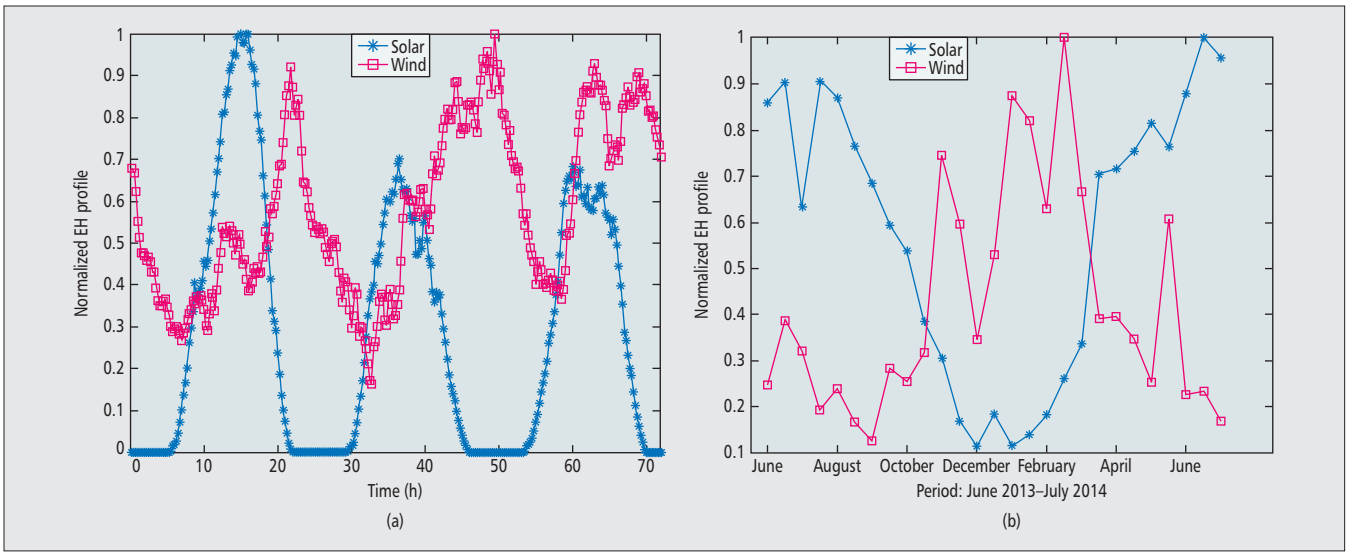


Figure 1. Normalized solar and wind energy profiles: a) short time horizon; b) long time horizon.

The Feasibility of EH-SCNs

The energy consumption models of SCBSs are fundamentally different from the macro BS, which are specified as follows:

- The communication distances from the SCBSs to mobile users will be significantly reduced compared to macro BSs, as SCBSs will be densely deployed [2]. Consequently, the transmit power of SCBSs will be greatly reduced. For example, the maximum transmit power of a typical femto BS is 17 dBm compared to 43 dBm for a macro BS [10].
- The baseband processing in SCBSs is much simpler than in the macro BS since several key operations are eliminated, such as digital predistortion.
- Cooling in the macro BS accounts for around 10 percent of the total BS power consumption [10], while SCBSs can be cooled by natural air circulation.

Given the above discussion, the power consumption of an SCBS is orders of magnitude smaller than the typical macro BS. Specifically, the power consumption of a typical macro BS is 225 W/transceiver. In contrast, it is 72.3 W/7.3 W/5.2 W for a micro/pico/femto BS [10]. Thus, it will be more feasible to power SCBSs by EH. On the other hand, as SCBSs need to be densely deployed, their total energy consumption may still be high. Thus, powering SCBSs by renewable energy sources is also motivated by environmental concerns.

To check the feasibility of powering SCNs via EH, we summarize the main energy sources of EH techniques in Table 1. We see that most of the existing applications of EH techniques are limited to low-power electronic devices, mainly due to the low EH rates. Considering the typical power consumption of an SCBS (> 5 W), only a few of the EH sources are applicable, of which solar and wind energy are the two most promising ones due to the following reasons:

- Sufficient harvested energy can be guaranteed with either a solar or wind energy harvester. For example, 100 W electric power can be generated by either a 121 cm × 53.6 cm solar panel under rated sunlight radiation, or by a rotor with a 1 m diameter under an 8 m/s wind speed.
- Such energy harvesters are cost-effective, due to their decades of lifetime and almost negligible maintenance expenditure. The main cost of solar/wind energy harvesters originates from the deployment stage, which has been decreasing dramatically in recent years.
- Many industrial companies are actively participating in developing solar and wind energy harvesters, such as Suntech, First Solar, Sunpower, and Trina Solar for solar energy, and GE Energy, Vestas, Siemens Wind Power, and Goldwind for wind energy.

Although solar and wind harvesters enjoy high harvesting rates and low cost, the time variation of the energy source poses challenges to solar/wind power generation. Fortunately, it turns out that solar and wind are a good complement to each other. On daily timescales, high pressure areas tend to bring clear skies and low surface winds, which is favorable for solar harvesters, whereas low pressure areas tend to be windier and cloudier, and thus are good for wind harvesters. On seasonal timescales, solar energy peaks in summer, whereas in many areas wind energy is lower in summer and higher in winter. We demonstrate such a complementary effect in the following case study.

A case study: We use real solar and wind power generation data by the Elia Group in Belgium.[1] The normalized energy profiles on daily timescales are shown first. Based on the measured data from 0:00 am, 15 June to 0:00 am, 17 June, 2014, the average solar/wind power is shown in Fig. 1a, where the EH rates are sampled (averaged) every 15 min. We see that the peak of solar power always coincides with the valley of wind power, and vice versa. Next we show the energy profiles on seasonal timescales. Based on the data from 0:00 am, 17 May 2013 to 0:00 am, 02 July 2014, the average solar/wind

[1] Elia, Power Generation, http://www.elia.be/en/grid-data/power-generation.

Energy sources	Characteristics	Implementation techniques	Amount of harvested energy	Typical applications
Solar [11]	Uncontrollable, predictable	Photovoltaic cells	15 mW/cm^2	Wireless sensor, household appliances
Wind [11]	Uncontrollable, unpredictable	Anemometer	100 W (rotor diameter 1 m, wind speed 8 m/s)	Wireless sensor, household appliances
Environmental vibration [11]	Uncontrollable, unpredictable	Electromagnetic induction	0.2 mW/cm^2	Wireless sensor, consumer electronic
Human motion [11]	Controllable, predictable	Piezoelectric	Finger motion: 2.1 mW; footfalls: 5 W	On-body monitoring, portable devices
Thermal [12]	Uncontrollable, unpredictable	Thermopiles	~ 40 mW	Wireless sensor
Ambient RF signal [13]	Uncontrollable, unpredictable	Rectification and filtering	< 0.2 mW	RFID, low-power device
Biomass [14]	Controllable, predictable	Microbial fuel cells	153 mW/m^2	Underwater sensor

Table 1. Existing energy harvesting techniques.

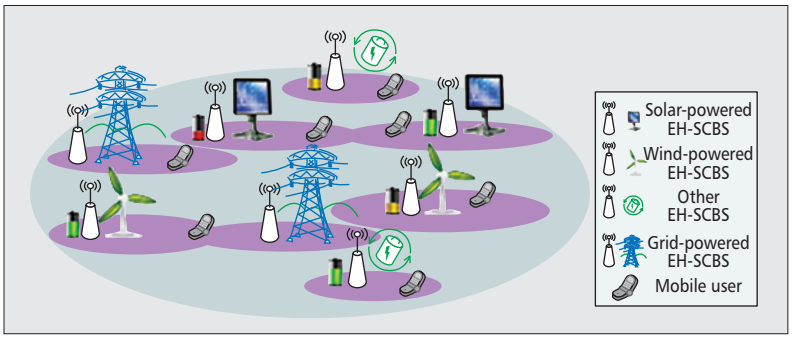


Figure 2. A sample EH-SCN.

power, averaged every 15 days, is shown in Fig. 1b. We see that the solar power achieves its peak during June-August, while the wind power reaches its bottom. An opposite trend is observed during December–February.

This case study reveals that a combination of solar and wind energy is a good candidate for the energy source of SCBSs. Actually, BSs powered by hybrid solar-wind energy have already drawn great attention from the industry. For example, the Turkish mobile operator Avea and the leading equipment vendor Huawei have shown great interest in such BSs. In particular, Wind-Fi, a renewable energy BS designed by the Centre for White Space Communications, enables wireless networks to operate entirely on solar and wind energy, and achieves 99.98 percent reliability.[2]

The Prospect of EH-SCNs

The above discussion demonstrates that EH technology, particularly solar and wind harvesters, is a viable green energy solution for SCNs. Therefore, EH-SCNs in the near future may consist of solar-powered SCBSs, wind-powered SCBSs, hybrid solar-wind powered SCBSs, EH-SCBSs powered by other energy sources, and conventional grid-powered SCBSs as well. EH-SCNs will not only reduce the deployment cost and energy bills of the operators, but will also be more environmentally friendly and thus can enable sustainable growth of wireless networks. An example of such a network is shown in Fig. 2.

Powering SCNs by EH sources will bring new design challenges. The network coverage and operating reliability will be difficult to guarantee since harvested energy varies daily and seasonally. Adjustments in the communication protocols and transmission strategies will be needed. In the following part of this article, special attention will be paid to addressing the upcoming design issues in network deployment as well as in network operation.

Network Deployment of EH-SCNs

Network deployment is the first step toward designing an effective EH-SCN. A key question to ask is how many EH-SCBSs are needed. Thus, in this section we investigate the impact of SCBS density on network performance and cost, which reveals some interesting trade-offs in EH-SCNs. Other deployment issues are then discussed.

Basic Trade-offs

The density of an EH-SCN will determine its performance as well as the network cost. Increasing EH-SCBS density can improve the coverage and throughput, but it will also increase the deployment cost. On the other hand, with a low density of EH-SCBSs, more grid-powered SCBSs will be needed to maintain the coverage, which will consume more non-renewable energy and increase the energy bills. In the following, we provide simulation results to illustrate these

[2] Centre for White Space Communications, Wind-Fi: Renewable-Energy Wireless Basestations, http://www.wireless-whitespace.org/projects/wind-fi-renewable-energy-basestation.aspx.

trade-offs. The outage probability is adopted as the performance metric, which is the portion of users that cannot be successfully served.

Trade-off between Outage Probability and EH-SCBS Density — We first consider providing network coverage only with off-grid EH-SCBSs (i.e., without any support of the grid). We assume each user is associated with its nearest SCBS, and we ignore co-channel interference as the main purpose is to guarantee network coverage. For each SCBS, energy arrives intermittently with an average EH rate P_{EH}. In each time slot, part of the harvested energy will be used to serve its users, while the remaining part will be stored in a battery with capacity C_B. To investigate the impact of C_B, we consider two extreme cases: $C_B = 0$ and $C_B = \infty$. The transmit power for each user is determined to satisfy its receive signal-to-noise ratio (SNR) requirement γ_{th}. We ignore the circuit power consumption of the SCBSs unless otherwise mentioned. Each SCBS will serve all of its associated users if the available energy is sufficient; otherwise, it maximizes the number of served users. The SCBS and user densities are denoted by λ_{BS} and λ_u, respectively. The trade-off between the outage probability p_{out} and λ_{BS} is shown in Fig. 3a. Key observations can be drawn:

- The outage probability decreases with λ_{BS}, but the decreasing rate reduces as λ_{BS} increases further. Specifically, to achieve $p_{out} = 10$ percent in this EH-SCN with $\lambda_u = 10^{-3}$ m^{-2}, we can deploy SCBSs with density $\lambda_{BS} \geq 1.7 \times 10^{-4}$ m^{-2} if each SCBS has $P_{EH} = 20$ mW when supported by a battery with large enough capacity, or $\lambda_{BS} \geq 2.1 \times 10^{-4}$ m^{-2} if each has no battery. Both are within the typical network density range of SCNs [15].
- The battery capacity has little influence on p_{out} when λ_{BS} is either very small or very large. When λ_{BS} is very small, as the harvested energy is insufficient almost all the time, the energy will be exhausted immediately after it arrives. On the other hand, if λ_{BS} is very large, the current harvested energy will be more than enough, and there is no need to consume the energy in the battery.
- Increasing either P_{EH} or λ_{BS} will reduce p_{out}. Interestingly, increasing λ_{BS} brings more performance improvement, which can be explained intuitively. Doubling λ_{BS} not only doubles the available energy in the whole network, but also reduces the transmission distances on average.

Trade-off between Grid Power Consumption and EH-SCBS Density — In this part, we assume that all the SCBSs are on-grid SCBSs, that is, the power grid is retained as the backup energy source for each SCBS. With a stable power supply, it is easy to guarantee coverage, and thus the focus is on the impact of EH-SCBS density on grid power consumption. At each SCBS, the harvested energy will be exhausted first, and the grid power will be used only when necessary. The trade-off between grid power consumption P_G and λ_{BS} is shown in Fig. 3b. Key observations can be drawn:

- The grid power consumption P_G decreases with the SCBS density λ_{BS}, and the optimal EH-SCBS density can be chosen to minimize the network deployment and operating expenditures. For example, assume the

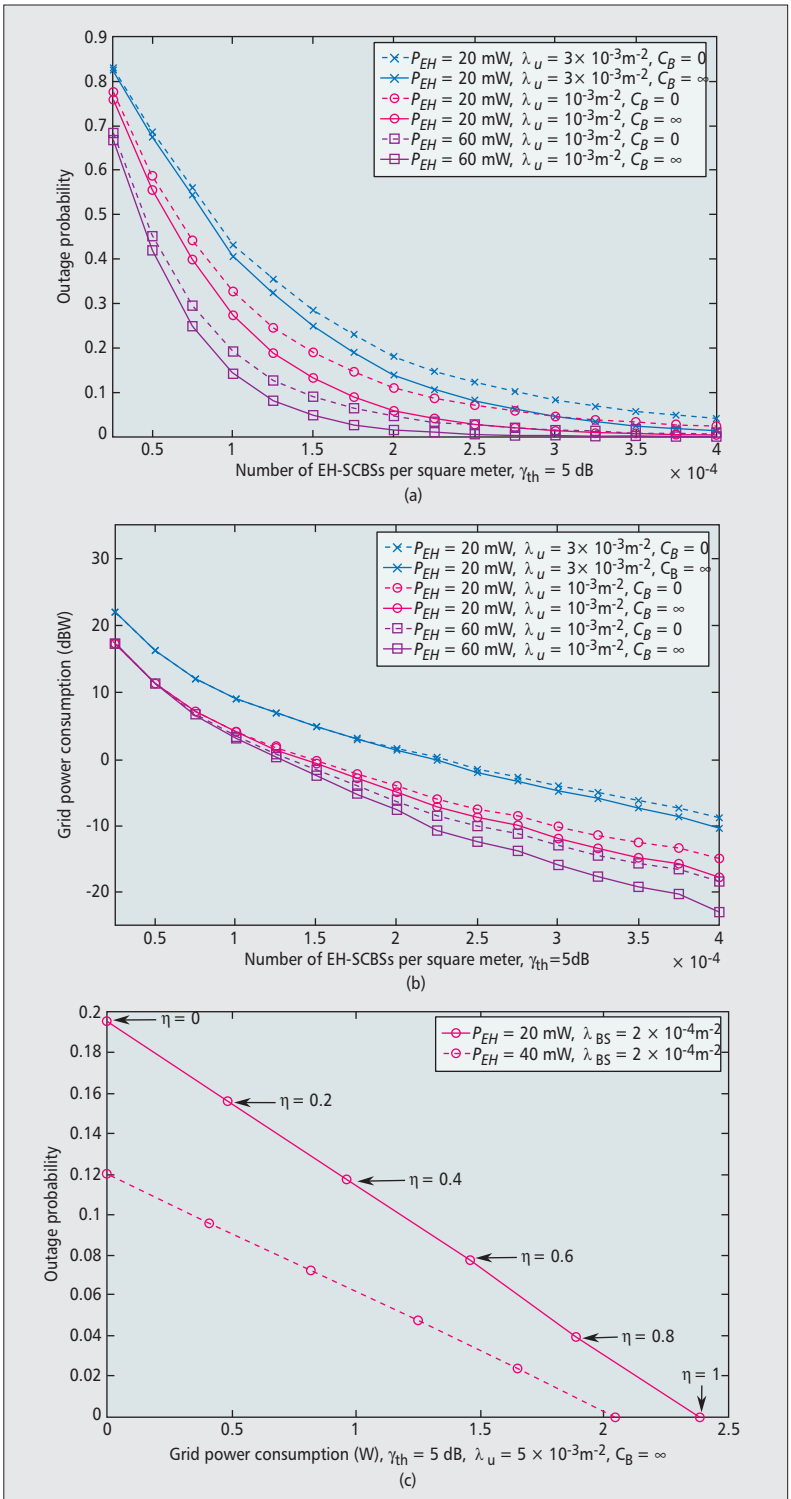


Figure 3. Basic trade-offs in the EH-SCNs: a) outage probability vs. SCBS density; b) grid power consumption vs. SCBS density; c) outage probability vs. grid power consumption.

[3] This value takes the circuit power consumption of an EH-SCBS into account.

Channel information is important for wireless communications. However, due to the limited available energy in EH-SCNs, the energy spent on channel estimation and data transmission should be balanced. It is also critical to decide when to perform channel training based on the time-varying EH profile.

electricity price is \$0.1971/kWh, while each SCBS costs \$135, of which \$35 is for the 10 W photovoltaic cell,[3] and \$100 is for the BS equipment. Normally, the lifetime of EH-SCNs is around 10 years. Then if $\lambda_u = 3 \times 10^{-3}$ m^{-2} and $P_{EH} = 20$ mW, we can find the optimal SCBS density as 7.5×10^{-5} m^{-2}.

- To reduce P_G, increasing λ_{BS} is more effective than increasing P_{EH}. The impact of C_B on P_G is negligible when λ_{BS} is extremely small or extremely large.

Trade-off between Outage Probability and Grid Power Consumption — Compared to off-grid SCBSs, on-grid SCBSs make it easy to guarantee coverage with a stable power supply, but they are more difficult to deploy due to the grid power supply, and they will also increase the non-renewable energy consumption. In this part, we consider an SCN with both off-grid and on-grid SCBSs, while the total density is fixed. By varying the density of off-grid SCBSs, we can achieve different trade-offs between the outage probability and the grid power consumption. With the ratio of on-grid SCBSs, denoted as η, increasing from 0 to 1, the outage probability will decrease, while the grid power consumption will increase. The relationship between p_{out} and P_G is shown in Fig. 3c, from which we can make the following observations:

- Changing η can adjust the trade-off between the outage probability and grid power consumption. For example, when $\lambda_{BS} = 2 \times 10^{-4}$ m^{-2}, $P_{EH} = 40$ mW, we can achieve the outage probability $p_{out} = 0$ with $P_G = 2.38$ W by setting $\eta = 1$. Alternatively, we can achieve $p_{out} = 8$ percent with $P_G = 1.46$ W by setting $\eta = 0.6$. That is, replacing 40 percent of the on-grid SCBSs with off-grid SCBSs, we can reduce the grid power consumption by ~40 percent with a slight performance degradation.
- The outage probability scales linearly with P_G, as both p_{out} and P_G scale linearly with η due to the independent and identical settings for different SCBSs, such as their locations and EH rates.

Deployment Issues

The previous discussions on the three basic trade-offs in EH-SCNs provide us with the following deployment guidelines:

- Satisfactory coverage can be guaranteed in EH-SCNs with a reasonable network density. By carefully determining the network density, we can not only balance between network performance and deployment cost, but also achieve a trade-off between performance and grid power consumption.
- To improve network performance or save grid power consumption, it is more effective to increase the SCBS density than to increase the EH rate of each SCBS (e.g., by deploying a larger solar panel).
- When the EH-SCBS density is extremely small or extremely large, battery capacity has little influence on network performance or grid power consumption.

So far, the considered scenarios are rather simplified. For example, co-channel interference between users is ignored. The BS power consumption model is also ideal, as only the transmit power is considered, while in practice, for a femto BS, when the transmit power is 25 mW, around 5.2 W is consumed by the whole BS [10]. Therefore, a more detailed investigation is needed. One useful tool for network deployment is the spatial network model, as adopted in [9, 15]. Such a network model can help to provide analytical results for performance evaluation, which may then provide guidelines for network deployment and avoid time consuming simulations. Moreover, we need to take realistic physical and social factors into consideration. Generally, BS locations can be adapted to the spatial traffic profile, that is, more SCBSs should be deployed in the traffic hotspots to meet the high communication demand. Moreover, for a given location, the EH source should be chosen according to the ambient energy availability and the economic costs. For instance, a wind-powered SCBS is preferred to a solar-powered SCBS at the seashore due to the abundant amount and installation convenience of wind energy.

Network Operation of EH-SCNs

In the previous section, we investigate the deployment issues in EH-SCNs with simplified network operations. In a practically deployed EH-SCN, network operation should be carefully designed to optimize network performance. Due to the spatial and temporal variations of EH conditions, the network operation strategies for conventional grid-powered SCNs are no longer applicable to EH-SCNs. In this section, with the joint power assignment and cell association problem as an example, we illustrate the unique design challenges and some promising methodologies for EH-SCN network operations.

Power Assignment and Cell Association in EH-SCNs

Introducing EH-SCBSs will bring unique challenges for the SCBS power assignment and cell association problem, that is, determining with which SCBS each mobile should be associated and at which power level each SCBS should choose to transmit the signal. In particular, the following aspects should be considered:

- Incorporating the temporal and spatial variation of the available energy. In conventional SCNs, fixed cell association is normally adopted (e.g., the users are associated with their nearest SCBSs) [15]. In SCNs, the temporal and spatial variation of the available EH source makes fixed association inapplicable, and a given user will need to be associated with different SCBSs during different periods. Thus, the design of cell association policies should balance the energy utilization of different SCBSs.
- Incorporating the coupling among different users/SCBSs. For a given SCBS, if the transmit power allocated to serve one user is too high, it may easily exhaust its available energy and not be able to serve other users. Thus, some of its users need to be offload-

ed to other SCBSs, the available energy of which may be quickly depleted. This coupling among users/SCBSs renders power assignment of each SCBS and cell association quite complicated in EH-SCNs.

To demonstrate these aspects in more detail, we next consider two specific design problems.

Performance Optimization for Off-Grid EH-SCNs — We first consider an EH-SCN with M off-grid EH SCBSs and K mobile users, where each user is served by one SCBS in each time slot. To provide satisfactory performance to these users, an efficient joint cell association and power assignment policy should be developed. For simplicity, we assume a constant EH rate for each SCBS, but different SCBSs may have different EH rates. The design objective is to maximize the minimum average SNR among the K users; thus, fair performance can be provided. This problem can be shown to be NP-hard. To reduce the computational complexity, we propose a low-complexity suboptimal solution based on the threshold-bisection algorithm proposed in [16]. The proposed method will not only balance the energy usage at different SCBSs, but also take the future available energy at each SCBS into consideration (i.e., it considers both spatial and temporal energy variation).

To show the effectiveness of the proposed method, we introduce a performance upper bound and two baseline policies. The upper bound is obtained by allowing multiple SCBSs to jointly serve all the users using distributed beamforming, denoted as "distributed BF". The first baseline policy adopts distance-based cell association, where each user is served by its nearest SCBS. The second one adopts SNR-based cell association, where each user is associated with the SCBS that provides the highest receive SNR with the available energy. The performances of different policies are shown in Fig. 4. Key observations can be drawn:

- The proposed solution greatly outperforms both baseline policies and achieves performance close to the upper bound.
- The distance-based policy suffers performance loss as it neglects the spatial variation of available energy at different SCBSs. Therefore, conventional cell association strategies cannot be directly adopted in EH-SCNs.
- The SNR-based policy performs better than the distance-based policy, as it utilizes information on both the distance and the current energy state. However, it still suffers performance degradation as it makes decisions based only on the current system state and neglects the coupling in different transmission blocks as well as among different users.

In summary, the cell association policies should be redesigned for EH-SCNs, and important aspects should be taken into consideration, including the temporal and spatial variation of the energy and the coupling among SCBSs/mobile users. It is difficult to obtain optimal solutions, but effective suboptimal solutions can be developed by considering the unique properties of EH-SCNs.

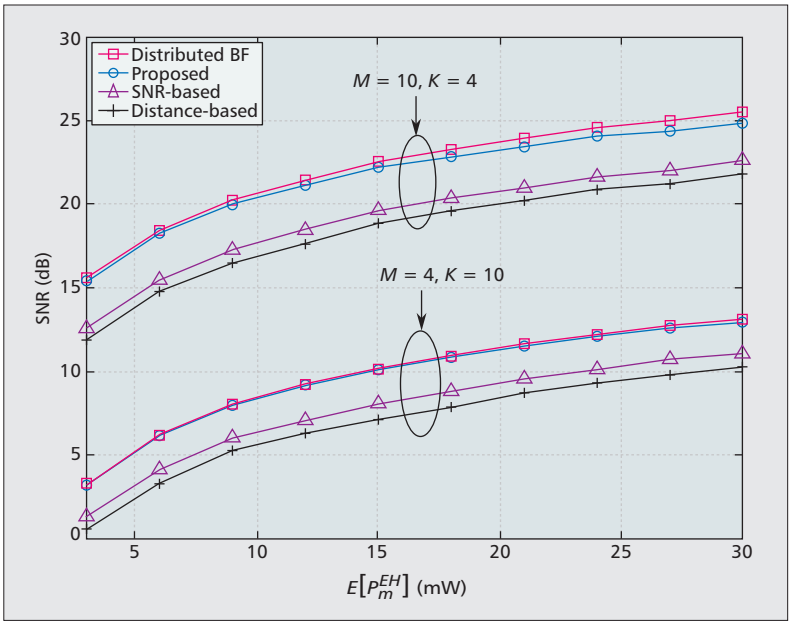


Figure 4. Comparisons of different power assignment and cell association policies.

Grid Power Minimization for On-Grid EH-SCBSs — In this part, we consider an SCN consisting of both EH-SCBSs and grid-powered SCBSs. The design objective is to minimize the power consumption of grid-powered SCBSs by adaptive cell association. For simplicity, we assume that only one of the M SCBSs is powered by the grid, and focus on a single-user case. All other assumptions are the same as in the previous design problem. For this problem, we have obtained the following two optimal transmission strategies:

- **The save-transmit strategy:** For this solution, there is a critical time slot, before which the user is served by the grid-powered SCBS, and after which the EH-SCBSs take turns to serve the user. This strategy reflects an innate characteristic of EH systems: with a given number of time slots to use EH-SCBSs, deferring these time slots will not deteriorate the performance. However, non-causal EH information is required to obtain the critical time slot index.
- **The greedy-transmit strategy:** For this solution, in each time slot, if possible, the user is served by one of the EH-SCBSs that has enough energy. Otherwise, the user is served by the grid-powered SCBS. This solution is extremely simple, as the decision in each time slot only depends on the current energy state of each EH-SCBS, irrespective of the future EH information.

We illustrate these two optimal transmission strategies in Fig. 5 with $M = 2$. The upper part of Fig. 5 shows the save-transmit strategy; that is, the user is served by the grid-powered SCBS from time slot 1 to 4, and then by the EH-SCBS from time slot 5 to 10 (the critical time slot index is 4). The lower part of Fig. 5 shows the greedy-transmit solution, where the user is served by the EH-SCBS as long as it has accumulated enough energy to support the transmit

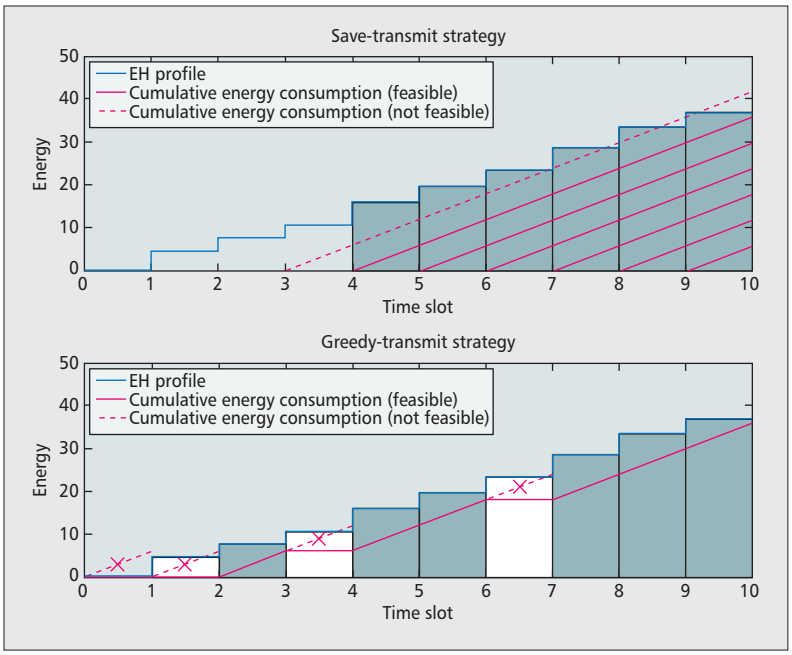


Figure 5. Illustration of the save-transmit and greedy-transmit strategies (the user is served by the EH-SCBS in the shaded time slots, and the slope of the curve represents the transmit power).

power. In both solutions, the user is served by the grid-powered SCBS in four time slots; that is, both consume the same amount of grid power.

These two solutions are typical transmission strategies for EH communication systems. With low complexity and simple operation, surprisingly, they are optimal for the considered problem. Although the optimality may be lost in more general cases, they can still serve as heuristic methods and provide low-complexity suboptimal solutions.

Other Design Problems

The above discussions of cell association shed light on the operation issues in EH-SCNs, and provide some potential solutions. In general, the design problems in EH-SCNs will be more challenging than in conventional SCNs, and their unique characteristics, especially the impact of energy profiles, should be taken into consideration. Such policies as save-transmit and greedy-transmit can help develop efficient transmission policies, which in certain cases can be shown to be optimal. There are many other design problems to be addressed, including, but not limited to, the following:

- **Sleep control:** When taking the circuit power of an SCBS into consideration, the energy efficiency of EH-SCNs can be effectively improved by sleep control, i.e., to adaptively switch off some SCBSs.
- **User scheduling:** When an EH-SCBS is serving multiple users, how to schedule these users is vital for the network performance. As the available energy of each EH-SCBS accumulates over time, probably the users with better channel conditions should be served earlier, while the optimal policy requires further investigation.
- **Channel estimation:** Channel information is important for wireless communications. However, due to the limited available energy in EH-SCNs, the energy spent on channel estimation and data transmission should be balanced. Moreover, it is also critical to decide when to perform channel training based on the time-varying EH profile.

Conclusions

In this article, we have conducted a comprehensive study of EH-SCNs, including feasibility analysis, network deployment investigation, and network operation design. Among potential EH sources, we have found that the combination of solar and wind energy is a good candidate to power SCNs. To provide network deployment guidelines for network deployment, three basic trade-offs between the network performance, EH-SCBS density, and grid power consumption are investigated. For a given deployed EH-SCN, in order to optimize network performance, special attention has been paid to the network operation designs in EH-SCNs. Throughout the article, distinctive challenges of EH-SCNs are highlighted, and novel design methodologies are proposed. Open research problems are identified that deserve unremitting efforts to promote faster, greener, and more flexible EH-SCNs.

References

[1] V. Chandrasekha and J. G. Andrews, "Femtocell Networks: A Survey," *IEEE Commun. Mag.*, vol. 46, no. 9, Sept. 2008, pp. 59–67.

[2] G. Bartoli *et al.*, "Beamforming for Small Cell Deployment in LTE-Advanced and Beyond," *IEEE Wireless Commun.*, vol. 21, no. 2, Feb. 2014, pp. 50–56.

[3] T. Han and N. Ansari, "Green-Energy Aware and Latency Aware User Association in Heterogeneous Cellular Networks," *Proc. IEEE GLOBECOM*, Atlanta, GA, Dec. 2013.

[4] G. Piro *et al.*, "Hetnets Powered by Renewable Energy Sources: Sustainable Next Generation Cellular Networks," *IEEE Internet Comp.*, vol. 17, no. 1, Jan. 2013, pp. 32–39.

[5] D. Gunduz *et al.*, "Designing Intelligent Energy Harvesting Communication Systems," *IEEE Commun. Mag.*, vol. 52, no. 1, Jan. 2014, pp. 210–16.

[6] C. K. Ho and R. Zhang, "Optimal Energy Allocation for Wireless Communications with Energy Harvesting Constraints," *IEEE Trans. Signal Processing*, vol. 60, no. 9, Sept. 2012, pp. 4808–18.

[7] Y. Luo, J. Zhang, and K. B. Letaief, "Optimal Scheduling and Power Allocation for Two-Hop Energy Harvesting Communication Systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, Sept. 2013, pp. 4729–41.

[8] J. Yang, O. Ozel, and S. Ulukus, "Broadcasting with an Energy Harvesting Rechargeable Transmitter," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, Feb. 2012, pp. 571–83.

[9] H. S. Dhillon *et al.*, "Fundamentals of Heterogeneous Cellular Networks with Energy Harvesting," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, May 2014, pp. 2782–97.

[10] G. Auer *et al.*, "How Much Energy Is Needed to Run A Wireless Network?," *IEEE Wireless Commun.*, vol. 18, no. 5, Oct. 2011, pp. 40–49.

[11] S. Sudevalayam and P. Kulkarni, "Energy Harvesting Sensor Nodes: Survey and Implications," *IEEE Commun. Surveys and Tutorials*, vol. 13, no. 3, Sept. 2011, pp. 443–61.

[12] X. Lu and S. Yang, "Thermal Energy Harvesting for WSNs," *Proc. IEEE Int'l. Conf. Systems Man and Cybernetics*, Istanbul, Turkey, Oct. 2010.

[13] X. Lu *et al.*, "Wireless Networks with RF Energy Harvesting: A Contemporary Survey," *IEEE Commun. Surveys and Tutorials*, to appear.

[14] G. Huang *et al.*, "A Biomass-Based Marine Sediment Energy Harvesting System," *Proc. IEEE Int'l. Symp. Low Power Electronics and Design*, Beijing, China, Sept. 2013.

[15] C. Li, J. Zhang, and K. B. Letaief, "Throughput and Energy Efficiency Analysis of Small Cell Networks with Multi-Antenna Base Stations," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, May 2014, pp. 2505–17.

[16] Y. Luo, J. Zhang, and K. B. Letaief, "Achieving Energy Diversity with Multiple Energy Harvesting Relays," *Int'l. Conf. Wireless Commun. and Signal Processing*, Hefei, China, Oct. 2014.

Biographies

YUYI MAO [S'14] (ymaoac@ust.hk) received his B.Eng degree in information and communication engineering from Zhejiang University, Hangzhou, China, in 2013. He is currently working toward a Ph.D. degree in the Department of Electronic and Computer Engineering at Hong Kong University of Science and Technology under the supervision of Prof. Khaled B. Letaief. His current research interests include energy harvesting cellular systems, cooperative systems, smart grid communications, and stochastic optimization.

YAMING LUO [S'11] (luoymhk@ust.hk) received his B.Eng. degree from the Department of Communication Engineering at Harbin Institute of Technology, China, in 2010. He is currently working toward a Ph.D. degree in the Department of Electronic and Computer Engineering at Hong Kong University of Science and Technology under the supervision of Prof. Khaled B. Letaief. His current research interests include energy harvesting networks, relay systems, and green communications.

JUN ZHANG [M'10] (eejzhang@ust.hk) received his Ph.D. degree in electrical and computer engineering from the University of Texas at Austin in 2009. He is currently a research assistant professor in the Department of Electronic and Computer Engineering at Hong Kong University of Science and Technology. He co-authored the book *Fundamentals of LTE* (Prentice-Hall, 2010). His research interests include wireless communications and networking, green communications, and signal processing.

KHALED B. LETAIEF [S'85, M'86, SM'97, F'03] (eekhaled@ust.hk) received his Ph.D. from Purdue University. He is currently Chair Professor and Dean of Engineering at Hong Kong University of Science and Technology. He is an internationally recognized leader in wireless communications with over 500 papers and 15 patents. He is founding Editor-in-Chief of *IEEE Transactions on Wireless Communications* and a recipient of many honors including the 2009 IEEE Marconi Prize Award in Wireless Communications and 12 IEEE Best Paper Awards. He is an ISI Highly Cited Researcher.

Wireless Energy Harvesting for the Internet of Things

Pouya Kamalinejad, Chinmaya Mahapatra, Zhengguo Sheng, Shahriar Mirabbasi, Victor C. M. Leung, and Yong Liang Guan

ABSTRACT

The Internet of Things (IoT) is an emerging computing concept that describes a structure in which everyday physical objects, each provided with unique identifiers, are connected to the Internet without requiring human interaction. Long-term and self-sustainable operation are key components for realization of such a complex network, and entail energy-aware devices that are potentially capable of harvesting their required energy from ambient sources. Among different energy harvesting methods, such as vibration, light, and thermal energy extraction, wireless energy harvesting (WEH) has proven to be one of the most promising solutions by virtue of its simplicity, ease of implementation, and availability. In this article, we present an overview of enabling technologies for efficient WEH, analyze the lifetime of WEH-enabled IoT devices, and briefly study the future trends in the design of efficient WEH systems and research challenges that lie ahead.

INTRODUCTION

The Internet of Things (IoT) is an intelligent infrastructure of uniquely identifiable devices capable of wirelessly communicating with each other, services, and people on a large scale through the Internet [1]. IoT aims to make the Internet ubiquitous and pervasive, and has the potential to affect many aspects of users' quality of life. The networked heterogeneous devices connected in an IoT structure are typically equipped with sensors, controlling processors, wireless transceivers, and an energy source (e.g., a battery) to monitor their environment and send/receive data. Applications envisioned for IoT span a wide range of fields including home automation, healthcare, surveillance, transportation, smart environments, and many more. One of the dominant barriers to implementing such a grandiose scheme is supplying adequate energy to operate the network in a self-sufficient manner without compromising quality of service (QoS). Therefore, it is imperative to improve the energy efficiency and longevity of devices in IoT.

Although there are numerous methods to achieve energy efficiency, such as using lightweight communication protocols [2] or adopting low-power radio transceivers [3], the recent technology trend in energy harvesting provides a fundamental method to prolong battery longevity. Thus, energy harvesting is a promising approach for the emerging IoT [4]. Practically, energy can be harvested from environmental sources: thermal, solar, vibration, and wireless RF energy sources [5]. While harvesting from the aforementioned environmental sources is dependent on the presence of the corresponding energy source, RF energy harvesting provides key benefits in terms of being wireless, readily available in the form of transmitted energy (TV/radio broadcasters, mobile base stations and handheld radios), low cost, and small form factor implementation. This article presents an overview of wireless energy harvesting units in the context of wireless energy harvesting IoT (WEH-IoT) systems. In this scenario, multiple sensor nodes typically transmit data to a common sink node. The sink node, also known as a gateway, is connected to the network and is accessible to the outside world over the Internet.

A WEH-enabled sensor device usually consists of an antenna, a transceiver, a WEH unit, a power management unit (PMU), a sensor/processor unit, and possibly an onboard battery. Among those components, there are two essential units for energy harvesting: unit and PMU:

- The WEH unit is in charge of harvesting the RF energy and producing a stable energy source for the rest of the device. It also interfaces with the PMU.
- The PMU controls the transceiver, sensing unit functionality, and manages the energy consumption of each unit and/or accommodates battery charging using the harvested energy.

In this article, we also focus on the enabling technologies including high-efficiency wireless/RF energy harvesting rectifiers and low-power wake-up radio for WEH units. We propose a PMU architecture that accommodates a battery charging scheme using the harvested energy through a WEH unit. Furthermore, we analyze

Pouya Kamalinejad, Chinmaya Mahapatra, Shahriar Mirabbasi, and Victor C. M. Leung are with the University of British Columbia.

Zhengguo Sheng is with the University of Sussex.

Yong Liang Guan is with Nanyang Technological University.

This work is supported by funding from the Natural Sciences and Engineering Research Council of Canada, the ICICS/TELUS People and Planet Friendly Home Initiative at The University of British Columbia, TELUS and other industry partners.

0163-6804/15/$25.00 © 2015 IEEE

the lifetime of the proposed WEH-IoT system in the context of two common scenarios in IoT networked systems. The energy cost model is described using uniform and random distribution topology of sensor devices. It is shown that the lifetime is increased substantially using the wireless harvesting techniques. Finally, we conclude with a discussion of future research challenges.

The Wireless Energy Harvesting Unit

The WEH receives the transmitted radio waves with an antenna and converts the received RF energy into a stable direct current (DC) energy source to supply the sensor device. Generally, in the context of IoT, wireless sensor networks (WSNs), and RF identification (RFID) tags, wireless energy sources can be classified into two categories [6]:

1. *Dedicated source*: Dedicated RF sources are deployed to provide a predictable energy supply to the device. Dedicated sources can be optimized in terms of frequency and maximum power to meet the requirements of the sensor devices. A sink node is an example of a dedicated source.
2. *Ambient source*: This type of source is further divided into two subcategories:
   - *Static or anticipated ambient sources*, which are transmitters that radiate stable power over time, are not optimized (e.g., in terms of frequency and transmitted power) to supply the sensor device. Mobile base stations, and broadcast radio and TV are examples of anticipated ambient sources.
   - *Dynamic or unknown ambient sources*, which are transmitters that transmit periodically in a fashion not controlled by the IoT system. Harvesting energy from such sources require an intelligent WEH to monitor the channel for harvesting opportunities. WiFi access points, microwave radio links and police radios are examples of unknown ambient sources.

Different ambient sources transmit at different frequency bands. Harvesting wireless energy at multiple frequency bands complicates the antenna geometry requirements and demands a sophisticated power converter. Therefore, WEH is typically optimized to harvest from the dedicated energy source (e.g., sink node) and may be devised so as to allow ambient energy harvesting as an auxiliary source.

RF-to-DC Rectifier

In practice, the conversion from the received RF power to the usable DC supply comes with a certain amount of power loss in the matching circuit and in the internal circuitry of the power converter. The power conversion efficiency (PCE) of the converter is the ratio of the generated usable DC output power to the input RF power. State-of-the-art RF-to-DC converters (also known as rectifiers) can achieve high PCE values, up to 70 percent or more [7]. PCE is an indication of the amount of harvested energy that is available for the sensor device. The available harvested power, P_H, is given by a Friis equation [7] and is directly proportional to the transmitted power, P_T, path loss, P_L, transmitter antenna gain, G_T, receiver antenna gain, G_R, power conversion efficiency of the converter, PCE_H, and the square of the wavelength, λ, and is inversely proportional to the square of the communication distance, r, between the source and the device (Fig. 1).

A schematic diagram of the WEH unit is shown in Fig. 1a. The transmission power, communication medium, antenna gains, and frequency of operation are typically dictated by the application requirements. Therefore, a viable design parameter to enhance the harvested power P_H or maximize the communication distance r is PCE. The PCE curve as a function of distance/input power level for a typical rectifier is shown in Fig. 1b. As shown, the PCE is optimized to peak at a certain input level that corresponds to a specific distance (i.e., PCE_{MAX} at r_{opt}).

In addition to a high PCE, other important characteristics of the WEH unit include high sensitivity, wide high-efficiency range, multiband operation, and ease of implementation. Extensive studies have been performed on techniques to improve the efficiency of the converter unit [7]. One of our recent efforts to enhance PCE for rectifiers operating at small input levels and a technique to enable harvesting RF energy at multiple frequencies with a single antenna (to facilitate energy harvesting from ambient sources) are described in [8].

Wake-Up Radio Scheme

The radio transceiver is typically the most power hungry block of a wireless sensor device. Although the transceiver is rarely called into action during each operation cycle, it has to keep monitoring the channel. This idle listening process is a significant contributor to the overall power consumption of the sensor device.

An efficient approach to address the idle-mode energy consumption is *duty-cycling*, in which the receiver on-demand switches between listening and sleeping states. Among the different categories of duty-cycling (synchronous, pseudo-asynchronous, and pure asynchronous), the latter provides the most efficient solution in terms of energy consumption [3].

In the asynchronous approach, the sensor device is in deep sleep mode and only wakes up when signaled by the sink node or its neighboring devices through an interrupt command generated by a low-power wake-up radio (WUR). The timing diagram of the asynchronous communication approach is shown in Fig. 1c. Since the WUR is constantly active to monitor the channel, this scheme outperforms other alternatives only if the energy consumption of the WUR is negligible compared to that of the main receiver. The block diagram of a WUR-enabled sensor device is shown in Fig. 1d.

A WUR is a simple receiver that receives the wake-up command (e.g., the device unique address) and generates an interrupt for the main receiver. In a WEH device, the implemented rectifier followed by a data slicer (comparator) can perform as a WUR with minimal complexity overhead. We advocate the use of such a WUR

An efficient approach to address the idle-mode energy consumption is duty-cycling, in which the receiver on-demand switches between listening and sleeping states. Among the different categories of duty-cycling (synchronous, pseudo-asynchronous, and pure asynchronous), the latter provides the most efficient solution in terms of energy consumption.

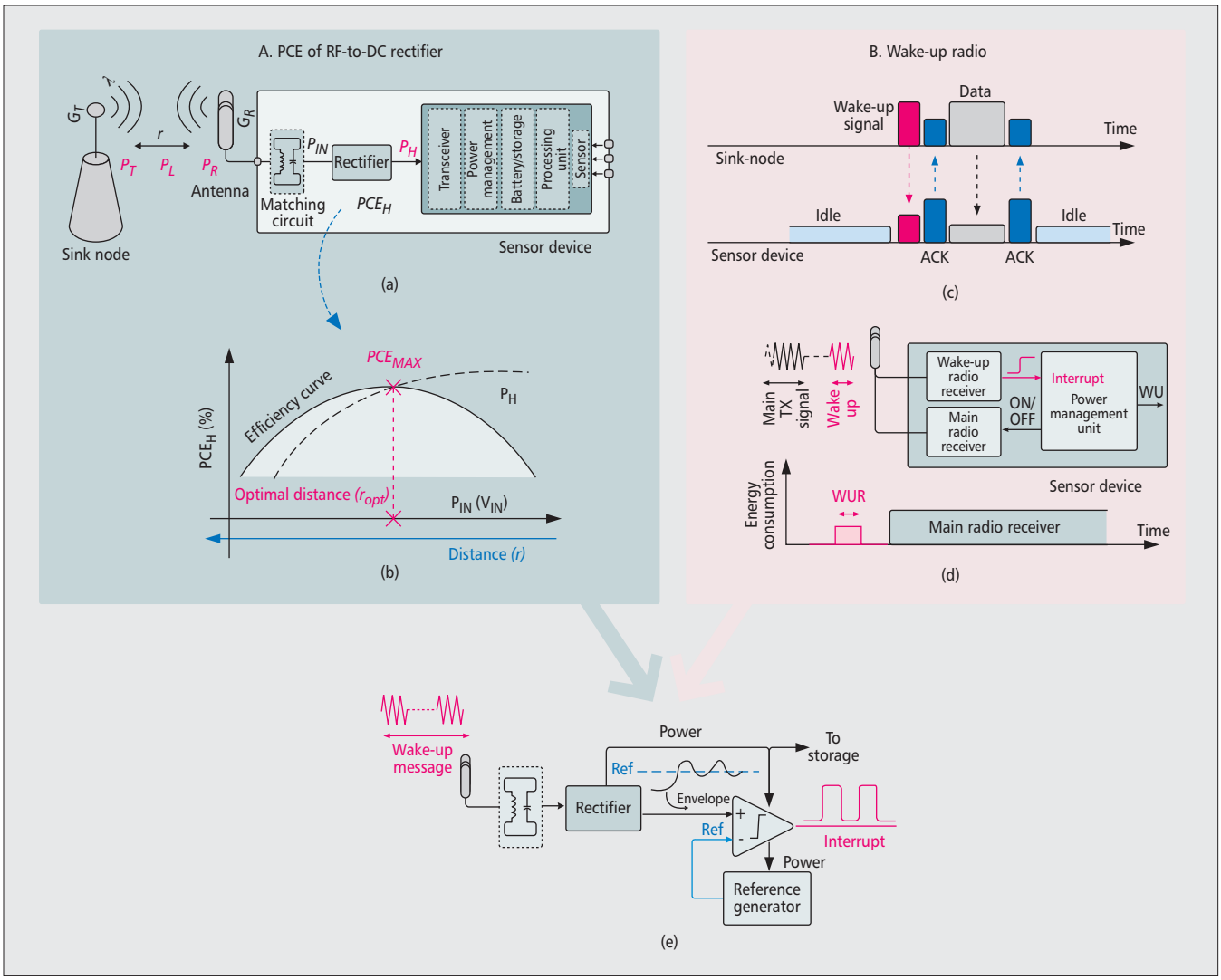


Figure 1. A WUR-enabled energy harvesting unit: a) block diagram of the WEH sensor device; b) efficiency curve (solid line) and harvested energy of the rectifier (dashed line); c) timing diagram of asynchronous wake-up scheme; d) block diagram and energy consumption of the WUR; e) zero-power interrupt generation unit.

(for detection of a simple on/off keying [OOK] wake-up message) as schematically shown in Fig. 1e [9]. If all the required energy of the WUR circuitry is supplied by the harvested energy from the wake-up message itself, the battery is not used during the idle listening mode (virtually zero energy consumption). Aside from energy consumption, high sensitivity and range, robustness to interferers, selectivity, and latency are also of paramount importance in designing a WUR for IoT.

Power Management Unit

An integral part of any energy harvesting system is its power management unit (PMU). The PMU is in charge of controlling the storage of the harvested energy. It also manages the distribution of the available energy among different consumers in an effort to maximize the lifetime of the device while maintaining high QoS. We extend the architecture of the PMU proposed in [10] to enable effective cooperation with the WEH unit. The architecture proposed is an event-triggered/asynchronous scheme based on the signal generated by WUR. The PMU architecture also detects/preempts the failure of a node in the event of energy deficiency.

A detailed block diagram of the PMU for the WEH sensor device is shown in Fig. 2. The PMU starts its operation by a trigger signal generated by the WUR of the WEH unit (*INTERRUPT*). The PMU first activates the main transceiver through (*ON*/*OFF*) and then sends a wake up signal (*WAKE UP*) to the sensing unit to start its operation. The sensing unit toggles the *STOP*/*RUN* to high, signifying the PMU that it is in running mode. The *REQ* signal indicates the amount of energy required by the sensing unit. The signals *BAT* and *SE* indicate the amount of energy left in the battery device and the WEH unit storage element, respectively. Accordingly, the PMU activates switch SW_1 through signal *SENSE* to fulfill the power requirements of the sensing unit. The sensor unit is in charge of sensing, data processing via a microprocessor (μp) and finally transmitting them to a low-power transceiver based on Bluetooth, WiFi,

IEEE 802.15.4, Zigbee, and so on. The sensor device requires a minimum power of P_{Dmin} to operate in sensing mode. When the energy in the battery device goes below a certain threshold, $P_{TH} < 1.5\ P_{Dmin}$, the PMU sends a *RECHARGE* command to the storage element by activating switch SW_2 of the WEH unit to charge the battery. When the energy level of the device remains 1.1 P_{Dmin}, the device sends an out of service (OUS) command to the sink node, signaling that it is going out of service until it recharges itself again to more than 1.5 P_{Dmin}. The sink node in turn sends a stop all service (SAS) signal to the device. The sink node/gateway puts the device out of the sensing service loop but keeps transmitting RF energy for harvesting. When the device is ready for service again, it sends a *READY* signal to the sink node, which in turn sends a resume all services (RAS) signal to the device.

Life-Time Prediction Through Energy Cost Model

Let us consider a network of k static and identical sensor devices. As in [11], WSNs are either uniformly distributed in a ring topology, communicating with the sink node in a peer-to-peer fashion, or randomly distributed in a multihop ad hoc topology. As WSNs are a subset of an IoT system, we base our analysis on these topologies. Table 1 delineates the parameters used for the analysis of our scenarios.

The power transmitted by the sink node has a certain maximum and minimum value of P_{Tmax} and P_{Tmin}, respectively, and the sensor devices require a minimum power of P_{Dmin} to function properly. The sink node communicates with sensor devices asynchronously. We define two operational modes for the sensor devices: *active* and *idle* modes.

The lifetime of an IoT system (battery operated only/battery with WEH unit) depends on the average energy consumption of the sensor devices E_D per active duty cycle. This involves the combined operations of sensing, processing, and communication (receive/transmit).

Let N be the total number of active duty cycles representing the lifetime of the sensor device. The communication energy consists of E_{LS} (listening energy), E_{RX} (receiver energy), and E_{TX} (transmitter energy). The computation energy includes E_{PR} (processing energy) and E_{SN} (sensing energy). To capture the energy distribution among the aforementioned energy consumers, weighting coefficients $\alpha_{LS} > \alpha_{TX} > \alpha_{RX} > \alpha_{PR} > \alpha_{SN}$ are assigned to them. The total average energy consumption $E_D = \alpha_{LS} E_{LS} + \alpha_{TX} E_{TX} + \alpha_{RX} E_{RX} + \alpha_{PR} E_{PR} + \alpha_{SN} E_{SN}$. E_B is the total energy stored in the battery, and E_H is the available harvested energy per active duty cycle. We assume constant energy consumptions for receiver, processor, and sensor. However, the energy consumption of the transmitter (E_{TX}) is directly proportional to r_{ij}^2, where r_{ij} is the distance between the originating device j and the sink node i (in ring topology) or the sink node/sensor device (in multihop topology).The harvested energy E_H is inversely proportional to r_{ij}^2 (here j is the sink node and $r_{ij} = r_{ji}$). Based on these assumptions, the lifetime estimation N_H for the sensor device operated with a battery assisted by a WEH unit can be formulated as

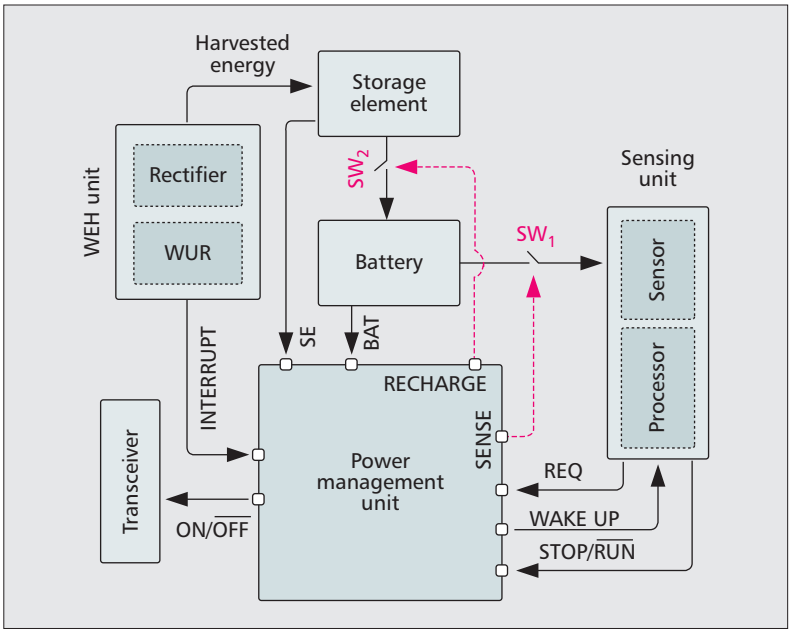


Figure 2. The proposed power management unit.

$$N_H = \frac{E_B}{C_D + CTX \cdot r_{ij}^2 - \frac{C_H}{r_{ji}^2}} \quad (1)$$

where $C_D = \alpha_{LS} E_{LS} + \alpha_{RX} E_{RX} + \alpha_{PR} E_{PR} + \alpha_{SN} E_{SN}$. C_{TX}, and C_H are the proportionality constants for E_{TX} and E_H as mentioned above. We analyze the lifetime of the sensor devices for two different scenarios. Note that the lifetime of the battery-only device (without WEH) is denoted as N_B.

Scenario I: Uniform Distribution in a Ring Topology

In the ring topology, sensor devices (js) are uniformly distributed around the sink node (i) at a distance r_{ij}, as shown in Fig. 3a. Since they are positioned equidistant from the sink node, all the devices receive a similar amount of wireless energy for harvesting. Assuming the channel is static between the sink node and the sensor devices, the transmitted energy from the device to the sink is the same for all nodes. Figure 3b schematically shows the distribution of energy among different consumers. The horizontal axis depicts the active duty cycle (T). Energy consumption for E_{TX}, E_{RX}, E_{PR}, and E_{SN} occur only for a fraction of the total active duty cycle, whereas energy consumption for E_{LS} and harvested energy E_H happen constantly throughout T. Based on these assumptions, the lifetime of the devices is estimated using Eq. 1 for N_B (battery only, E_H is zero) and N_H. Figure 3c shows the estimated lifetime vs. the power consumption of the devices (E_D). By incorporating the WUR scheme, we reduce the $\alpha_{LS} E_{LS}$ to have an almost negligible effect on C_D.

Parameter	Description
E_D	Total energy consumption of device
E_B	Energy of the battery
E_H	Harvested energy
E_{LS}	Listen mode energy
E_{RX}	Receive mode energy
E_{TX}	Transmit mode energy
E_{PR}	Processing energy
E_{SN}	Sensing energy
$\alpha_{LS} > \alpha_{TX} > \alpha_{RX} > \alpha_{PR} > \alpha_{SN}$	Weighting coefficients of the respective energy parameters
r_{ij}	Distance of originating device j to sink node i
N	Estimated device lifetime
N_B	Estimated device lifetime (battery only)
N_H	Estimated device lifetime (WEH-assisted battery)
$N_H(WUR, r)$	Estimated device lifetime (WUR-enabled WEH assisted battery) at distance $r = r_{ij}$

Table1. Parameters used for lifetime analysis.

Typical values for the distribution of different energy consumers are from [9, 12]. Energy harvesting increases the lifetime of the battery assisted device (N_H) by ~30 percent for low-power sensor devices (e.g., temperature, pressure, and light sensors). The WUR scheme enhances the lifetime by a further ~110 percent as depicted by $N_H(WUR, r)$. Setting $r_{ij} = r/2$, the lifetime of an energy harvesting device increases with the reduction in distance between sink node and sensor node as shown in $N_H(WUR, r/2)$.

Scenario II: Randomly Distributed Multihop Topology

In the multihop topology, sensor devices (j_Ns) are randomly distributed following a Poisson distribution. The maximum distance of the kth-farthest sensor device from the sink node is r_{ijk}. In a multihop transmission, the j_k sensor node's data hops $k - 1$ times before it reaches the sink node/gateway. The farthest node j_k only acts as a sensor and transmits its own data. However, the remaining $j_1, j_2, \ldots, j_k$ nodes act as both a sensor for their own data and a relay for data coming from farther nodes. Thus, to compute their lifetime, it is required to add the energy consumption of the relay cycle to that of the sensor cycle. The relay cycle energy consumption is E_D-$\alpha_{PR}E_{PR}$-$\alpha_{SN}E_{SN}$, which is the total energy consumption for a device operation minus the processing and sensing energy. Figure 4a depicts a two-hop scenario. Figure 4b shows the distribution of energy among different consumers. The transmit mode energy is smaller in this case than in Fig. 3b in the ring topology as the transmission distance is reduced to r_{j2j1}. Figure 4c shows the energy distribution for the sensor node j_1 operating as a sensor and as a relay for j_2. The estimated lifetime of the sensor devices in the two-hop topology is shown in Fig. 4d for nodes j_2 and j_1 (lifetime for battery only, N_B, and for WEH assisted battery, N_H). As shown in the figure, energy harvesting extends the lifetime of the nodes. Node j_1 is exercised twice as often as j_2, so it consumes more energy and has a shorter lifetime than node j_2. The lifetime enhancement through WEH for a node at distance similar to that of the ring topology is approximately 5 times larger (~510 percent as compared to ~110 percent in ring topology) for low-power sensor devices. Therefore, for the N-hop topology, the lifetime of the devices is further enhanced.

Future Trends and Research Challenges

The approaches presented in this work are not exhaustive. For the proposed system to become more practically viable, there are research challenges ahead that need to be addressed.

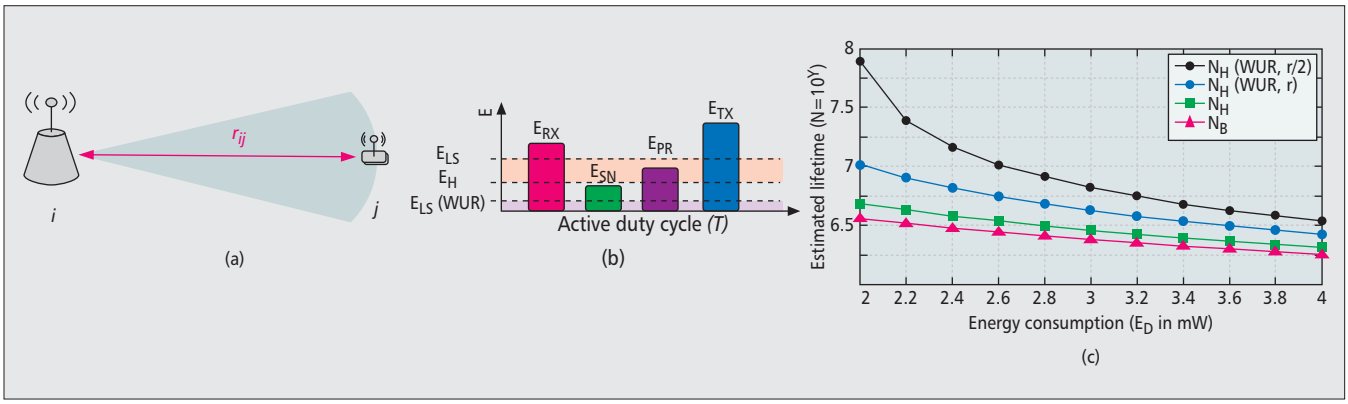


Figure 3. Lifetime prediction through energy cost model: a) ring topology architecture; b) energy distribution; c) estimated lifetime vs. energy consumption.

Highly Efficient, Low-Cost, and Small-Form-Factor Wireless Energy Harvesting System

The key challenge in successful large-scale deployment of sensor devices in an IoT infrastructure is to minimize their impact on users and the environment. Non-intrusive devices need to be small, be fabricated and deployed at very low cost, and are expected to operate in a self-sufficient manner for a long time. A WEH unit as an integral part of such devices must comply with such cost and size requirements. Efficiency is another crucial factor for a WEH system. High efficiency becomes increasingly relevant considering that the transmitted power by the dedicated source is usually limited due to health issues and interference constraints.

Commercial RF harvesting systems currently existing in the market enable single-band RF harvesting at sub-milliwatt power levels with efficiencies as high as 50 percent. However, extensive studies are still being carried out to improve the performance of WEH systems at the circuit and system levels. Energy beamforming [5], high gain antennas, and multi-band harvesting are among the other hot topics in the context of WEH systems for IoT.

Channel Statistics for Wireless Energy Harvesting IoT Systems

The scenarios and their respective analysis in our article assume the channel as static and time invariant. Practically, channel characteristics vary depending on the environment in which the number of interferers and the number of paths available from source device to sink. Harvested energy depends on the distance between sink and sensor node. In the presence of fading or multipath, the received energy for the purpose of harvesting and the transmitted data are adversely affected. In [13], a compressive sensing based approach is proposed to recover sparse signals from multiple spatially correlated data transmitted to a fusion center. Recently, in [14], researchers have proposed techniques to reduce the amount of packets to be retransmitted in case of faulty transmission, eventually saving energy.

Cross-Layer Design of Wireless Energy Harvesting

Although the recent development of energy harvesting technologies mitigates the energy scarcity issue, the sensor device still has to operate in duty-cycled mode due to limited energy collection from the environment, and dynamically adjust duty cycles to adapt to the availability of environmental energy. Such dynamic duty cycles pose challenges for medium access control (MAC) layer protocol design in terms of synchronization, reliability, efficiency of utilizing channel resource and energy, and so on. Therefore, solutions of duty-cycling-aware middleware between MAC and physical layer power management are highly desired. Moreover, dynamic duty cycling also has nontrivial impact on the end-to-end performance of the network layer,

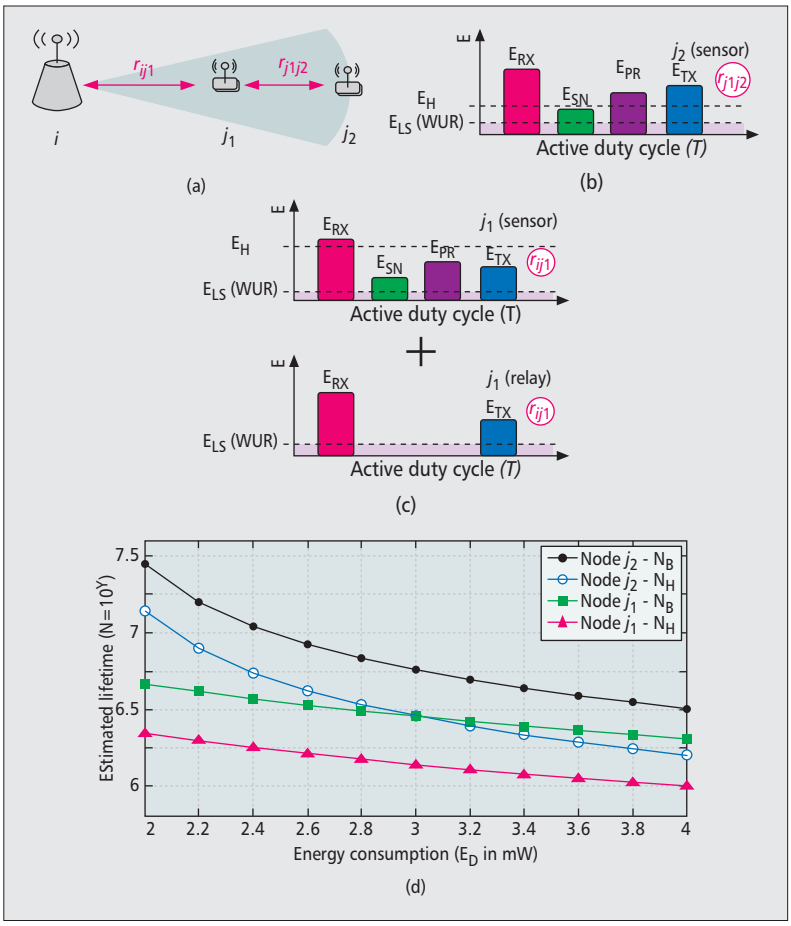


Figure 4. Life-time prediction through energy cost model: a) random distribution architecture; b) energy distribution of sensor j_2; c) energy distribution of sensor j_1 as sensor and relay; d) estimated lifetime vs. energy consumption.

including end-to-end delay, throughput, and so on. However, the current routing protocol design for IoT has paid very little attention to duty cycling. The problem of seamlessly integrating duty-cycle awareness into the multi-path routing scenario has been dealt with in [15] using a sleep scheduling mechanism; however, it still remains an open question.

Smartphone Relays: Wireless Energy Harvesting in 5G Systems

Fueled by increases in traffic and data demand, mobile technology is moving toward the fifth generation (5G), where everything will be connected via the Internet and accessed through the cloud. Relay techniques utilized by 5G can benefit wireless/RF energy harvesting [6]. Also, smartphones, due to their mobility, can act as gateways/sink nodes, relaying data for sensors in personal area networks and as a source of RF energy for harvesting purposes.

Concluding Remarks

In this article, we have overviewed technologies and schemes to enable WEH for IoT systems. With an emphasis on improving the efficiency of the WEH unit and reducing the energy con-

Although the recent development of energy harvesting technologies mitigates the energy scarcity issue, the sensor device still has to operate in duty-cycled mode due to limited energy collection from the environment, and dynamically adjust duty cycles to adapt to the availability of the environmental energy.

sumption of the devices, the lifetime of WEH-assisted battery-operated systems in an IoT architecture are analyzed for two different scenarios. A study of the specific energy requirements of IoT devices reveals that achieving self-sustainability requires improved design techniques at both the circuit and system levels.

References

[1] Z. Sheng *et al.*, "A Survey on the IETF Protocol Suite for the Internet of Things: Standards, Challenges, and Opportunities," *IEEE Wireless Commun.*, vol. 20, no. 6, Dec. 2013, pp. 91–98.

[2] Z. Sheng, C. Zhu, and V. C. M. Leung, "Surfing the Internet-of-Things: Lightweight Access and Control of Wireless Sensor Networks Using Industrial Low Power Protocols," *EAI Endorsed Trans. Industrial Networks and Intelligent Systems*, vol. 14, no. 1, 2014.

[3] V. Jelicic *et al.*, "Analytic Comparison of Wake-Up Receivers for WSNs and Benefits Over the Wake-On Radio Scheme," *Proc. 7th ACM Wksp. Performance Monitoring and Measurement of Heterogeneous Wireless and Wired Networks*, 2012, pp. 99–106.

[4] S. Sudevalayam and P. Kulkarni, "Energy Harvesting Sensor Nodes: Survey and Implications," *IEEE Commun. Surveys and Tutorials*, vol. 13, no. 3, 3rd 2011, pp. 443–61.

[5] G. Yang, C. K. Ho, and Y. L. Guan, "Dynamic Resource Allocation for Multiple-Antenna Wireless Power Transfer," *IEEE Trans. Signal Processing*, vol. 62, no. 14, July 2014, pp. 3565–77.

[6] L. Xiao *et al.*, "Wireless Networks with RF Energy Harvesting: A Contemporary Survey," *IEEE Commun. Surveys and Tutorials*, 2014, pp. 1–1.

[7] M. Russo, P. Šolić, and M. Stella, "Probabilistic Modeling of Harvested GSM Energy and Its Application in Extending UHF RFID Tags Reading Range," *J. Electromagnetic Waves and Applications*, vol. 27, no. 4, 2013, pp. 473–84.

[8] P. Kamalinejad *et al.*, "Efficiency Enhancement Techniques and A Dual-Band Approach in RF Rectifiers for Wireless Power Harvesting," *Proc. IEEE Int'l. Symp. Circuits and Systems*, June 2014, pp. 2049–52.

[9] P. Kamalinejad *et al.*, "A High-Sensitivity Fully Passive Wake-Up Radio Front-End for Wireless Sensor Nodes," *Proc. IEEE Int'l. Conf. Consumer Electronics*, Jan. 2014, pp. 209–10.

[10] S. S. Kumar and D. K. Kashwan, "Research Study of Energy Harvesting in Wireless Sensor Networks," *Int'l. J. Renewable Energy Research*, vol. 3, no. 3, 2013, pp. 745–53.

[11] C. Tunca *et al.*, "Distributed Mobile Sink Routing for Wireless Sensor Networks: A Survey," *IEEE Commun. Surveys and Tutorials*, vol. 16, no. 2, 2014, pp. 877–97.

[12] M. A. Razzaque and S. Dobson, "Energy-Efficient Sensing in Wireless Sensor Networks Using Compressed Sensing," *Sensors*, vol. 14, no. 2, 2014, pp. 2822–59.

[13] G. Yang *et al.*, "Wireless Compressive Sensing for Energy Harvesting Sensor Nodes," *IEEE Trans. Signal Processing*, vol. 61, no. 18, Sept. 2013, pp. 4491–4505.

[14] H. Sharma and P. Balamuralidhar, "A Transmission Scheme for Robust Delivery of Urgent/Critical Data in Internet of Things," *Proc. 5th Int'l. Conf. Commun. Systems and Networks*, Jan. 2013, pp. 1–7.

[15] L. Shu *et al.*, "Impacts of Duty-Cycle on TPGF Geographical Multipath Routing in Wireless Sensor Networks," *Proc. 18th Int'l. Wksp. Quality of Service*, June 2010, pp. 1–2.

Biographies

Pouya Kamalinejad (pkamali@ece.ubc.ca) received his B.Sc. and M.Sc. degrees in electrical and computer engineering from the University of Tehran, Iran, in 2006 and 2008, respectively, and his Ph.D. degree in electrical and computer engineering from the University of British Columbia (UBC) in 2014. He is currently a post-doctoral fellow at the University of British Columbia. His current interests include RF and low-power integrated circuit design, wireless energy harvesting for RFID tags and wireless sensor networks, and sensor interface design.

Chinmaya Mahapatra (chinmaya@ece.ubc.ca) received his B.Tech. degree in electronics and communication engineering from N.I.T Rourkela, India, in 2009, and his M.A.Sc. degree in electrical and computer engineering from UBC in 2013. He is currently a Ph.D. student of electrical and computer engineering at the University of British Columbia. His current interests include the Internet of Things, body sensor area networks, embedded systems, sensor cloud, and smartphone energy optimization.

Zhengguo Sheng (z.sheng@sussex.ac.uk) is a lecturer at the School of Engineering and Informatics, University of Sussex, United Kingdom. He is also a visiting faculty member at UBC and the co-founder of WRTnode. Previously, he was with UBC as a research associate, and with France Telecom Orange Labs as a senior researcher and project manager in M2M/IoT. He also worked as a research intern with the IBM T. J. Watson Research Center and U.S. Army Research Labs. Before joining Orange Labs, he received his Ph.D. and M.S. with distinction from Imperial College London in 2011 and 2007, respectively, and his B.Sc. from the University of Electronic Science and Technology of China (UESTC) in 2006. He has published more than 30 international conference and journal papers. He is also the recipient of the Auto21 TestDRIVE Competition Award 2014 and Orange Outstanding Researcher Award 2012. His current research interests cover IoT/M2M, cloud/edge computing, vehicular communications, and power line communication.

Shahriar Mirabbasi (shahriar@ece.ubc.ca) received his B.Sc. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 1990, and his M.A.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Toronto, Ontario, Canada, in 1997 and 2002, respectively. Since August 2002, he has been with the Department of Electrical and Computer Engineering, UBC, where he is currently a professor. His current research interests include analog, mixed-signal, RF, and mmWave integrated circuit and system design with particular emphasis on communication, sensor interface, and biomedical applications.

Victor C. M. Leung [F] (vleung@ece.ubc.ca) is a professor and holder of the TELUS Mobility Research Chair in Advanced Telecommunications Engineering in the Department of Electrical and Computer Engineering at UBC, where he completed his B.A.Sc. and Ph.D. degrees in 1977 and 1981, respectively. He has been involved in telecommunications research with a focus on wireless networks and mobile systems for more than 30 years, which has resulted in more than 700 journal and conference papers co-authored with his students and collaborators, including several papers that won best paper awards. He is a Fellow of EIC and CAE. He was a Distinguished Lecturer of the IEEE Communications Society. He has served/is serving on the Editorial Boards of many journals. He has contributed to the organization committees and technical program committees of numerous conferences. He was the winner of an APEBC Gold Medal in 1977, an NSERC Postgraduate Scholarship for 1977–1981, an IEEE Vancouver Section Centennial Award in 2011, and a UBC Killam Research Prize in 2012.

Guan Yong Liang (eylguan@ntu.edu.sg) obtained his Ph.D. from the Imperial College of London, United Kingdom, and his B.Eng. with first class honors from the National University of Singapore. He is an associate professor at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests broadly include modulation, coding and signal processing for communication, and storage and information security systems.

Joint Wireless Information and Energy Transfer in Massive Distributed Antenna Systems

Fangchao Yuan, Shi Jin, Yongming Huang, Kai-Kit Wong, Q. T. Zhang, and Hongbo Zhu

ABSTRACT

In mobile communications, two potentially ground-breaking ideas have emerged in recent years: massive MIMO antenna technology, which delivers enormous information rates, and wireless energy transfer (WET), which makes remote charging of mobile users possible. This article realizes the huge potential of combining the two for high wireless information transfer (WIT) and WET. In particular, to maximize the synergy, the distributed version of massive MIMO (known as massive distributed antenna system, MDAS) is advocated to go with the joint wireless information and energy transfer (JWIET) system capable of both WIT and WET. We present the opportunities in MDAS-JWIET, and discuss research trends in MDAS with several architectures involving WET and WIT.

INTRODUCTION AND MOTIVATION

Future-generation mobile communication systems ought to provide data-rich wireless services anytime anywhere [1]. It is estimated that in the next decade, there will be a 1000× increase in the mobile traffic we experience today [1]. To meet this explosive growth, the spectral efficiency and energy efficiency of commercial mobile communications must be increased accordingly. However, it is a fundamental challenge to increase both spectral efficiency and energy efficiency simultaneously, as they are often conflicting goals. As a result, one would expect that increasing spectral efficiency would naturally come with decreasing energy efficiency.

A radical approach to achieve simultaneous enhancement in spectral efficiency and energy efficiency is to deploy an excessively large number of base station (BS) antennas (e.g., 100 or more), resulting in a massive multiple-input multiple-output (MIMO) system [2]. Excessive use of BS antennas makes possible the radio energy to focus on extremely small regions of space to yield a huge gain in the achievable rate of wireless information transfer (WIT).

It has been reported in [2] that massive MIMO can increase capacity 10× or more while simultaneously enhancing the radiated energy efficiency 100×. Remarkably, it was shown in [3] that every mobile user in a massive MIMO system can scale down its transmit power proportionally to the number of BS antennas (or the square root of the number of antennas at the BS if the channel state information, CSI, is imperfect) in order to achieve the same performance as a single-input single-output (SISO) system. This exceptional energy efficiency has made massive MIMO one of the candidate technologies for future-generation wireless communications networks.

Even if mobile systems are able to maximize the delivery for every drop of energy spent, this would still be insufficient because mobile users (or user equipments, UEs, in Third Generation Partnership Project, 3GPP, terminology) operate with finite battery power, and it is important to find an alternative source of power to prolong the life of UEs and provide them with more flexible power constraints. Wireless energy transfer (WET) has long been regarded as a possibility, dating back to as early as 1891 in Tesla's demonstration [4]. However, not until recently has WET become widely recognized as "doable" for mobile communications.

The concept of WET brings numerous new opportunities to mobile communications. With WET, UEs can now be charged wirelessly from ambient signals or deliberate remote charging signals. This enables proactive and controllable energy replenishment of mobile devices for genuine mobility where mobile devices do not depend on centralized power sources. Besides, more sophisticated power constraints that take into account future intake of remote energy can be adopted in order to more aggressively improve system performance.

Indeed, attention on WET in mobile communications networks has been growing recently, and has generated a large body of literature. For example, a hybrid network for powering the UEs, overlaid with an uplink cellular network with randomly deployed power beacons, was proposed in [5]. Also, in [6], a harvest-then-transmit protocol was introduced for wireless energy receivers to transmit information in the

Fangchao Yuan, Q. T. Zhang, and Hongbo Zhu are with Nanjing University of Posts and Telecommunications.

Shi Jin, Yongming Huang, are with Southeast University.

Kai-Kit Wong is with University College London.

0163-6804/15/$25.00

The large number of antennas of a BS are in groups of RAUs and are quite evenly distributed in a cell. The RAUs within a cell are all connected to a baseband processing unit using high-quality bidirectional wired or wireless links. The RAUs are functionally simple, whereas the BPU has all the necessary baseband processing capability of a BS.

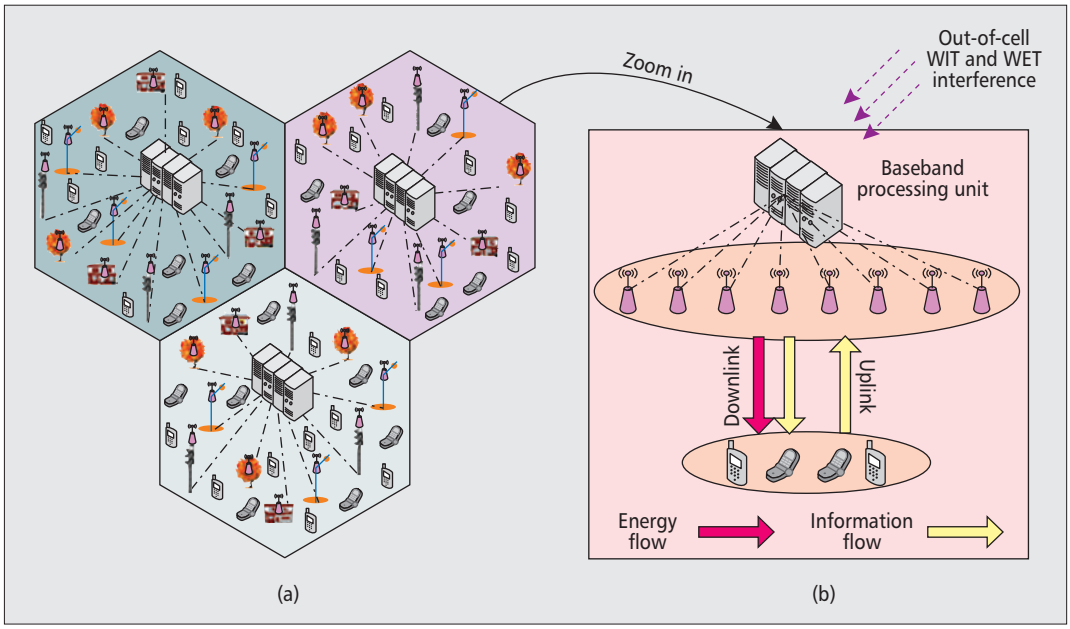


Figure 1. The architecture for integrating MDAS and JWIET: a) MDAS; b)WIT and WET.

uplink. This protocol was recently applied to the massive MIMO setting in [7], anticipating a huge gain in transmission capacity and reduction in the transmission power through precoding and coherent combining. Dual use of RF signals for information as well as energy transfer, widely known as simultaneous wireless information and energy transfer (SWIPT), also provides a low-cost solution for sustainable operation of wireless systems [8].

A major issue for WET is the fast decay in energy transfer efficiency over the transmission distance due to propagation loss. Thus, various MIMO beamforming techniques have been proposed to help restore the energy transfer efficiency [6–9]. Clearly, there is possible synergy between massive MIMO and WET (or more generally SWIPT), but marriage of the two is not at all straightforward. In particular, user fairness could become a critical issue, as the energy transfer efficiency for the UEs might be significantly different due to their different locations. According to relational expression of the transmit power and the receiver power in [6–9], we can easily see that the user far from the transmitter harvests noticeably less power. Only the user close to the transmitter can receive considerable energy. Moreover, pilot contamination is another issue for the achievable WET efficiency.

Interestingly, the above-mentioned problems will disappear in massive distributed antenna systems (MDASs) in which the remote antenna units (RAUs) are more arbitrarily distributed over the cells, and the distance from any given UE to its nearby RAU(s) is much smaller, making WET more feasible. Unlike the centralized version of massive MIMO systems, in MDAS, multiple RAUs consisting of one or more antennas are distributed geographically and connected with the BSs via dedicated high-speed backhaul links, leading to significant distance reduction of the radio transmission between the antennas and the UEs, and improvement in the radiated energy efficiency. Of course, MDAS possesses the same superpower of ultra-high energy efficiency and spectral efficiency as do standard massive MIMO systems. When the number of distributed antennas becomes massive, by virtue of the law of large numbers, the effects of uncorrelated noise and fading vanish, and intra-cell interference (ICI) can be mitigated.

Although MDAS for high WIT is preliminarily well understood [10, 11], using it for WET or generally JWIET demands a rethink in many design and analytical aspects. Despite the enormous potential of marrying MDAS and JWIET, there are obstacles ahead. The aim of this article is to provide several intuitions on coordinating WET and WIT in MDAS systems, and reveal the associated challenges and opportunities.

Architecture of the MDAS-JWIET Paradigm

In this section, we first describe the general architecture of MDASs for WIT and WET, and introduce the RAU transmitter and UE receiver designs in the downlink. A typical distributed architecture for WIT and WET is shown in Fig. 1. The large number of antennas of a BS are in groups of RAUs, and they are quite evenly distributed in a cell. The RAUs within a cell are all connected to a baseband processing unit (BPU) using high-quality bidirectional wired (e.g., RoF) or wireless (e.g., microwave repeater) links. The RAUs are functionally simple, whereas the BPU has all the necessary baseband processing capability of a BS. Figure 1b shows that the distributed RAUs support JWIET. The UEs enjoy the information and/or energy services from the network. The traditional WIT-only setup will not be optimal, and we discuss the key differences below.

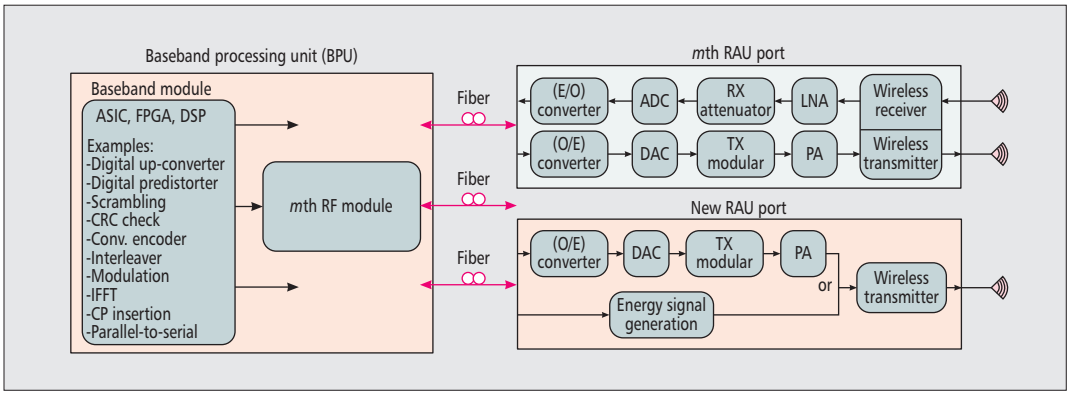


Figure 2. A block diagram showing the interaction between the BPU and RAU.

Since electrical energy is isotropically propagated from an isotropic antenna, a UE can only harvest a small fraction of the transmitted energy, resulting in low WET efficiency. This can be tackled by a distributed energy beamforming approach where a cluster of distributed energy sources mimic an antenna array.

Transmitter Circuitry

Figure 2 illustrates the blocks concerning the interplay between the BPU and the RAUs for an MDAS transmitter with, say, M RAUs. The BPU is comprised of a baseband module and M RF modules. The baseband module performs the needed digital signal processing, such as digital up-conversion, encoding, and modulation (more examples in Fig. 2), and distributes the digital signals to the M RF modules. Each RF module converts the digitized electrical signals into optical signals that can be sent to the remote RF module at the RAU port over an optical fibre. Each RAU port then emits the electrical signals through a digital-to-analog converter (DAC), a TxModular, and a power amplifier (PA) followed by an RF antenna. Signal reception is a similar process but in a reverse fashion.

The WET concept suggests that some of the RAUs transfer energy-only RF signals, which is not needed to be modulated, processed, and so on at the same level as the information-bearing signal. This leads to great simplification in the design of BPU and RAU for power saving. For example, BPU can now have a simple baseband process for the energy-only RF signal, or the RAU may generate the RF signal itself because the energy-only signal is easy to produce, which can reduce considerable computation at the BPU and the burden of the fronthaul. Thus, for WET, the information receiving module is not required in this kind of RAU, greatly reducing the hardware cost.

Transmit Beamforming

Since electrical energy is isotropically propagated from an isotropic antenna, a UE can only harvest a small fraction of the transmitted energy, resulting in low WET efficiency. This can be tackled by a distributed energy beamforming approach where a cluster of distributed energy sources mimic an antenna array by transmitting RF energy collaboratively for a focused directional energy beam to an intended energy receiver. The potential energy gains at the receiver from distributed energy beamforming are expected to be of the same magnitude as the information gains from information beamforming. However, challenges arise in the implementation. For instance, timing synchronization among distributed energy sources and coordination of distributed carriers in phase and frequency so that RF signals can be combined constructively at the receiver side are some of the problems. This section presents the beamforming designs for MDAS-WET (and also MDAS-SWIPT) classified into the following four categories.

Downlink SWIPT Beamforming: In SWIPT, beamforming designs have to be made to steer the RF signals toward the target receivers with different information and energy harvesting (EH) requirements. SWIPT beamforming is presently a hot topic of research, and was first introduced in a three-node MIMO network [8] where the optimal transmission strategies were investigated in order to achieve the best trade-off between WIT and WET assuming the availability of perfect channel state information at the transmitter (CSIT). In MDAS-WIT, however, the major difference from the co-located MIMO systems when designing beamforming lies in the heterogeneity of large-scale channel effects from different RAUs for a given user, which also affects the WET beamforming. Note that the ICI and out-of-cell interference (OCI) should be taken as a kind of energy source, leading to a totally different metric for WIT.

Moreover, the knowledge of CSI in MDAS plays an important role in the beamforming optimization. Perfect acquisition of CSI is practically impossible due to factors such as channel estimation error, feedback error, and time-varying channel. To accurately estimate a channel state, a significant overhead will be incurred at a UE. Although longer time for channel estimation exploiting channel reciprocity results in more accurate CSI, this also leads to reduced time for transmission, and less harvested energy as well as information. Hence, beamforming optimization for SWIPT should consider the trade-off between durations for transmission and CSI estimation. Furthermore, in order to get perfect CSI, backhaul is a key factor for the MDASs. In practical implementations, the large amounts of received signals that need to be quantized and transmitted via an additional backhaul between the involved cells to central processing points will be a non-negligible issue [15]. Its ability determines the difficulty of sharing EH and information signals.

In an MDAS, each of the transmission points may serve different UEs, achieving the benefits of cell splitting while keeping mutual interference inside the cell. There may be advantages if some of the transmission points collaboratively serve a single UE, while others collaborate to serve different UEs.

In order to accommodate more users, more effort is required to see how MDAS-SWIPT beamforming can work in multicell with OCI. Note that no more orthogonal pilot sequences can be employed in neighboring cells in the setup of an MDAS than in massive MIMO systems. Reusing the same set of orthogonal pilot sequences in all cells causes pilot contamination. In the downlink, this implies that the beamformed SWIPT signals in one cell will jam users in the neighboring cells. Such strong directional interference, unlike ICI, will not disappear even if the number of BS antennas increases to infinity. Therefore, in designing SWIPT beamforming, one should balance all the positive and negative effects. To support multiple users with different WET and WIT requirements, beamforming design is believed to be key for feasible implementation of SWIPT.

Downlink WET Beamforming for Uplink WIT: In the above consideration of downlink SWIPT, the harvested energy is not used for information transmission. An emerging trend concentrates on the study of using wireless power to support wireless communications, a harvest-then-transmit protocol [7], which allows the energy receiver to also play a role as the information transmitter. In the application of an implanted device, the energy receiver transmits information using the harvested energy.

In centralized multi-antenna systems, there is a "doubly near-far" phenomenon, in which a UE farther from the BS receives less wireless energy than a nearer UE in the downlink and also has to transmit with more power in the uplink for reliable WIT. For a fair network, more energy should be allocated to the farther UEs to have the same performance as the nearer UEs, but this approach will yield poor energy efficiency. Fortunately, in MDAS, this is no longer a problem because RAUs are located geographically such that the distances to the UEs are averaged. The RF signal will experience near path loss in both the downlink and the uplink, and the energy efficiency will thus be greatly improved. Without the fairness issue, MDAS only concerns maximizing the WIT in the uplink via downlink WET beamforming.

Knowledge of CSI is essential for both downlink energy beamforming and uplink information decoding (ID) in MDAS. More accurate CSI contributes to higher efficiency of energy transfer and higher uplink information rate. Similarly, although a longer time duration for channel estimation leads to more accurate CSIT, it reduces the time available for downlink WET and uplink WIT, and hence less harvested energy and lower rate. Also, it should be pointed out that the duration for WET and the WET beamforming optimization affect both the energy and time available for WIT in the uplink. To maximize the WIT achievable rate, one needs to optimize the time for channel estimation, WET and WIT, and WET beamforming.

RAU Selection: The cost of the RF chains (which include PAs, microprocessers, and analog-to-digital converters, ADCs) associated with the RAUs might make MDAS an expensive technology. Moreover, another problem connected to distributed antenna systems is the large amount of information that has to be gathered from or conveyed to involved receive or transmit antennas, requiring a large backhaul infrastructure between the cells, with a cost that is also a limitation for MDAS implementation. RAU selection is an effective approach to solve these problems while preserving the diversity and multiplexing gains. It can also reduce the signaling overhead and the cost to operate radio over fiber transmissions. Indeed, in [12], selecting a subset of RAUs for transmission was found to improve the performance of the blanket transmission. Although it decreases performance of full transmission, it greatly reduces the hardward costs.

In SWIPT, RAU selection should take the energy transfer efficiency metric into consideration to balance the energy efficiency for WIT and WET. Under the MDAS-SWIPT architecture, the RAUs can be classified into three groups: EH group, ID group (conventional RAUs), and EH-ID group. When a UE in the network requires energy and/or information, it issues a request to the BS, which begins a protocol to seek to which RAU groups to transfer the energy and to which to transfer the information. The MDAS switches the RAUs modes (i.e., wake or sleep) dynamically according to the needs and location information of the UEs. In other words, the MDAS will select those RAUs that are good for the active UEs and put others to sleep to save power. As shown in Fig. 3, some RAUs send only energy, some only information, and some both, representing the three groups. Coordinating the different applications (EH and/or ID) of UEs, the BS selects the best RAUs to deliver the corresponding functions.

User-Centric Coverage: In an MDAS, each of the transmission points may serve different UEs, achieving the benefits of cell splitting while keeping mutual interference inside the cell. There may be advantages if some of the transmission points collaboratively serve a single UE, while others collaborate to serve different UEs. This user-centric architecture, as opposed to being cell-centric, will be better in terms of routing information and energy flows with different priorities and purposes toward different sets of UEs within the network. A given UE should be able to communicate by exchanging information and energy flows through several possible sets of RAUs (including the RAUs from other cells). In other words, the set of RAUs providing connectivity to a given device (i.e., UE) and the functions of these nodes in a particular session (i.e., the RAUs send energy or information or both) should be tailored to that specific device and session. Under this vision, the problem is then split into multiple subproblems by a user-centric cluster, and each cluster's precoding and power control are managed in parallel for high energy efficiency. To do so, the concepts of up/downlink and control/data/energy channel should all be revisited. At the cell edge, the RAUs from other cells that have good channel gains can be selected to serve the UEs in the current cell, tackling the problem of poor cell edge performance and

providing uniform performance for all users, but this will burden the cost of more complex backhaul management.

Receiver Circuitry

Since RF signals that carry energy can at the same time be used as a vehicle for transporting information, theoretically both EH and ID can be performed from the same RF signal. This concept allows the information receiver and RF energy receiver to share the same RAUs. The receiver circuitry should simultaneously take energy and information, as illustrated in Fig. 4a. The antenna can work on multiple frequency bands in order to harvest energy from in-band and out-of-band energy sources. Impedance matching is achieved by a resonator circuit operating at the designed frequency to maximize the energy transfer. The energy that can be utilized should be converted from RF signals (AC signals in nature) to DC voltage by a rectifier circuit. Power management can adopt harvest-use (HU), harvest-store-use (HSU), and harvest-use-store (HUS) strategies [13, references therein] to optimize the use of incoming energy, coordinated by the energy storage. However, current designs are not yet able to extract RF energy directly from the decoded information carrier. In other words, the energy is lost during ID, and a new receiver architecture has yet to be designed.

Receiver Architecture

Power sensitivity varies for ID and RF EH (e.g., –10 dBm for EH vs. –60 dBm for ID) [8].

Receiver architectures can be broadly divided into two types:

- Co-located receivers: This means that the EH receiver and ID receiver share the same receiver antenna. This architecture can further be divided into two models,
  –Time switching (Fig. 4b1)
  –Power splitting (Fig. 4b2)
  The focus of the time switching architecture is to coordinate the time for periodically switching the receiving antenna of each UE between the EH receiver and ID receiver. In contrast, for the power splitting architecture, the received RF signal is split into two separate signal streams with different power levels, respectively, to send to the EH receiver and the ID receiver.
- Separate receivers: This architecture allows the EH receiver and the ID receiver to be equipped with separate dedicated antennas (Fig. 4b3), which means that the two receivers can function independently. Although this can maximize the utilization of the RF signal, the size of the device will be larger than that of the co-located receivers.

Note that with the explosive growth in wireless users, the applications of the users will be very different. At any given time, a user may only require energy or data service, or both services simultaneously. This results in an investigation of the operation policy to find an optimal ratio to split the time or received RF signals of each user, and also the design of different trade-offs between WIT and WET.

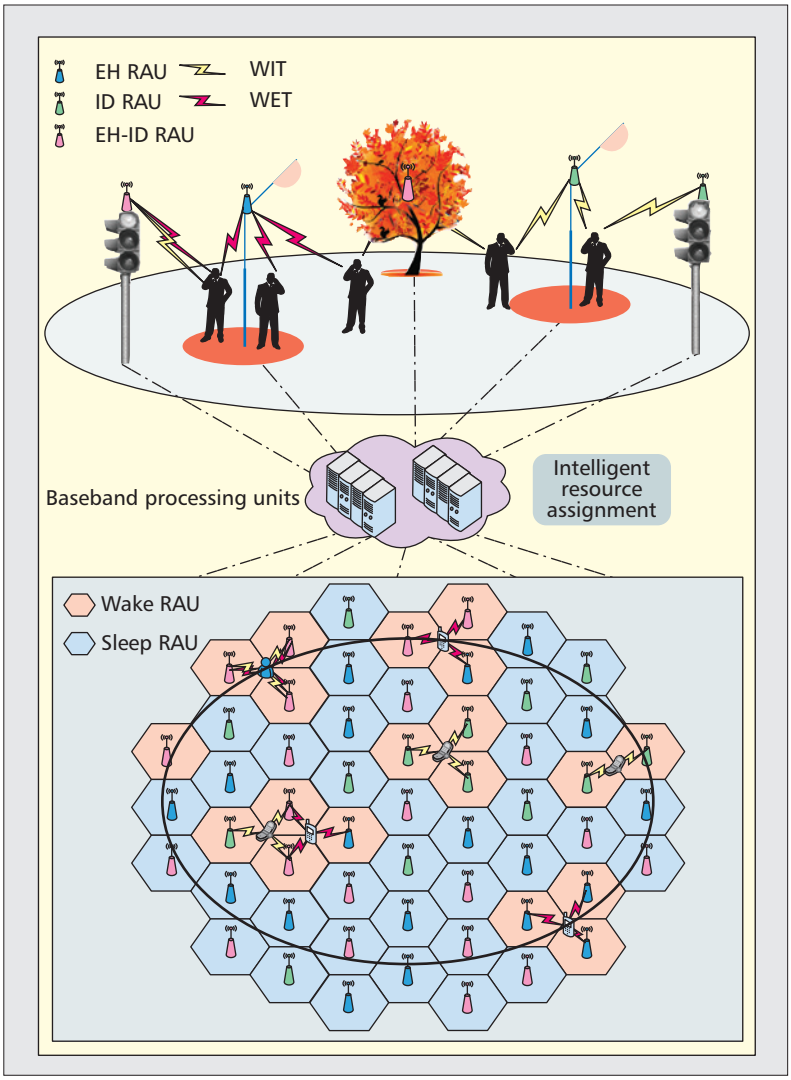


Figure 3. Advanced RAU selection and user-centric coverage.

Future Directions

Interference Management

Simulcast transmission, where the antennas transmit identical signals on the downlink, was one of the first MDAS transmission strategies. Simulcast MDASs were found to improve the signal-to-interference ratio (SIR), need fewer handoffs, and allow smoother scaling of network infrastructure costs with the amount of traffic. Unfortunately, these benefits came with a cost that includes increasing the total amount of interference in the network, which complicates cell design.

Intelligent transmission strategies can reduce the impact of excess interference. If transmissions are designed to minimize the transmit power with a fixed channel quality, MDAS can reduce the OCI relative to the centralized approaches. For the ID receiver, lower interference means higher SIR and higher data rate, but for the EH receiver, the interference signals are additional energy sources. With EH, harmful interference can be turned into useful energy through a scheduling policy. As a result, how to mitigate interference as well as facilitate WET is an important problem.

Recent research has recognized that optimizing SWIPT brings tradeoff on the wireless system design [8]. As a consequence, the amount of WET and WIT generally cannot be maximized at the same time. This calls for a redesign of transmission strategies in MDAS.

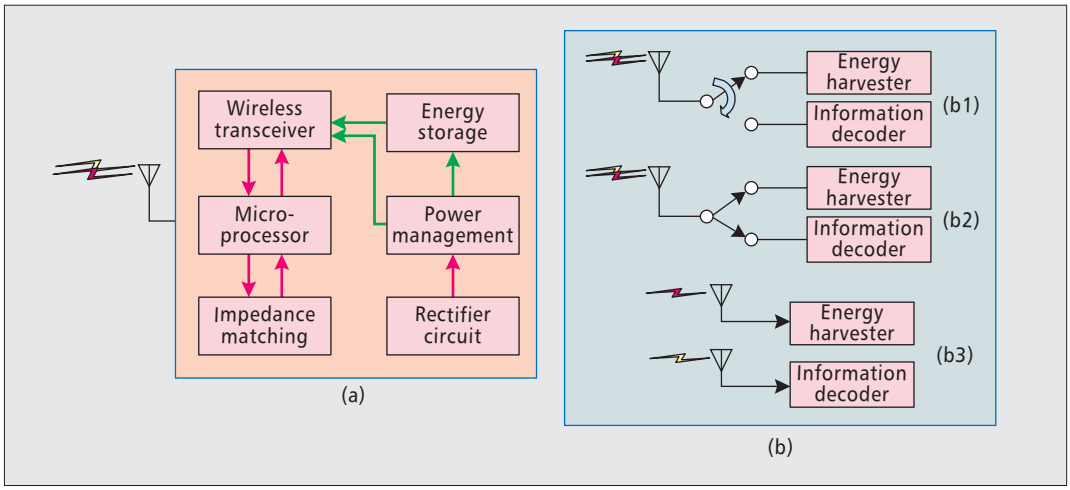


Figure 4. A block diagram of the receiver circuitry and architectures.

Recent research has recognized that optimizing SWIPT brings a trade-off on the wireless system design [8]. As a consequence, the amount of WET and WIT generally cannot be maximized at the same time. This calls for a redesign of transmission strategies in MDAS.

Power Allocation

Another major issue in MDASs is transmit power allocation for the downlink, which is usually necessary due to the imbalance of the microscopic and macroscopic fading between each RAU and the UE. In centralized multi-antenna systems, the optimal power allocation is the water-filling (WF) solution, and equal power allocation (EPA), which maximizes the ergodic capacity, will degrade significantly when the distances between each RAU and the reference UE are quite different. Therefore, it is necessary to develop an efficient power allocation scheme for MDASs. As the signals transmitted from different RAUs experience different macroscopic fading, efficient power allocation over the RAUs will significantly enhance the signal-to-noise ratio (SNR) at the UEs and the capacity or diversity gain of the MDAS [14]. Earlier, in [15], the authors proposed the maximal path loss ratio (MPR) approach to allocate more transmit power to the RAUs having larger channel gains. However, [15] did not include the metric of WET. In MDAS-JWIET, the trade-off between WIT and WET restricts the power allocation problem. In the harvest-then-transmit protocol, the double gains phenomenon will encourage more transmit power to be allocated to the RAU of larger channel gain than the MPR approach. When the user-centric architecture is considered, the power allocation strategy becomes much more complicated.

Resource Scheduling and Performance Evaluation

Due to the complex nature of MDAS-JWIET, intelligent resource management is needed. Emerging technologies such as software-defined networking (SDN) and network function virtualization (NFV) can be used for applying intelligence for WIT and WET in MDASs. Both technologies are anticipated to have a significant impact on forthcoming research.

Regarding the evaluation of MDAS-JWIET networks, a fair assessment is only possible by considering a range of performance metrics including the spectral efficiency, energy efficiency, delay, reliability, user fairness, implementation complexity, and so on. A unified framework should be developed for characterizing the trade-off among all the metrics. This will require high-complexity joint optimization algorithms and long simulation times.

Performance Comparison

In this section, we show the performance comparison of WET and WIT in a single cell respectively, as shown in Figs. 5 and 6. The metric for the WET is the average received power (i.e., harvesting energy per UE), while the metric for the WIT is sum rate among all UEs.

Parameter Setting — The RAUs are uniformly distributed in the cell varying from 50 to 600, serving users placed in a series of circle radii (i.e., 15, 40, 65, 90), and each circle has three users assuming equal distance, that is, a total of 12 users. The carrier frequency is 5 GHz, and the noise power is –70 dBm. We assume the Rayleigh fading of the propagation path (i.e., $h_{ij} \sim \mathcal{CN}(0, 1)$), and set the large-scale fading to be $\beta_{ij} = 10^{-3} d_{ij}^{-3}$, where the path loss exponent is 3 and d_{ij} is the distance between the ith RAU and jth UE. For a fair comparison, the total number of RAUs, total transmit power, and coverage area are assumed to be the same for all cases.

WET Performance — We consider 10 scenarios including the centralized antenna systems (CAS) broadcast (BC)/matched-filter beamforming (MFBF), distributed antenna systems (DASs) BC/MFBF, user-centric (UC) BC/MFBF (which means the RAUs within $area = \{x, y \in \mathbb{R}^2 : |y - x| \leq 10\}$ are activated to serve the UEs where x and y is the location of the user and RAU respectively, while RAUs outside the area are switched to sleep mode), the nearest one RAU selection (NRS-1) BC/MFBF, and the nearest three RAUs selection (NRS-3) BC/MFBF. It

can be seen from Fig. 5 that there is a large gain in average received power by deploying DASs over CASs, which means that the transmit power efficiency of DASs is larger than that of CASs as we expected. When the system broadcasts the energy signal, full transmission (i.e., DAS-BC) performs worse than the UC-BC and RAU selection cases (i.e., NRS-1-BC and NRS-3-BC), and the performance decreases with the the number of selected RAUs. For the beamforming cases, the situation is the other way around. Although the optimal beamforming is not used (we assume equal power for each user), all the scenarios still benefit a lot from the beamforming. The gap between BF and BC is gradually reduced vs. the number of active RAUs. In particular, the minimum gap scenario is NRS-1, where only 12 RAUs are active, while the maximum gap scenario is DAS, where all the RAUs are active. The UC scenario performs poorly at the beginning, and becomes better when the RAU number is large, accounting for that the RAUs which can be selected by the users are few or even none at the very start and becomes more with the increasing of the total RAU number. Overall, the harvested energy in the DASs is higher than the CASs, demonstrating that MDASs are more suitable for WET.

WIT Performance — Five scenarios are considered, including CAS-MFBF, DAS-MFBF, UC- MFBF, NRS-1-MFBF, and NRS-3-MFBF. In terms of sum-rate, all the scenarios provided growth with increasing number of RAUs, where the information transfer of the DAS-MFBF is superior to CAS-MFBF vs. all numbers of RAUs because of less correlation of the antennas with each other. Moreover, selecting a subset of RAUs for transmission was found to decrease the performance of full transmission (i.e., DAS-MFBF [12]). This inferiority will be improved with increasing the selected subset size of RAUs, as seen from NRS-1-MFBF and NRS-3-MFBF in Fig. 6. Although the performance in the selected cases is not as good as DAS-MFBF, they can reduce the signaling burden overhead on the BPU, the cost to operate radio over fiber transmissions, and RAU processing. This is due to the fact that the RAUs that are not selected can be switched to sleep mode. In particular, UC-MFBF performs worst at the beginning, accounting for the fact that when the RAU number is low, all the RAUs may be located out of the selected area of the user; hence, none of the RAUs can be selected by the user. It will improve with the RAU number increasing, eventually being better than NRS-1-MFBF, resulting in more than one RAU being selected by the user. In summary, full transmission has the best overall performance for the DAS architecture considered in this article.

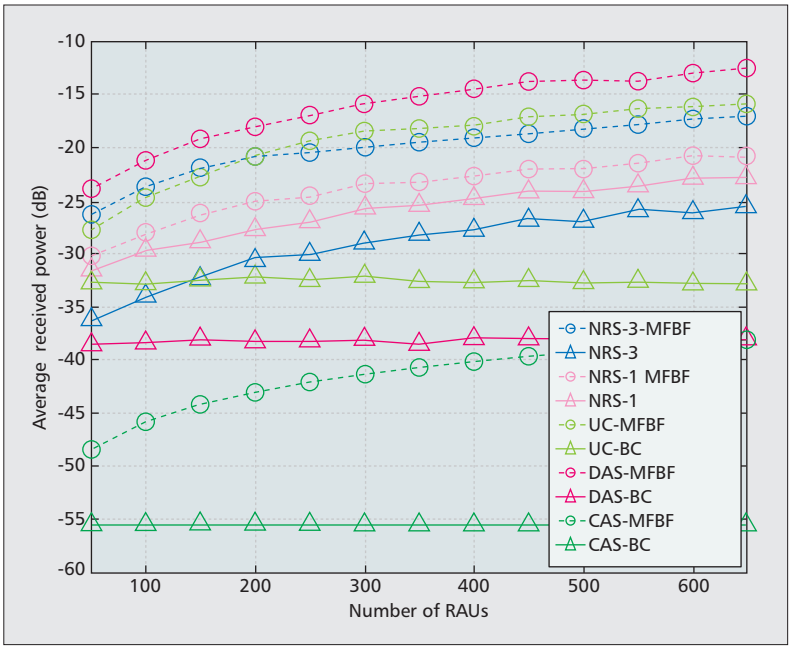


Figure 5. Wireless energy transfer performance vs. the number of RAUs under different scenarios.

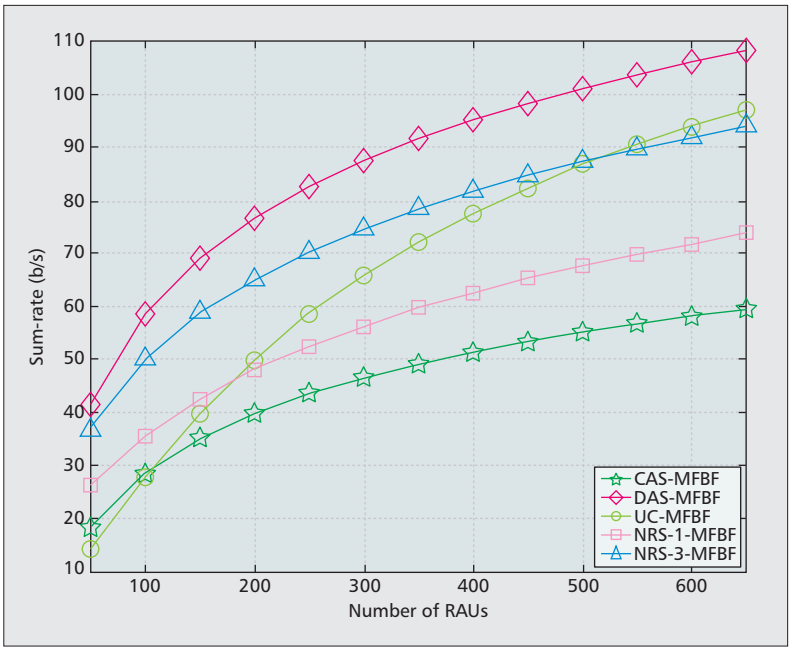


Figure 6. Wireless information transfer performance vs. the number of RAUs under different scenarios.

Conclusion

In this article, the concept of joint wireless information and energy transfer (JWIET) in MDASs has been proposed to improve meeting the current mobile network demand. We have discussed the key architecture designs in both the transmitter and receiver ends, and focused on technologies that could lead to the development of both architectures for the improvement of WET and WIT:

- At the transmitter end, transmitter circuitry and distributed energy beamforming (SWIPT, energy beforming, RAU selection, user-centric)
- At the receiver end, receiver circuitry and receiver architecture (co-located time switching or power splitting design, and separate receiver design)

Also, various potential issues that should be rethought have been pointed out to satisfy the expected performance requirements, such as

Recent research has recognized that optimizing SWIPT brings a trade-off on the wireless system design. As a consequence, the amount of WET and WIT generally cannot be maximized at the same time. This calls for a redesign of transmission strategies in MDAS.

interference management, power allocation, and resource scheduling. Numerical results presented in this article show that the MDAS is a very promising network architecture for future information and energy transfer systems.

Acknowledgment

The authors would like to thank the anonymous reviewers for providing very detailed suggestions for revision, which have considerably helped to improve the presentation of the article. This material is supported by the National Basic Research Program of China (973 Program) under Grant no. (2013CB329005) , the National Natural Science Foundation of China under Grant no. (61271237, 61222102, 61401235), and the Natural Science Foundation of Jiangsu Province under Grant no. (BK2012021).

References

[1] S. Z. Chen and J. Zhao "The Requirements, Challenges, and Technologies for 5G of Terrestrial Mobile Telecommunication," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 36-43.
[2] E. G. Larsson *et al.*, "Massive MIMO for Next Generation Wireless Systems," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 186–95.
[3] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and Spectral Efficiency of Very Large Multiuser MIMO Systems," *IEEE Trans.Commun.*, vol. 61, Apr. 2013, pp. 1436–49.
[4] W. C. Brown, "The History of Power Transmission by Radio Waves," *IEEE Trans. Microwave Theory Tech. 1984*, 32, 1230–42.
[5] K. Huang and V. K. N. Lau, "Enabling Wireless Power Transfer in Cellular Networks: Architecture, Modeling and Deployment," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, Feb. 2014, pp. 902–12.
[6] L. Liu, R. Zhang, and K. C. Chua, "Multi-Antenna Wireless Powered Communication with Energy Beamforming," *IEEE Trans. Commun.*, vol. 26, no. 12, Nov. 2014, pp. 4349–61.
[7] G. Yang *et al*, "Throughput Optimization for Massive MIMO Systems Powered by Wireless Energy Transfer," accepted for publication, *IEEE JSAC*, Jan. 2015, 10.1109/JSAC.2015.2391835.
[8] R. Zhang and C. K. Ho, "MIMO Broadcasting for Simultaneous Wireless Information and Power Transfer," *IEEE Trans. Wireless Commun.*, vol. 12 , no. 5, May 2013, pp. 1989–2001.
[9] X. Chen, X. Wang, and X. Chen, "Energy-efficient Optimization for Wireless Information and Power Transfer in Large-Scale MIMO Systems Employing Energy Beamforming," *IEEE Wireless Commun. Letters*, vol. 2, no. 6, Dec. 2013, pp. 667–70.
[10] K. T. Truong and R. W. Heath, "The Viability of Distributed Antennas for Massive MIMO Systems," *Asilomar Conf. Signals, Systems and Computers*, Nov. 2013, pp. 1318–23.
[11] J. Zhang *et al.*, "On Capacity of Large-scale MIMO Multiple Access Channels with Distributed Sets of Correlated Antennas," *IEEE JSAC*, vol. 31, no. 2, Feb. 2013, pp. 133–48.
[12] R. W. Heath *et al.*, "Multiuser MIMO in Distributed Antenna Systems with Out-of-Cell Interference," *IEEE Trans. Signal Processing*, vol. 59, no. 10, Oct. 2011, pp. 4885–99.
[13] F. Yuan *et al.*, "Optimal Harvest-Use-Store Strategy for Energy Harvesting Wireless Systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, Sept. 2014, pp. 698–710.
[14] S. Mahboob, C. Mahapatra, and V. C. M. Leung, "Energy-Efficient Multiuser MIMO Downlink Transmissions in Massively Distributed Antenna Systems with Predefined Capacity Constraints," *Proc. Broadband, Wireless Computing, Commun. and Applications*, Victoria, BC, Canada, Nov. 2012, pp. 208–11.
[15] X.-H. You *et al.*, "Cooperative Distributed Antenna Systems for Mobile Communications," *IEEE Wireless Commun.*, vol. 17, no. 3, Jun. 2010, pp. 35–43.

Biographies

FANGCHAO YUAN (dylanyfc@163.com) received his B.S. degree in communications engineering from Nanjing University of Posts and Telecommunications (NJUPT), China, in 2011. He is currently working toward his Ph.D degree in communications and information systems at NJUPT. His research interests include green communications, joint wireless information and energy transfer, and energy harvesting wireless communications.

SHI JIN (jinshi@seu.edu.cn) received his B.S. degree in communications engineering from Guilin University of Electronic Technology, China, in 1996, his M.S. degree from NJUPT in 2003, and his Ph.D. degree in communications and information systems from Southeast University, Nanjing, in 2007. He is currently with the faculty of the National Mobile Communications Research Laboratory, Southeast University. His research interests include space time wireless communications, random matrix theory, and information theory.

YONGMING HUANG (huangym@seu.edu.cn) received his Ph.D. degree in electrical engineering from Southeast University. During 2008–2009, he visitied the Signal Processing Lab, Electrical Engineering, Royal Institute of Technology (KTH), Stockholm, Sweden. He is currently a professor at the School of Information Science and Engineering, Southeast University. His current research interests include space-time wireless communications, cooperative wireless communications, energy-efficient wireless communications, and optimization theory.

KAI-KIT WONG (kai-kit.wong@ucl.ac.uk) received his B.Eng., M.Phil., and Ph.D. degrees, all in electrical and electronic engineering, from Hong Kong University of Science and Technology, Hong Kong, in 1996, 1998, and 2001, respectively. Since August 2006, he has been with University College London, first at the Adastral Park Campus and at present the Department of Electronic and Electrical Engineering, where he is a reader in wireless communications. His current research interests include game-theoretic cognitive radio networks, cooperative communications, multiuser communications theory, physical-layer security, massive MIMO, and energy-harvesting wireless communications.

Q. T. ZHANG [F] (qtzhang@ieee.org) received his B.Eng. from Tsinghua University, Beijing, and M.Eng. from South China University of Technology, Guangzhou, China, both in wireless communications, and his Ph.D. degree in electrical engineering from McMaster University, Hamilton, Ontario, Canada. He is now an Honorary Professor with NJUPT. His research interest is in wireless communications, currently focusing on wireless MIMO, 5G cellular, and green communications.

HONGBO ZHU (zhuhb@njupt.edu.cn) received his B.S. degree in communications engineering from NJUPT and his Ph.D. degree in information and communications engineering from Beijing University of Posts and Telecommunications in 1982 and 1996, respectively. He is presently working as a professor and vice-president at NJUPT. His research interests include mobile communications, wireless communication theory, and electromagnetic compatibility.

When Telecommunications Networks Meet Energy Grids: Cellular Networks with Energy Harvesting and Trading Capabilities

Davide Zordan, Marco Miozzo, Paolo Dini, and Michele Rossi

ABSTRACT

In this article, we cover eco-friendly cellular networks, discussing the benefits that ambient energy harvesting offers in terms of energy consumption and profit. We advocate for future networks where energy harvesting will be massively employed to power network elements; even further, communication networks will seamlessly blend with future power grids. This vision entails the fact that future base stations may trade some of the excess energy they harvest so as to make a profit and provide ancillary services to the electricity grid. We start by discussing recent developments in the energy harvesting field, and then deliberate on the way future energy markets are expected to evolve and the new fundamental trade-offs that arise when energy can be traded. Performance estimates are given throughout to support our arguments, and open research issues in this emerging field are discussed.

INTRODUCTION

Energy efficiency in cellular networks is becoming a key requirement for network operators to reduce their operative expenditure (OPEX) and mitigate the footprint of information and communications technologies (ICT) on the environment. Costs and greenhouse gas emissions due to ICT have grown in the last few years due to the escalation of traffic demand from mobile devices such as smartphones and tablets. Cloud-based and Internet of Things (IoT) services are expected to exacerbate this negative trend [1].

Network designers have been addressing this by considering hierarchical cell structures through so-called heterogeneous networks (HetNets), where small cells (i.e., micro or pico) are deployed to increase network coverage and capacity, and decrease overall energy consumption [2].

In this article, we deliberate on scenarios where harvested ambient energy is employed to steer HetNets toward *nearly zero* energy consumption and, more than that, where communication networks blend with future electricity grids. By nearly zero we mean that, in the long run, the monetary costs incurred in operating the network are counterbalanced by the revenue from either grid energy savings or energy trading. This vision means that future network elements may trade some of the energy they harvest to make a profit and provide ancillary services to the power grid. In pico deployments, for instance, this may occur in the form of supporting connected loads, such as street lighting or weather stations. Instead, selling energy to the grid operator may make sense for micro and macrocells where the amount of energy harvested easily matches or surpasses that required by residential users.

Among other harvesting technologies, solar power is deemed the most appropriate due to the good efficiency of commercial photovoltaic (PV) panels [3] as well as the wide availability of solar energy for typical installations. The idea of using solar harvesters to power base stations (BSs) has been around since 2001, starting with second generation (2G) technology [4]. However, initial studies focused on rural scenarios, where connectivity to the power grid is impracticable. In the last few years, thanks to the reduced power constraints of small BSs and the reduction of the costs of energy harvesting technologies, PV sources are also becoming appealing for urban scenarios, as testified by the vivid literature on this topic [5]. The exploitation of other types of renewable energy, such as wind, is also possible. Due to limited space, in this article we focus on PV technology, but we remark that most of the following discussion applies as well to other energy sources.

The viability of a self-sustainable small cell network was discussed in [6], and can be achieved through the addition of rechargeable batteries and energy harvesting devices, such as solar panels. This makes it possible to reduce the OPEX concerning the cost associated with the purchase of energy from the electricity grid.

Davide Zordan and Michele Rossi are with the University of Padova.

Marco Miozzo and Paolo Dini are with Centre Tecnológic de Telecomunicacions de Catalunya.

0163-6804/15/$25.00

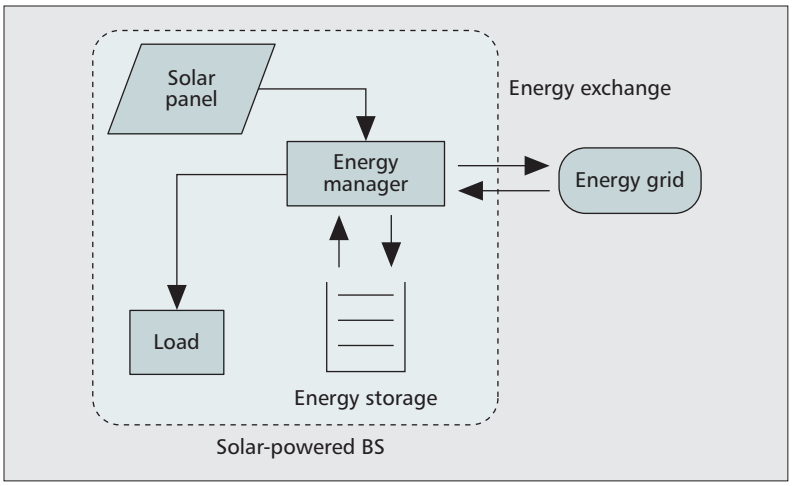


Figure 1. Reference diagram for a solar-powered BS.

These facts were analytically studied in [7] to identify availability regions for uncoordinated BS operating strategies, where the network is energetically self-sufficient. Coordinated control algorithms for hybrid powered architectures are discussed in [8], where the consumption from the electricity grid is reduced through power control strategies and load balancing. Sleep mode control has been investigated in [9] to regulate the BS transmission power.

Here, we approach HetNet design from a fundamentally different angle. Following [10], we advocate for nearly zero energy networks where the energy, besides being used for self-powering [6–9], can also be traded with the electricity grid or other network elements. Moreover, we foresee that these HetNets will operate within an energy market where the price of energy changes hourly and is set a day ahead (a contemporary example for this market is discussed below). This allows for several new optimizations related to the way energy is consumed, purchased, and possibly sold to the grid operator. The nature, benefits, and open challenges related to these optimizations constitute the main objective of this article. We stress while energy harvesting technology for BSs has been investigated in the literature, energy trading with the smart grid entails a new design philosophy for energy management and networking algorithms, which should interact with or be aware of smart grid optimization procedures.

In the following sections, we start discussing the requirements in terms of panel and battery size where BSs are operated off-grid and where they are grid-connected. We examine the related tradeoffs using realistic traces for the harvested energy, the price of the energy purchased from and sold to the grid, and the users' demand profile.

Thus, we analyze whether the capital expenditure (CAPEX) of providing these BSs with the needed solar add-on can be amortized and how long it takes to return the initial investments.

Finally, we present open research challenges concerning the integration of the foreseen BSs into future energy grids. These involve the study of a system of systems that arises from the combination of telecommunications networks and electricity grids. According to this vision, telecommunications networks may become active players in the electricity market, providing ancillary services for the energy grid and/or support to connected loads [11]. These aspects are currently unexplored and require a great deal of work, both theoretical and applied.

Reference Framework

In Fig. 1 we show our reference diagram for a solar-powered BS. The solar energy is harvested through a solar panel. A management module orchestrates the use of this energy and, in particular, decides whether to accumulate it in local energy storage or sell it to the energy grid. The load corresponds to the BS elements in charge of managing the transmission and reception activities, and their power consumption can be modulated as a function of the users' demand, as we detail shortly. Energy can be purchased from the electricity grid when the energy stored in the energy storage and that harvested is not sufficient to satisfy the current cell's load, but can additionally be sold if the energy inflow from the solar module is abundant. Similarly, energy can be locally stored, for example, when it is not convenient to sell it (e.g., when the energy price is high) and can subsequently be used to power the BS or sold to the grid operator.

For the moment, we assume that energy can be "sold" without specifying the nature of this action (i.e., whether energy is injected into the power grid or used to power a connected load). We elaborate on this below, where we identify open research challenges. For energy storage, in the following discussion we consider lithium ion rechargeable batteries, which is the technology of choice at the time of writing. Nevertheless, we note that other options are viable, such as molten salt batteries and fuel cells.

A detailed energy consumption model of Long Term Evolution (LTE) BSs has been proposed in the context of the European project EARTH [2], where the main sources of energy consumption have been identified and evaluated. The power consumption, P_c, of different types of BS considerably differs, ranging from about 1 kW for a macro BS to about 100 W for micro and 10 W for pico (at full system load). Following [2], for P_c we consider a linear dependence on the (normalized) load factor ρ,

$$P_c(\rho) = P_0 + \alpha\rho, \ \rho \in [0,1], \quad (1)$$

where P_0 W is the minimum power consumption for the BS. Denoting the parameter vector by $\overline{p} = (P_0, \alpha)$, for macro, micro, and pico cells, we have $\overline{p}_{\text{macro}} = (750,600)$, $\overline{p}_{\text{micro}} = (105.6,39)$, $\overline{p}_{\text{pico}} = (13.6, 1.1)$, respectively.

In the following sections, we discuss suitable models for the hourly price of energy, the amount of energy harvested, and the demand profile.

Hourly Price of Energy and Energy Market

Most likely, the energy price in future power grids will change hourly. This practice is not yet adopted worldwide, but there are relevant pro-

grams that already use it. A relevant example can be found in Illinois, where electrical companies offer new hourly electricity pricing programs where energy prices are set a day ahead by the hourly wholesale electricity market run by the midcontinent independent system operator (MISO). In this way, customers can optimize their usage patterns, saving money on their energy bills. In this article, we use publicly available historical energy price data from these programs to discuss suitable energy management policies.

In Fig. 2, we show a typical profile of the hourly energy prices for the first week of November 2010. The price dynamics follow a regular pattern with a bimodal shape within each day. Note that the price significantly increases during peak hours. In the summer months the price shows a different behavior, (i.e., it is bell-shaped with a single maximum around midday). This is due to the impact of air conditioning, which is heavily used during the warmest hours.

A further important observation is in order. Up to now, communications networks have been mostly optimized for communication performance or in terms of energy efficiency, which entails a diminished cost in purchasing the energy required for operating the communications apparati. However, energy harvesting and future market policies will permit at least two additional optimization strategies. First, the system could adapt its behavior to the energy price (i.e., it could be energy frugal when the energy cost is high, while adopting more aggressive policies when the cost drops). Second, part of the energy that is accumulated could be sold or redistributed among other network elements. As shown in Fig. 2, the peak in harvested energy may not always occur when the price is maximum. Thus, as we show below, it may be worth temporarily saving the harvested energy to sell later, thus achieving a higher revenue.

These facts may revolutionize the way we design communications systems, going from a solely communication-performance-oriented approach to an energy-market-oriented one.

Harvested Energy and Demand Profiles

Hourly energy generation traces from a solar source have been obtained for the cities of Los Angeles, California, and Chicago, Illinois. For the solar modules, we have considered commercially available Panasonic N235B photovoltaic (PV) technology. These panels have single-cell efficiencies as high as 21.1 percent, which ranks them among the most efficient solar modules at the time of writing, delivering about 186 W/m^2. For the results discussed in this article, raw irradiance data were collected from the National Renewable Energy Laboratory [12] and converted, accounting for this solar power technology, into harvested energy traces using the SolarStat tool of [13].

We note that energy harvesting traces are generally bell-shaped with a peak around midday, whereas the energy harvested during the night is negligible. We also note, as discussed in [13], that there may be high variability in the energy harvested during the day, and this also holds for the summer months. This means that although the energy inflow pattern can be known to a certain extent, intelligent and adaptive algorithms that make their decisions based on current and past inflow patterns as well as predictions of future energy arrivals have to be designed. While these are being extensively studied by the wireless sensor networks community, much still has to be done for cellular systems.

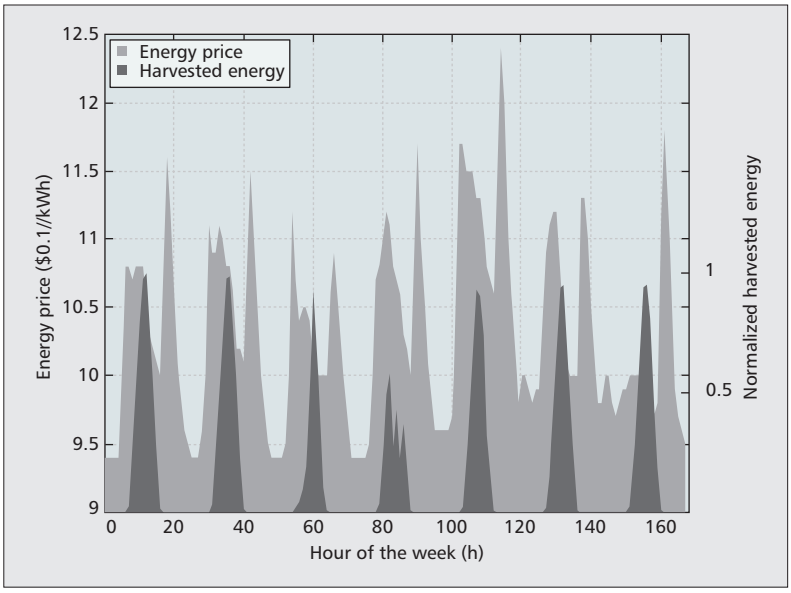


Figure 2. Hourly energy price and harvested energy for the state of Illinois during the first week of November 2010 (Monday goes from hour 0 to hour 23, Sunday goes from hour 145 to hour 168). The maximum harvested energy during the week is 123 Wh for a solar panel of 1.2 m^2.

For the demand profile, it is commonly accepted and confirmed by measurements that the energy use of base stations is time-correlated and daily periodic. In this article, we use the load profiles obtained within the EARTH project [14, p. 25].

Off-Grid Deployments of Solar-Powered Base Stations

Next, we consider the setup of Fig. 1, where a BS collects renewable energy from a solar panel and uses it for self-powering. Some of the excess energy can temporarily be accumulated in a local energy buffer (battery) and used at a later time when harvested energy is scarce or none. The BS operates off-grid, and the above models account for the energy harvested and cell load. Here, we are concerned with the right sizing of solar panel and battery so that the BS can operate perpetually without having to buy energy from the electricity grid.

By *outage probability* we refer to the fraction of time during which the BS is unable to serve users' demands due to an insufficient energy reserve. In that case, the BS has to be momentarily switched off or put into a power saving mode. The contour plot for the outage probability for microcells is shown in Fig. 3 considering solar traces from Los Angeles. Different

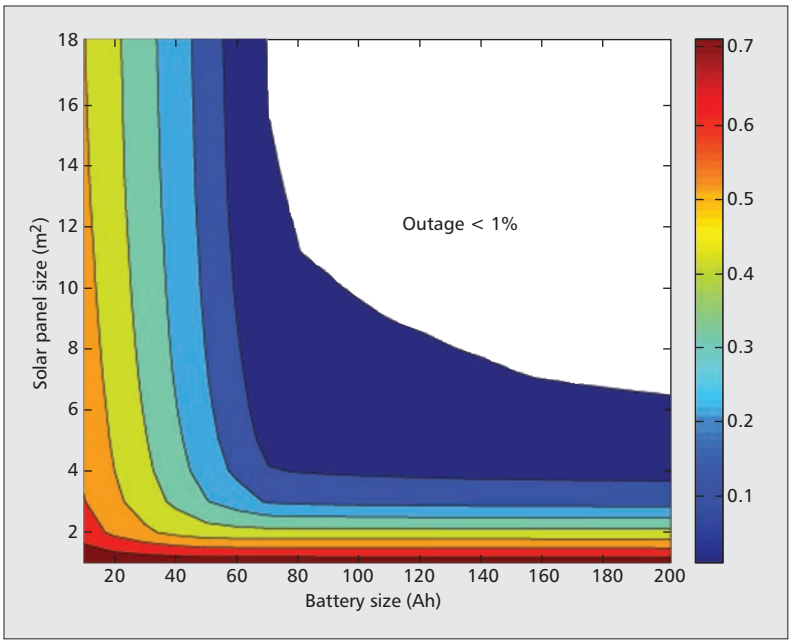


Figure 3. Contour plot of the outage probability for a micro cell operated off-grid (battery voltage is 24 V). Different colors indicate outage probability regions, whose maximum outage is specified in the color map on the right of the plot.

colors are used to indicate outage probability regions (maximum outages are specified in the associated color map). The white-filled area indicates the parameter region where the outage probability is smaller than 1 percent. The outage probability graphs for pico and macrocells are not given due to space constraints; the corresponding plots show a similar trend, rescaled to higher (macro) or smaller (pico) values along both axes.

From Fig. 3, we see that panels of size smaller than 15 m^2 and battery capacities of at most 150 Ah at 24 V suffice for microcells. For pico and macro deployments, solar panels range in size from 0.7 to 1.4 m^2 (pico) and from 40 to 80 m^2 (macro), and battery capacities from 20 to 90 Ah at 12V (pico) and from 300 to 1500 Ah at 48 V (macro). Taking an outage of 1 percent as our design parameter, all the points on the boundary of the white-filled region are equally good. If we further consider the harvesting hardware cost (CAPEX), the best solution also depends on the revenue that can be accrued when the BS is grid-connected and energy can be traded, as we see below.

The results for the city of Los Angeles are rather good, indicating that nearly zero energy is indeed a feasible goal. In fact, both battery and panel sizes are acceptable given the dimensions of typical installation sites for the considered BSs. Instead, for the city of Chicago the energy inflow is less abundant, and this is especially so during the winter months. In that case, reasonable panel and battery sizes (even slightly higher than those discussed for Los Angeles) lead to outages of 10 percent or higher. Due to this, grid-connected operation is required for locations where the energy inflow is moderate (especially during the winter).

Grid-Connected Deployments: Energy Trading and Optimization

Energy Trading Policy

Now, we take hourly energy prices into account (Fig. 2) and consider a grid-connected BS. The system evolves in slotted time t, where the slot duration is 1 h. At any given time t, the BS may sell or buy a certain amount of energy e_t, which is positive when energy is sold and negative when purchased. When energy $e_t < 0$ is purchased from the grid operator, a monetary cost $C(e_t)$ is incurred, which corresponds to the price of energy in slot t. Instead, when energy $e_t > 0$ is sold, a reward $R(e_t) = rC(-e_t)$ is accrued, with $r \leq 1$ being a discount factor. This means that the energy sold is paid less than that purchased, as this is usually the case in current energy markets and is expected to remain so for future ones. Also, we use $C(e_t) = 0$ for $e_t \geq 0$ and $R(e_t) = 0$ for $e_t \leq 0$, meaning that no cost is incurred when selling and no reward is accrued when buying.

At each time t, the demand d_t has to be fully served, and the energy required to do so is harvested, taken from the battery, or bought from the grid. The energy manager of Fig. 1 intelligently chooses in which amounts and when energy e_t (the decision variable) has to be purchased or sold so that the system maximizes its profit. This corresponds to maximizing the total monetary reward, expressed as $f(T) = \Sigma_{t=0}^{T} [R(e_t) - C(e_t)]$ over the time horizon of interest $t \in \mathcal{T}$ (with $\mathcal{T} = \{0, 1, \ldots, \mathcal{T}\}$). The solution to this problem amounts to finding the optimal allocation $\{e_t^*\}_{t \in \mathcal{T}}$ for all time slots $\in \mathcal{T}$. Here, we do so through dynamic programming considering the above traces for hourly energy prices, user demand and harvested energy.

In Fig. 4, we show the optimal allocation e_t^* for the third week of November 2010 for the city of Los Angeles, considering a discount factor $r = 0.5$ for the energy sold and a microcell with a panel of 10 m^2 and a battery of 90 Ah (at 24 V). For the sake of illustration, the temporal traces of energy price ($0.01/kWh) battery state (Ah) and harvested current (Ah) are also shown. From these results, various interesting observations can be made.

During Monday and Wednesday, labels (a) and (c) in the figure, it is optimal to sell energy during the day; correspondingly, e_t^* shows two positive peaks. The first occurs because the energy inflow is abundant, and the price of energy is also high. Note that not selling in this case would imply discarding the excess energy as the battery size is insufficient to store it in full; this is inefficient and would lead to a loss of revenue. For the second (smaller) peak, the BS sells part of the energy previously accumulated, and this is done in correspondence with the second maximum for the price. However, not all the energy is sold, but a certain amount of it is retained to power the BS during the night, where the harvested energy inflow is null (d). For Tuesday (b), we have an additional small peak in e_t^* in the early morning as the price of energy is high and there is some residual energy in the battery. For

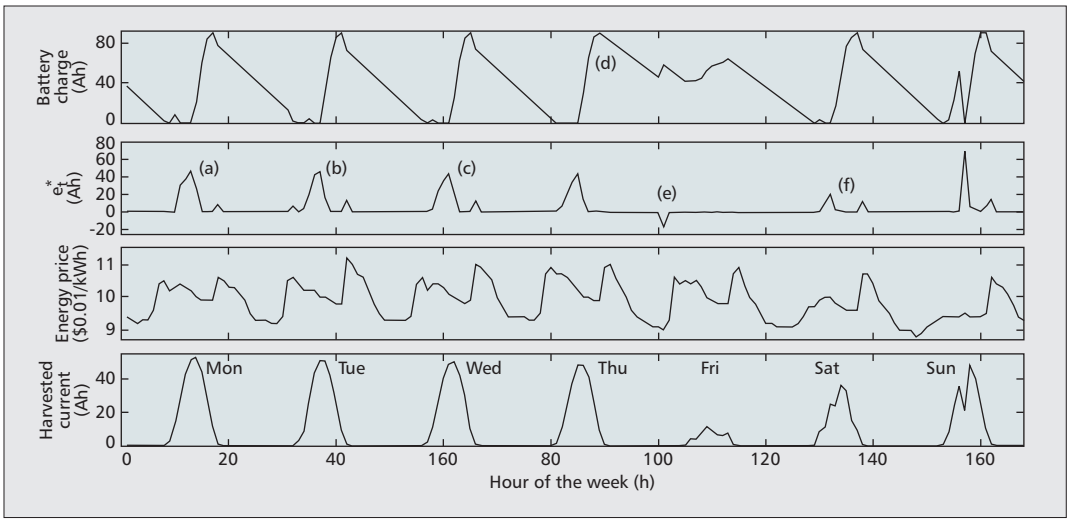


Figure 4. Optimal energy trading policy for the third week of November 2010 (Monday goes from hour 0 to hour 23, Sunday goes from hour 145 to hour 168). Price data has been taken from the Power Smart Pricing program. Energy harvesting traces are for the city of Los Angeles.

As one may expect, the actual sizing for the solar add-on depends on the energy selling price as well as on the location. Nevertheless, the rather good results that we have shown here are encouraging. These are due to the modest cost of PV technology, that has been plummeting over the last decade (10-fold reduction).

Thursday, labels (d) and (e), e_t^* shows a single positive peak: here, the inflow is so abundant that it cannot be entirely stored in the battery, and not selling it would imply a loss of revenue, as seen for (a) and (c). Moreover, there are no further selling events because the energy harvested during the next two days is modest. Thus, the optimal policy prefers to retain energy in the battery (d) to keep the BS operational. Note also (e) that some additional energy has to be purchased (see the negative peak in e_t^*), and this occurs during the night, where the price of energy is minimum. Finally, for Saturday (f) we again have two positive peaks, but the first is considerably smaller due to the smaller amount of energy harvested in this day.

Note that these results are obtained offline, given full knowledge of all processes for the considered time horizon. Online and predictive algorithms are still an open research topic.

Amortizing the CAPEX

Here, we additionally consider the solar panel size and battery capacity as optimization variables. Our task is to dimension the solar add-on in order to maximize the net profit, considering an amortization period T of 10 years and given that the optimal policy e_t^* is used throughout. The net profit over this period is obtained summing the revenue $f(T)$ to the cost incurred when the BS is powered in full by the energy grid, and subtracting the CAPEX associated with the resulting harvesting hardware.

For the following example results, we have accounted for the current price of solar panels, which is about \$0.5/kWh and a battery cost of \$300/kWh. Table 1 shows the 10-year net income for pico, micro, and macrocells. According to the considered CAPEX cost, optimal designs tend to pick smaller battery capacities and invest more in solar modules. In the table, two designs, D1 and D2, are shown for each type of BS, where D2 returns the maximum net profit within the considered parameter range. Notably, a positive income is accrued in almost all cases. As expected, Los Angeles allows for higher revenues due to the more abundant energy inflow experienced at that location. D1 was added to show that even a suboptimal design, which may be required due to space limitations, still provides positive incomes and is a sensible alternative. The only case returning a negative net profit is Chicago for macro BSs, where an additional year (11 years) would be required to amortize the CAPEX.

As one may expect, the actual sizing for the solar add-on depends on the energy selling price as well as on the location. Nevertheless, the rather good results we have shown here are encouraging. These are due to the modest cost of PV technology, which has been plummeting over the last decade (10-fold reduction). In addition, we observe that while commercial panels at the time of writing have maximum efficiencies of about 21 percent, new developments with efficiencies as high as 44 percent are on the way [3]. The battery cost is still rather high, but trends are encouraging for it as well. As an example, since 2008, the cost reduction has been about one third for lithium ion cells, which is the technology of choice at the time of writing. These facts can be found in numerous reports (e.g., [15]), and allow us to assert that the scenarios envisioned here are already feasible and expected to become even more appealing in the near future, as the harvesting CAPEX drops further and PV efficiencies improve.

The following observations are in order. First, we note that further optimizations in the BS energy consumption model, although not considered here, are possible. For example, the design of energy-efficient sleeping modes is expected to be a very effective means to further reduce the energy consumption figure. Their exploitation could lead to further savings by selectively switching off some of the BSs (note that sleeping modes should be designed to allow for quick transitions into the active state as soon as new traffic is detected). Second, we remark that with the advent of 5G mobile systems, the cell demand is expected to further increase with

	Chicago			Los Angeles		
BS type	D1 (net income)	D2 (net income)	D2 (annual revenue)	D1 (net income)	D2 (net income)	D2 (annual revenue)
Pico	$19 (1, 20)	$58 (2, 20)	$71	$51 (1, 20)	$117 (2, 20)	$130
Micro	$232 (10, 80)	$607 (20, 80)	$709	$544 (10, 80)	$1193 (20, 80)	$1295
Macro	–$1566 (60, 500)	–$695 (80, 500)	$1395	$446 (60, 500)	$1813 (80, 500)	$2568

Table 1. Net income and annual revenue for different configurations. for the net income the notation is "$\$X$ (Y, Z)," where X is the net income in U.S. dollars, Y is the solar panel size (square meters), and Z is the battery size (Ah). 12, 24, and 48 V batteries are implied for pico, micro, and macro BSs, respectively.

respect to the traffic volume we consider here. This will not affect the energy consumption of small cells, as their energy expenditure only marginally depends on the load ρ, and will have a minor impact on microcells (Eq. 1). The results for macro deployment may, however, be affected by a higher load due to their larger α parameter (Eq. 1). However, this may be partially mitigated by the massive deployment of small cells, which are expected to offload macro BSs and by further optimizations that are among the objectives of fifth generation (5G) systems.

Research Challenges

Building on the above discussion, we believe that nearly zero energy cellular networks are fertile ground for research, both theoretical and applied.

Research challenges go from the optimal energy management of federations of small cells to their active involvement as actors of future energy grids and new energy trading models. Moreover, while some technical work has been performed on energy management (in terms of, e.g., load balancing and BS ON/OFF switching) for federations of small cells, much still has to be done for the scenario of this article, where cells harvest ambient energy and sell or transfer part of it to other network elements. To the best of our knowledge, these aspects are currently unexplored. Next, a few fundamental problems are identified and discussed.

Optimal Energy Management of Federations of Small Cells

Here, we envision a network setup where hierarchical cell structures are deployed within the same geographical area (e.g., a district) and possibly operated off-grid. Modulating the user profile has a modest impact for picocells (variations in the load ρ only marginally affect their energy consumption). Thus, it may make sense to think of installations where some BSs are temporarily put into some power saving state. The cells that are still active will provide the needed coverage, whereas the remaining ones will recharge their batteries. This dynamic management is required as the renewable energy inflow may differ from cell to cell depending on obstructions from objects, which will cause partial shading of the irradiated solar power.

This calls for smart algorithms that take into consideration aspects such as load balancing among the cells and the need for offloading some of the traffic from the macrocells in the area. Note also that, although no energy can be traded by the cells when these are operated off-grid, their presence is expected to relieve some demand from the macrocells in the area by substantially reducing the energy that the latter purchase from the power grid. Hence, although no direct energy transfer occurs, the discussed offloading mechanism is expected to provide economical savings for macro-BS operators.

QoE and Energy Harvesting Aware Streaming

There is a large body of work concerning the optimization of routing for multimedia traffic for increased quality of experience (QoE). This has to do with the selection of the access point, but also with routing strategies within the fixed portion of the network. These algorithms are usually based on measures of congestion and the on-the-fly estimation of the quality perceived by end users. Current adaptation strategies act on the degree of compression of audio/video streams (video coding rate), their transmission rate, power, and possibly their routing path.

Here, we support the use of additional information to model the residual energy at the mobile user and the base station(s), and the status of the corresponding energy harvesting process(es). Future adaptation policies could then take these further aspects into account to deliver the expected target QoE at the end user side under the energy sustainability constraints of the system.

Blending Mobile Networks with Power Grids

With the possibility of trading energy with the grid, future communication networks will become active actors in future smart energy grids. The energy injected by BS deployments could be used to provide ancillary services for power grids such as load balancing and compensation, peak shaving, or, for example, the reduction of power distribution losses through the injection of controlled amounts of power [16]. This especially holds for micro and macrocells, the generated energy of which will easily surpass that of residential users. For pico deployments, we may imagine the support of connected loads, such as weather stations, street lightning, or IoT networks for traffic con-

trol in smart cities. In addition, we may think of massive installations where energy aggregators will be in charge of trading the energy generated by a myriad of distributed small cells.

According to this vision, telecommunication infrastructures will additionally manage their internal resources in terms of trading with the electricity grid, possibly benefiting its efficiency. This may benefit from the adoption of demand/response strategies managed by electric utilities [11]. Moreover, this will only work if proper pricing schemes will are in place, which should incentivize BSs to sell their excess energy while also making these transactions convenient for the electricity grid.

These aspects are still unexplored, and their study requires optimization frameworks that jointly account for telecommunication and grid aspects under radically new market models. In addition, co-simulation tools are essential to assess the performance of such a complex system.

Concluding Remarks

In this article we have made the case for nearly zero energy cellular networks, where excess energy can be traded with the electricity grid to make a profit and provide ancillary services. To support this vision, we have provided quantitative performance examples, using real data traces, and have elaborated on challenging and new research issues. The foreseen technology positions itself at the intersection between the telecommunications and power grid fields, and its success requires cross-disciplinary research involving tools from telecommunications, operations research, economics, and smart grids. The authors believe that this technology holds huge potential and will lead to an exciting new field.

References

[1] Cisco Systems Inc., "Cisco Visual Networking Index Global Mobile Data Traffic Forecast Update 2012–2017," white paper, http://www.cisco.com/, Feb. 2013.

[2] G. Auer *et al.*, "How Much Energy Is Needed to Run a Wireless Network?" *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[3] NREL, National Renewable Energy Laboratory, "Best Research-Cell Efficiencies," http://www.nrel.gov/ncpv/images/efficiency_chart.jpg.

[4] B. Lindemark and G. Oberg, "Solar Power For Radio Base Station (RBS) Sites Applications Including System Dimensioning, Cell Planning and Operation," *Int'l. Telecommun. Energy Conference*, Edinburgh, U.K., Oct. 2001.

[5] H. Al Haj Hassan, L. Nuaymi, and A. Pelov, "Renewable Energy in Cellular Networks: A Survey," *2013 IEEE Online Conf. Green Commun.*, Oct 2013, pp. 1–7.

[6] G. Piro *et al.*, "HetNets Powered by Renewable Energy Sources: Sustainable Next-Generation Cellular Networks," *IEEE Internet Computing* , vol. 17, no. 1, Jan. 2013, pp. 32–39.

[7] H. S. Dhillon *et al.*, "Fundamentals of Heterogeneous Cellular Networks with Energy Harvesting," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, May 2014, pp. 2782–97.

[8] T. Han and N. Ansari, "On Optimizing Green Energy Utilization for Cellular Networks with Hybrid Energy Supplies," *IEEE Trans. Wireless*, vol. 12, no. 8, Aug. 2013, pp. 3872–82.

[9] T. Pamuklu and C. Ersoy, "Optimization of Renewable Green Base Station Deployment," *IEEE Int'l. Conf. Green Computing and Commun.*, Beijing, China, Aug. 2013.

[10] M. S. Zefreh, T. D. Todd, and G. Karakostas, "Energy Provisioning and Operating Costs in Hybrid Solar-Powered Infrastructure," *IEEE Trans. Sustainable Energy*, vol. 5, no. 3, July 2014, pp. 986–94.

[11] F. Rahimi and A. Ipakchi, "Demand Response as a Market Resource under the Smart Grid Paradigm," *IEEE Trans. Smart Grids*, vol. 1, no. 1, June 2010, pp. 82–88.

[12] NREL, "Renewable Resource Data Center," http://www.nrel.gov/rredc/.

[13] M. Miozzo *et al.*, "SolarStat: Modeling Photovoltaic Sources through Stochastic Markov Processes," *IEEE Energy Conf.*, Dubrovnik, Croatia, May 2014.

[14] EARTH, "D2.3: Energy Efficiency Analysis of the Reference Systems, Areas of Improvements and Target Breakdown," Project Deliv. D2.3, http://www.ict-earth.eu, 2010.

[15] Bloomerg New Energy Finance, "World Energy Perspective — The Cost of Energy Technologies," World Energy Council's white paper, http://about.bnef.com, Oct. 2013.

[16] P. Tenti *et al.*, "Distribution Loss Minimization by Token Ring Control of Power Electronic Interfaces in Residential Microgrids," *IEEE Trans. Industrial Electronics*, vol. 59, no. 10, Oct. 2012, pp. 167–78.

Biographies

Davide Zordan (zordanda@dei.unipd.it) received his M.Sc. in telecommunications engineering and Ph.D. from the University of Padova, Italy, in 2010 and 2014, respectively. He is currently a postodoctoral researcher at the Department of Information Engineering, University of Padova. His research interests include stochastic modeling and optimization, protocol design, and performance evaluation for wireless networks, and in-network processing techniques (including compressive sensing), energy-efficient protocols, and energy harvesting techniques for WSNs and wearable IoT devices.

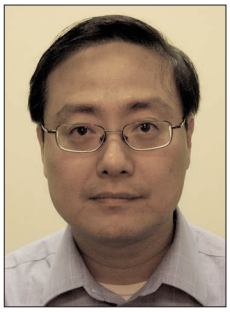
Marco Miozzo (mmiozzo@cttc.es) is a researcher at the Centre Tecnológic de Telecomunicacions de Catalunya (CTTC)and a Ph.D. student at the Technical University of Catalonia, Barcelona, Spain. His main research interests are in mobile networks, green communications, and mobile networks with energy harvesting capabilities. He holds a Master's in telecommunications engineering from the University of Ferrara, Italy.

Paolo Dini (pdini@cttc.es) has been with CTTC from 2006. His research interests encompass wireless networks modeling and optimization, with particular emphasis on energy saving and energy efficiency, integration of renewable energy sources in mobile network equipment, energy sustainable network architectures and protocols, self-organizing networks, cognitive networking, radio resource and mobility management, and quality of service/experience provision. He is author of around 40 papers in scientific journals and international conferences.

Michele Rossi [SM] (rossi@dei.unipd.it) is an assistant professor with the Department of Information Engineering, University of Padova. His current research interests are centered around wireless sensor networks (WSNs), IoT, telecommunication and optimization aspects of smart power grids, and energy harvesting mobile networks. He recently initiated a research line on signal processing (feature extraction and compression) for wearable IoT devices. In the last few years, he has been actively involved in EU projects on IoT technology and has collaborated with Worldsensing (Barcelona, Spain) on the design of optimized WSN solutions for smart cities. He is the author of more than 100 scientific papers published in international conferences, book chapters, and journals, and has been the recipient of four best paper awards from the IEEE. He currently serves on the Editorial Board of *IEEE Transactions on Wireless Communications*.

The foreseen technology positions itself at the intersection between the telecommunications and power grid fields, and its success requires cross-disciplinary research involving tools from telecommunications, operations research, economics, and smart grids.

AUTOMOTIVE NETWORKING AND APPLICATIONS


Wai Chen


Luca Delgrossi


Timo Kosch


Tadao Saito

In this 15th issue of the Automotive Networking and Applications Series, we are pleased to present two papers that address security vulnerabilities of connected vehicle streams in cooperative driving and web access for passengers of public transit systems in the presence of intermittently connected networks.

Autonomous vehicles capable of navigating the roadways with little feedback from humans are maturing rapidly. With the increased reliance on onboard sensors, computer software, and wireless communications to make cooperative driving decisions for passengers, ensuring the security of autonomous driving systems (e.g., autonomous vehicle streams) will become increasingly important. The first article, "Security Vulnerabilities of Connected Vehicles Streams and Their Impact on Cooperative Driving" by M. Amoozadeh *et al.*, presents a first look at the effects of security attacks on the communication channel as well as sensor tampering of a connected vehicle stream equipped to achieve cooperative adaptive cruise control (CACC). The authors first give a comprehensive overview of the security attacks on the CACC vehicle stream, with particular focus on attacks in the application and network layers. The article then describes a simulation implementation of CACC based on one-vehicle look-ahead communication using IEEE 802.11p, and presents the impacts on the application and network layers under various what-if scenarios of malicious insider attacks (e.g., message falsification and radio jamming). The authors then discuss potential countermeasures for detecting malicious behavior to secure the CACC vehicle streams to mitigate the impacts of attacks (e.g., rear-end collision or instability of the vehicle stream).

Providing web access to passengers of public transit systems has received much attention from the research community with the objective of making public transit a more attractive commuting alternative compared to individual driving. The second article, "ICOW: Internet Access in Public Transit Systems" by S. G. Hong *et al.*, proposes a cache-based system that can be deployed in public transit vehicles to enhance quality of experience for passengers. In comparison to existing caching systems, the authors propose new and practical approaches to improve bandwidth availability, system efficiency, and content delivery performance. The article provides a comprehensive overview of their core system features such as cache-based central control and aggregation of user requests, queuing of user requests (during network disconnections), asynchronous notification of content availability, popularity-based scheduling of retrieval for requested content, and passenger privacy protection, among others. The authors then describe the system architecture and prototype implementation, and discuss the performance of their proposed system. Finally, the article outlines some open issues of deployment and further research.

We thank all contributors who submitted manuscripts for this series, as well as all the reviewers who helped with thoughtful and timely reviews. We thank Dr. Osman Gebizlioglu, Editor-in-Chief, for his support, guidance, and suggestions throughout the process of putting together this issue. We also thank the IEEE publication staff, particularly Ms. Charis Scoggins and Ms. Jennifer Porcello, for their assistance and diligence in preparing the issue for publication.

BIOGRAPHIES

WAI CHEN (waichen@ieee.org) received his B.S. degree from Zhejiang University, and M.S., M.Phil., and Ph.D. degrees from Columbia University, New York. He is Chief Scientist of China Mobile Research and General Manager of China Mobile Internet-of-Things Research Institute. Previously he was Vice President and Group Director of ASTRI, Hong Kong; and Chief Scientist and Director at Telcordia (formerly known as Bellcore), New Jersey, USA. While at Telcordia, he led a vehicular communications research program over 10 years in collaboration with a major automaker on automotive networking technologies for vehicle safety and information applications. He was Principal Investigator of several government funded projects on advanced networking technologies research. He was the General Co-Chair for the IEEE Vehicular Networking Conference (IEEE VNC 2009 – 2013) and a Guest Editor for Special Issue on Vehicular Communications and Networks for the IEEE Journal on Selected Areas in Communications (2011). He has also served as a Guest Editor for Special Issue on Inter Vehicular Communication (IVC) for the IEEE Wireless Communications Magazine (2006), the IEEE Distinguished Lecturer (2004-2006), the Co-Chair for Vehicle-to-Vehicle Communications Workshop (IEEE V2VCOM 2005–2008) co-located at IEEE Intelligent Vehicles Symposium, the Co-Chair for the IEEE Workshop on Automotive Networking and Applications (IEEE AutoNet 2006–2008) co-located with IEEE Globecom, and the Vice Chair of Technical Program Committee for Vehicular Communications of the IEEE Vehicular Technology Conference (IEEE VTC Spring 2009).

LUCA DELGROSSI is manager of the Vehicle-Centric Communications Group at Mercedes-Benz Research & Development North America Inc., Palo Alto, California. He started as a researcher at the International Computer Science Institute (ICSI) of the University of California at Berkeley and received his Ph.D. in computer science from the Technical University of Berlin, Germany. He served for many years as professor and associate director of the Centre for Research on the Applications of Telematics to Organizations and Society (CRATOS) of the Catholic University at Milan, Italy, where he helped create and manage the Master's in Network Economy (MiNE) program. In the area of vehicle safety communications, he coordinated the Dedicated Short Range Communications (DSRC) Radio and On-Board Equipment work orders to produce the DSRC specifications and build the first prototype DSRC equipment as part of the Vehicle Infrastructure Integration (VII) initiative of the U.S. Department of Transportation. The Mercedes-Benz team in Palo Alto is a recognized leader in the R&D of vehicle-to-infrastructure as well as vehicle-to-vehicle communications safety systems.

TIMO KOSCH works as a team manager for BMW Group Research and Technology, where he is responsible for projects on distributed information systems, including such topics as cooperative systems for active safety and automotive IT security. He has been active in a number of national and international research programs, and serves as coordinator for the European project COMeSafety, co-financed by the European Commission. He Is also currently heading the system development for a large German Car2X field test. For more than three years, until recently, he chaired the Architecture working group and was a member of the Technical Committee of the Car-to-Car Communication Consortium. He studied computer science and economics at Darmstadt University of Technology and the University of British Columbia in Vancouver with scholarships from the German National Merit Foundation and the German Academic Exchange Service. He received his Ph.D. from the Computer Science Faculty of the Munich University of Technology.

TADAO SAITO [LF] received a Ph. D degree in electronics from the University of Tokyo in 1968. Since then he has been a lecturer, an associate professor, and a professor at the University of Tokyo, where he is now a Professor Emeritus. From April 2001 to 2013, he was chief scientist and CTO of Toyota InfoTechnology Center, where he studied future ubiquitous information services around automobiles. His research includes a variety of communication networks and their social applications such as ITS. Included in his past study, in the 1970s he was a member of the design group of the Tokyo Metropolitan Area Traffic Signal Control System, designed to control 7000 intersections under Tokyo Police Authority. Since 1990 he is the Chairman of the Multi Media Promotion Forum, organizing 30 industry-related meetings every year. Now he is the Chairman of the New Generation IP Network Promotion Forum of Japan. He has written two books on electronic circuitry, four books on computers, and two books on digital communication and multimedia. From 1998 to 2002 he was the Chairman of the Telecommunication Business Committee of the Telecommunication Council of the Japanese government and contributed to regulatory policy of telecommunication business for broadband network deployment in Japan. He is an honorary member and Fellow of IEICEJ.

Security Vulnerabilities of Connected Vehicle Streams and Their Impact on Cooperative Driving

Mani Amoozadeh, Arun Raghuramu, Chen-Nee Chuah, Dipak Ghosal, H. Michael Zhang, Jeff Rowe, and Karl Levitt

ABSTRACT

Autonomous vehicles capable of navigating unpredictable real-world environments with little human feedback are a reality today. Such systems rely heavily on onboard sensors such as cameras, radar/LIDAR, and GPS as well as capabilities such as 3G/4G connectivity and V2V/V2I communication to make real-time maneuvering decisions. Autonomous vehicle control imposes very strict requirements on the security of the communication channels used by the vehicle to exchange information as well as the control logic that performs complex driving tasks such as adapting vehicle velocity or changing lanes. This study presents a first look at the effects of security attacks on the communication channel as well as sensor tampering of a connected vehicle stream equipped to achieve CACC. Our simulation results show that an insider attack can cause significant instability in the CACC vehicle stream. We also illustrate how different countermeasures, such as downgrading to ACC mode, could potentially be used to improve the security and safety of the connected vehicle streams.

INTRODUCTION

Autonomous driverless cars have recently received much publicity with successful demonstrations by Google, whose self-driving car has completed over 700,000 autonomous-driving miles across cities in the United States.[1] Levinson *et al*., [1] present algorithms used in "Junior," Stanford's autonomous vehicle that is capable of performing complex driving tasks in real time in unpredictable traffic conditions. While these ambitious forward-looking projects are currently in the proof-of-concept stage, lower levels of *function-specific automation* like cruise control (CC), and *combined function automation*[2] such as adaptive CC (ACC) are already finding their way into deployment in certain high-end automobiles. Cooperative ACC (CACC) is an extension to ACC that leverages inter-vehicle communications to create tightly coupled vehicle stream [2].

To achieve automated cooperative driving, vehicles need to have access to each other's information. Such information enhances the ability of the autonomous vehicle to plan ahead and make better decisions to improve the overall safety and performance of the vehicle. One possible way of achieving inter-vehicle communication is through vehicular ad hoc networks (VANETs). We refer to vehicles that participate in information sharing and cooperative driving through VANET communications as a *connected vehicle stream*. Although VANET technologies have matured over time, they are still riddled with security issues. Engoulou *et al.* [3] present the most recent survey of various security issues in VANETs including security requirements, attacks, and privacy protection. Faults in software components can potentially lead to devastating effects for autonomous vehicles and other vehicles sharing the roadway. Thus, it is important to design the system to be robust, secure against malicious attacks, privacy preserving, and fault tolerant.

The *Security Credential Management System* (SCMS), developed under a cooperative agreement with the United States Department of Transportation (USDOT), is currently the leading candidate for vehicle-to-vehicle (V2V) security backend design in the United States [4]. These efforts focus on securing the V2X communication channels over the *Wireless Access in Vehicular Environment* (WAVE) standard. Relatively little attention has been given to the safety and security of VANET control protocols in autonomous vehicles, and the impact of different security attacks in the presence of an adversary and bad-faith participants. Compromising sensor input or control protocols can lead to incorrect decisions being made by autonomous vehicles, leading to dynamic interactions between vehicles that threaten the stability and safety of autonomous vehicle streams.

Blum *et al*. [5] argue that jamming attacks can cause collisions in a vehicle platoon, leading

The authors are with the University of California, Davis.

This work is partly supported by NSF CMMI-1301496 grant and Intel Science and Technology Center for Secure Computing.

[1] Google Official Blog, The latest chapter for the self-driving car: mastering city street driving.

[2] U.S. Department of Transportation, Policy on Automated Vehicle Development.

0163-6804/15/$25.00 © 2015 IEEE

to serious multicar pile-ups. Guette *et al.* [6] present a security model based on the trusted platform module (TPM), and describe an application of cooperative driving and its associated threat model. These prior works only discuss a limited number of attacks, and none of them consider actual vehicle longitudinal control in CACC. As a result, these studies ignore other impacts that a security attack might have on a platoon such as string instability. Moreover, none of these studies have carried out any quantitative analysis of the impact of the attacks. Our main contributions in this work are the two areas below.

Analysis of Security Risks in an Autonomous Vehicle Stream: We perform a study of the security vulnerabilities and risks associated with deploying VANET communication in connected vehicle streams to achieve automated sensing and control such as CACC. We consider various types of what-if scenarios when communications between autonomous vehicles participating in CACC are compromised. In addition, we examine existing countermeasures, and explore the limitations of these methods and possible ways to alleviate negative effects.

Quantitative Analysis of Security Attacks: Vehicular Network Open Simulator (VENTOS) is an integrated open source simulator we have developed to study VANET-based traffic applications. A CACC car-following model based on one-vehicle look-ahead communication utilizing IEEE 802.11p was implemented in VENTOS in order to study various what-if scenarios involving security attacks on a CACC vehicle stream. In particular, we present simulation results on the impact of application and network layer attacks by a malicious insider on the performance of a CACC vehicle stream.

Cooperative Adaptive Cruise Control

CACC is an enhancement of ACC, and we consider it as a case study of cooperative driving of autonomous vehicles. CACC incorporates wireless V2V communication to access rich preview information about surrounding vehicles. This leads to tighter following gaps and faster response to changes than ACC, and makes collaborative driving such as platooning feasible. In this study, parameter exchange in a CACC vehicle stream is done through beacons, which are periodic single-hop messages broadcast by each vehicle.

We consider a CACC vehicle stream that is moving on a straight single-lane highway. The vehicles utilize a simple one-vehicle look-ahead communication scheme shown in Fig. 1a. Each CACC vehicle listens to beacon messages sent wirelessly using IEEE 802.11p from its immediately preceding vehicle. The vehicles then utilize the speed, position, acceleration, and other information embedded in these beacon messages to achieve distributed longitudinal control. Details on the CACC control design are discussed in [2].

We assume that the platoon of vehicles is already formed and is traveling on a straight single-lane highway with no need to change the platoon size or perform maneuvers (split, merge, leave, etc). There will be no need for a *platoon management protocol*, and the platoon leader acts only as the first vehicle in the platoon and creates disturbance by slowing down or speeding up. Thus, the only active communication between CACC vehicles is beaconing used to exchange necessary parameters for a longitudinal controller.

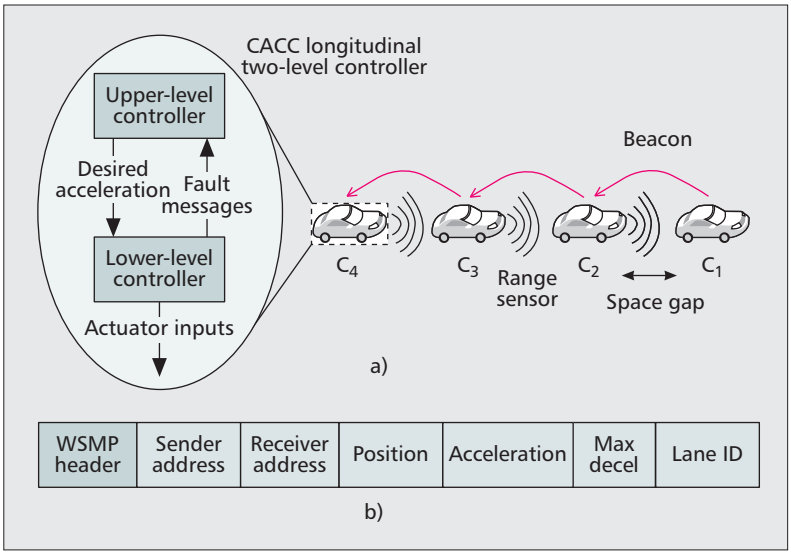


Figure 1. One-vehicle look-ahead communication in a CACC vehicle stream. The ith vehicle receives information embedded in beacon messages from the $(i - 1)$th vehicle using V2V wireless communication, and feeds it into the longitudinal control system to maintain a safe gap from the preceding vehicle. The longitudinal control system in each vehicle is typically designed as a hierarchical two-level controller [2]: a) V2V communications; b) beacon format.

Security Attacks on a CACC Vehicle Stream

We group the security attacks on a CACC vehicle stream as application layer, network layer, system level, and privacy leakage attacks. All these attacks can potentially impact the string stability of the system, and compromise the safety and privacy of the passengers of the CACC vehicle stream. Such attacks can be launched by either an outsider or insider adversary. While leveraging state-of-the-art security architectures can potentially limit the capabilities of outsider attacks, there can still be disruptive insider attacks. In the following subsections, we discuss the various categories of security attacks in more depth.

Application Layer Attacks

Application layer attacks affect the functionality of a particular application such as CACC beaconing, or message exchange in the platoon management protocol. The adversary can use message falsification (modification), spoofing (masquerading), or replay attacks to maliciously affect the vehicle stream. The impact of such attacks can be temporary instability in the vehi-

Unlike application layer attacks, network layer attacks have the potential to affect the functioning of multiple user applications. For instance, the adversary can attempt a DoS attack or DDoS attack to overwhelm the communication capability of a vehicle or a group of vehicles, and make them unable to participate properly in a CACC vehicle stream.

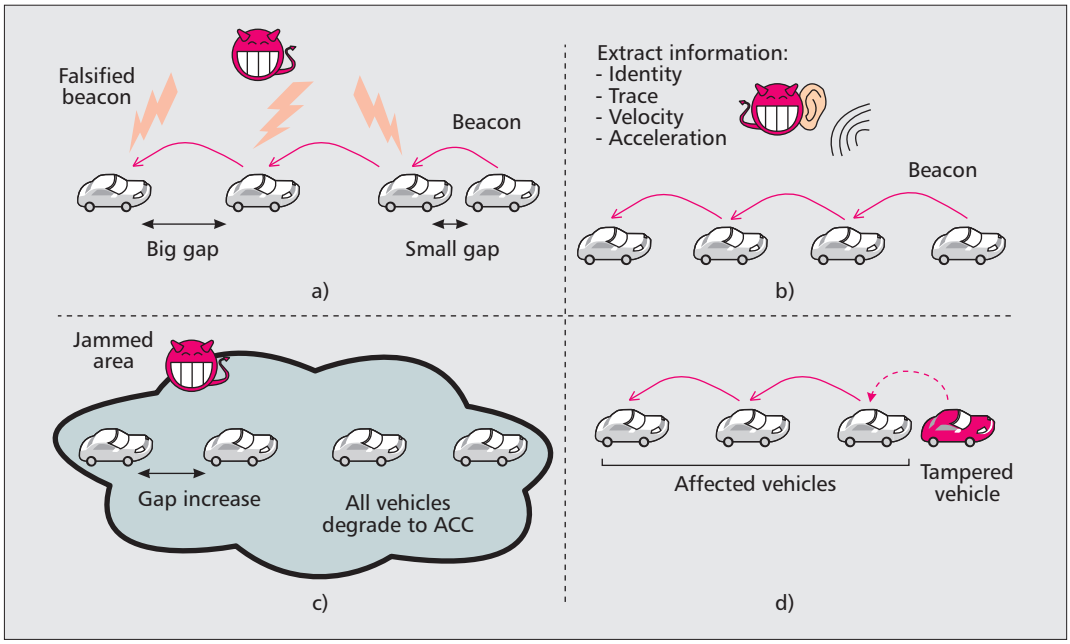


Figure 2. Security attacks on a CACC vehicle stream: a) falsification attack; b) eavesdropping attack; c) radio jamming attack; d) tampering attack.

cle stream, and in severe cases can lead to rear-end collision.

In a *message falsification attack*, the adversary starts listening to the wireless medium and, upon receiving each beacon, manipulates the content meaningfully and rebroadcasts it as depicted in Fig. 2a. Changing the value of different fields in a beacon might have different effects on the system depending on the implementation of the vehicle's longitudinal control system. For instance, changing the acceleration field might have a more significant effect than changing the velocity.

In a *spoofing attack*, the adversary impersonates another vehicle in the stream in order to inject fraudulent information into a specific vehicle. In one-vehicle look-ahead communication, the adversary can impersonate the vehicle preceding the target vehicle even when the attacker is physically distant from the target vehicle. Note that an *in-transit traffic tampering attack* in order to modify, delay, or drop messages is not applicable in a CACC vehicle stream. This is due to the fact that all communications are single-hop, and no vehicle acts as a relay node in the system.

In a *replay attack*, the adversary receives and stores a beacon sent by a member of the stream and tries to replay it at a later time with malicious intent. The replayed beacon contains old information, which can lead to hazardous effects. For instance, consider a scenario where a CACC vehicle stream is moving forward with speed of 30 m/s (108 km/h). The adversary captures a beacon, and stores it for later use. When the leading vehicle slows down, the adversary injects the old beacon into the system periodically. Following vehicles still think that the lead vehicle is driving at 30 m/s and do not slow down, potentially leading to a collision.

State-of-the-art security architectures employing a strong cryptographic system have the potential to effectively thwart application layer attacks in the case where the adversary is an untrusted outsider. Digital signatures provide data integrity for beacon messages and protect them from unauthorized change. In addition to *data integrity*, digital signatures also provide *authentication* (both peer entity and data-origin authentication), as well as *non-repudiation* (with the help of a trusted third party) services. Moreover, using *nonce* in the messages, which is an arbitrary number (chosen in a pseudo-random process) used only once in communication, is a technique to prevent replay attacks.

Although the above methods are well known, there are practical challenges involved in deployment, implementation, and standardization of such security architectures in VANETs. This is due to the scale of the proposed VANET and varying interpretations of what constitutes "security and privacy" in various areas of the world [3]. Furthermore, in the case where the adversary is a trusted insider such as a compromised vehicle with a valid certificate, the problem is much harder to solve. Typical approaches to handling this issue are via misbehavior/anomaly detection techniques such as [7], which require multiple sources of data that may not always be available. In addition, such techniques do not guarantee perfect detection in all circumstances, and are usually associated with finite false negative and false positive rates. There is much left to be done in the anomaly detection field to ensure acceptable performance of these algorithms to ensure the safety of passengers in the fully autonomous driving scenario.

Network Layer Attacks

Unlike application layer attacks, network layer attacks have the potential to affect the functioning of multiple user applications. For instance, the adversary can attempt a denial-of-service (DoS) or distributed DoS (DDoS) attack to overwhelm the communication capability of a

vehicle or a group of vehicles, and make them unable to participate properly in a CACC vehicle stream.

A known method to realize DoS in the VANET scenario is by using a vehicular botnet. Mevlut *et al*. [8] demonstrate the problems vehicular botnets introduce to the autonomous car setting via a simulation study. In their work, they focus on producing physical congestion in a given road segment and do not discuss the effects of network level congestion through botnets of compromised cars. It is envisioned that autonomous cars are equipped with a tamper-proof *hardware security module* (HSM), which is responsible for storing digital keys as well as performing all cryptographic operations, such as message signing/verification, encryption, and hashing [9]. These cryptographic operations are complex and CPU-intensive; hence, there is an upper bound on the number of cryptographic operations a HSM can perform at a time. A DoS/DDoS attack can target this limitation to overwhelm an autonomous vehicle and make its HSM unavailable.

Radio jamming to deliberately disrupt communications over small or wide geographic areas, as depicted in Fig. 2c, is another possible network layer DoS attack. IEEE 802.11p standard uses one *control channel* (CCH) with multiple *service channels* (SCHs). The adversary can use different jamming techniques like one-channel jamming or swiping between all channels and trying to jam them all. We would need a system such as [10] to help detect and mitigate such attacks. Other countermeasures for DoS attacks in CACC vehicle streams involve traditional solutions such as channel switching, technology switching, frequency hopping, and utilizing multiple radio transceivers. If none of these techniques are feasible, a CACC vehicle may need to downgrade to the ACC system to help avoid a rear-end collision.

System-Level Attacks

All presented attacks so far have been centered around exploiting V2V communication in a CACC vehicle stream. Another type of attack is tampering with vehicle hardware or software, which can be done by a malicious insider at the manufacturing level or by an outsider in an unattended vehicle (e.g., by replacing or altering certain vehicle sensors). Even if the communication channel is secure, and a state-of-the-art security architecture is deployed in the VANET if the onboard hardware/software is tampered with or faulty, the input information to the system will not be accurate. This affects the operation of the high-level protocols as illustrated in Fig. 2d. Hence, the risk of tampering should not be neglected.

One possible solution is to use tamper-proof sensors. If a tamper-proof version is not available or too expensive to deploy on a large scale, the misbehavior detection techniques discussed below can be useful. It is worth mentioning that in general, the term *tamper-proof hardware* means 100 percent secure against tampering, which is mostly used by marketing specialists and does not exist in practice. There are different technologies to make a device less vulnerable against tampering, and *tamper resistance*, *tamper evidence*, *tamper detection*, and *tamper response* are more accurate terms used to characterize such technologies [11], but these are out of the scope of this article.

Privacy Leakage Attacks

CACC vehicles periodically broadcast beacons that contain various types of information such as vehicle identity, current vehicle position, speed, and acceleration. The availability of this information can comprise the privacy. The adversary can carry out an *eavesdropping* attack, as shown in Fig. 2b, to extract valuable information about the vehicle stream such as its trace by linking position data, and use it for her own benefit. The presence of signatures in the beacon messages can worsen the situation, and allow the adversary to easily identify the participating vehicles in the CACC stream.

Eavesdropping is a type of passive attack, and hence is difficult to detect, especially in broadcast wireless communication. However, it is possible to prevent the success of eavesdropping by using encryption to achieve data privacy or using anonymity techniques to achieve identity and location privacy. Anonymity is typically implemented using *group signatures* [12] or *short-term certificates* (*pseudonyms*) [13]. Many privacy-preserving security architectures in VANETs use pseudonym-based schemes to keep the information private, but their applicability in a platoon requires more detailed investigation, and is out of the scope of this article.

Simulation Study

Simulation Setting

In order to study some of the discussed attacks on a CACC vehicle system, we utilize the VENTOS platform.[3] VENTOS is an integrated simulator, and is made up of many different modules, including Simulation of Urban Mobility (SUMO) and OMNET++/Veins. SUMO[4] is adopted as our traffic simulator, while OMNET++[5] is used to simulate the wireless communication. Our ACC and CACC car-following models are implemented to replace the default car-following model in SUMO. Moreover, the traffic control interface (TraCI), which is responsible for data/command exchange between SUMO and OMNET++, is extended with a new set of commands to gain necessary control over parameters exchange for ACC/CACC vehicles. Detailed packet-level simulation is performed in OMNET++, and IEEE 802.11p protocol, the standard protocol adopted for V2V communication in Vehicles in Network Simulation (Veins) framework,[6] is used for wireless communication between CACC vehicles.

VENTOS allows us to study situations where driverless vehicles utilize V2V communication to achieve CACC, and analyze the local and string stability of the system under different speed profiles. Stability is a critical requirement for both ACC and CACC control system design, and is achieved by dampening traffic flow disturbances [14]. *Local stability* concerns one vehicle following a preceding vehicle. A system is said to be local stable if the magnitude of disturbance

Other countermeasures for DoS attack in a CACC vehicle stream involve traditional solutions such as channel switching, technology switching, frequency hopping, and utilizing multiple radio transceivers. If none of these techniques are feasible, a CACC vehicle may need to downgrade to the ACC system to help avoid a rear-end collision.

[3] http://rubinet.ece.ucdavis.edu/projects/ventos

[4] http://goo.gl/Al4w07

[5] OMNet++ Network Simulation, http://www.omnetpp.org/

[6] http://veins.car2x.org/

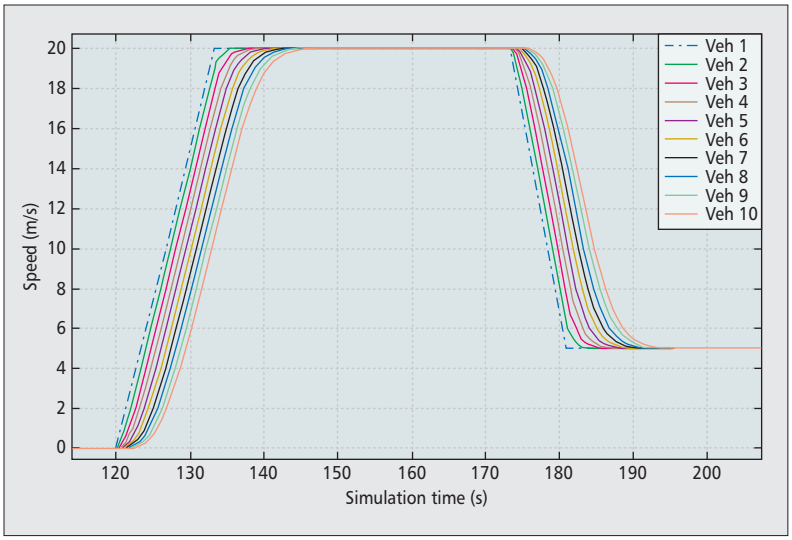


Figure 3. Speed profile of CACC vehicle stream with no adversary. The system is stable, and as "Veh 1" speeds up and slows down, all vehicles follow each other smoothly.

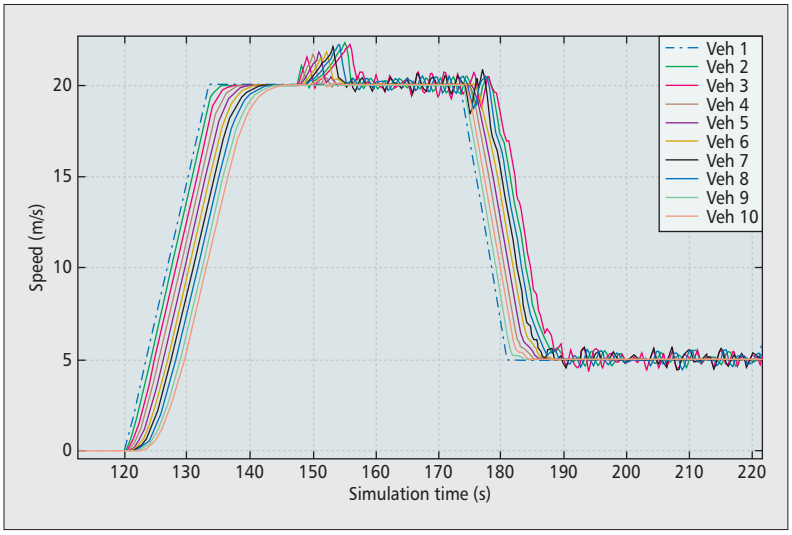


Figure 4. The effect of a message falsification attack on a CACC vehicle stream. String stability is not maintained, and the disturbance magnifies through the stream.

decreases with time. *String stability*, also referred to as *platoon stability* and *asymptotic stability*, concerns the propagation of disturbance in a string of vehicles. String stable means disturbance damps out when propagating to upstream vehicles.

We present simulation results of performing application and network layer security attacks with a malicious insider adversary. The insider adversary is on the side of the road with fixed position and is equipped with a radio to communicate with other vehicles in the network. In the application layer attack, we consider message falsification through which the adversary modifies the CACC beacon messages in order to affect the stream. In the network layer attack, we consider radio jamming through which all wireless communications are disrupted.

Results

Figure 3 shows the speed profile of 10 autonomous/driveless CACC vehicles moving on a straight single-lane highway when no insider adversary is present. The first vehicle (platoon leader) is referred to as Veh 1, and its speed profile is given and shown with a dashed blue line. As Veh 1 accelerates and decelerates, the disturbance propagates to the following vehicles. At t = 120 s, Veh 1 accelerates at 1.5 m/s^2 to a cruise speed of 20 m/s. Then it decelerates to 5 m/s with deceleration of –2 m/s^2. The other nine vehicles follow with uniform time gap setting. As can be seen, the system is stable, and vehicles follow each other smoothly.

Figure 4 presents the speed profile of the same vehicle stream in the presence of a malicious insider vehicle that is using a message falsification attack in order to compromise the system. At t = 147 s, the CACC vehicle stream enters into the radio range of the adversary and is affected by it. The adversary tries to manipulate the acceleration field of beacon messages to a fixed value of 6 m/s^2. Due to the fact that no security features are implemented in vehicles, they accept the falsified beacons and use them for longitudinal control, which leads to string instability in the stream, and this disturbance magnifies through the stream. Here, a falsification attack is most effective when a sudden change in acceleration occurs.

We note that since the attacker in this case is an insider, cryptographic security in terms of digital signatures/certificates by itself will not be able to prevent the attack. We will need more elaborate misbehavior detection measures, as mentioned earlier, to help mitigate this problem. In essence, the speed profile of the vehicle stream will look similar to the one in Fig. 4 even in the presence of security measures such as digital signatures/certificates.

Figure 5 illustrates the space gap between vehicles before and after a radio jamming attack. Before radio jamming, the space gap is 16 m when the CACC vehicle stream is traveling at a speed of 20 m/s (mark 1) and 5.5 m at a speed of 5 m/s (mark 2). The insider adversary disrupts all communications in CCH starting from t = 308 s (mark 3). When failing to receive beacons, CACC vehicles downgrade to ACC mode with larger time gap and delay settings. As a result, the space gap is increased to 26 m (mark 4), and the reaction of followers to speed changes becomes relatively slower (mark 5). Downgrading to ACC is a simple countermeasure that diminishes the impact of radio jamming from a rear-end collision to reduction in CACC performance. We leave designs of more elaborate countermeasures to the insider jamming attack for future work.

Potential Countermeasures for Detecting Malicious Behavior

In this section, we explore some possibilities to help secure CACC vehicle streams by detecting compromised or faulty sensors/vehicles. These could have implications for vehicles with varying

levels of automation, from WAVE-enabled vehicles with ACC/CACC capabilities to fully automated driverless vehicles.

Local Plausibility Check

A simple approach to detecting a faulty sensor is to check whether or not the incoming information is plausible [15]. For instance, if a sensor is not reading within its normal range, the sensor may be faulty or tampered with. The incorrect information can be either discarded or interpolated from the past correct information. Another possibility is deriving the information from other relevant sensors. For instance, if the wheel speed sensor is compromised and faulty, the velocity can be derived from the engine speed sensor.

Wearables and Mobile Devices

Wearable devices such as the Google Glass and mobile devices such as smartphones and tablets carry a wide array of sensors such as cameras, accelerometers, and GPS units along with wireless communication capability. These devices are carried by the driver or passengers of a vehicle. This opens up rich opportunity for developing applications that can potentially improve the security and safety of the system. For instance, the wearable/mobile device of a driver or passenger can act as a verifier for the sensing data generated or received by the vehicle. The wearable device can construct a "belief" from its sensor data about the position of the vehicle, velocity, or acceleration, and cross-check this with the belief computed by the vehicle. If there is a discrepancy in the beliefs as seen from the wearable device and the vehicle, it might be an indicator of a security compromise of the hardware or software in the car, or in the communication channel. If the passenger has multiple devices, it might be possible to fuse the sensor information from these devices to construct a more well formed belief, which can then be checked against the vehicle's belief.

Voting

The two approaches described above are based on local detection of misbehaving sensors/software or compromised communication. If this fails, collaborative decision-making techniques such as voting could be carried out that enable vehicles to collectively shield themselves against a misbehaving vehicle. Voting is most effective in scenarios where there are multiple vehicles in a group that are coordinating with one another. A group of vehicles can be defined as nearby vehicles driving within a geographic region or members of a vehicle platoon. Vehicles in a group keep track of each other's behavior, and check for anomalies in the data received from the members of the group and possibly other vehicles on the road. The vehicles then perform a trust computation and vote for/against keeping the vehicle in the group. This process needs to be done at regular intervals and therefore incurs communication overheads. More detailed study is needed to understand the trade-offs between performance and cost.

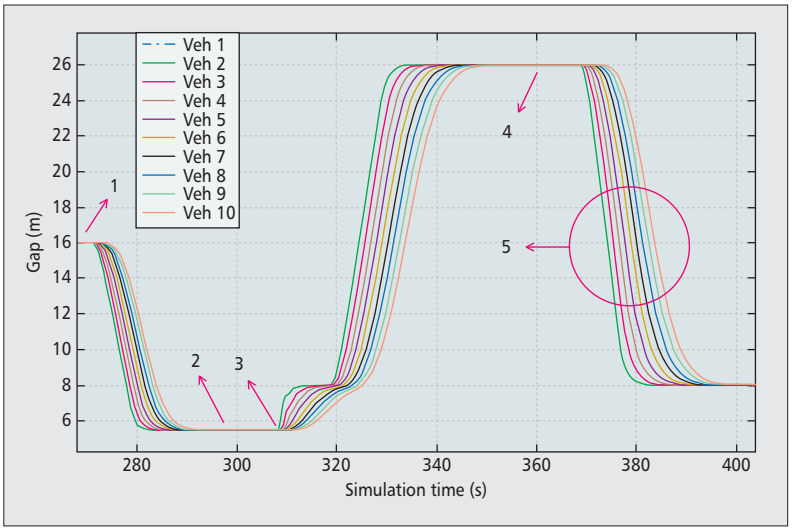


Figure 5. The effect of a radio jamming attack on the CACC vehicle stream. The CACC controller detects communication loss at $t = 308$ s and downgrades to ACC mode.

Conclusion

Mainstream car manufacturers such as Tesla, BMW, and Audi are all working on commercial versions of vehicles with varying levels of automation embedded in them. With the increased reliance on sensors and computer software to make driving decisions for passengers, addressing security issues will become very important with time. This work is an attempt to gain better understanding of the security risks involved in connected vehicle streams, from WAVE-enabled vehicles with ACC/CACC capabilities to fully automated driverless vehicles. We analyze different security attacks on a vehicle stream and discuss the possible security design decisions that will need to be taken to ensure the safety of the system. The impact of a security attack, such as rear-end collision or instability of the stream, is shown with the help of simulation where we implement the ACC/CACC longitudinal control and perform a detailed packet-level simulation.

References

[1] J. Levinson *et al.*, "Towards Fully Autonomous Driving: Systems and Algorithms," *Proc. IEEE Intelligent Vehicles Symp.*, 2011, pp. 163–68.

2] M. Amoozadeh *et al.*, "Platoon Management with Cooperative Adaptive Cruise Control Enabled by VANET," *Vehic. Commun. J.*, 2015.

[3] R. G. Engoulou *et al.*, "Vanet Security Surveys," *Computer Commun.*, vol. 44, 2014, pp. 1–13.

[4] W. Whyte *et al.*, "A Security Credential Management System for V2v Communications," *Proc. IEEE VNC*, 2013, pp. 1–8.

[5] J. Blum and A. Eskandarian, "The Threat of Intelligent Collisions," *IT Professional*, vol. 6, no. 1, 2004, pp. 24–29.

[6] G. Guette and C. Bryce, "Using TPMS to Secure Vehicular Ad-Hoc Networks (Vanets)," *Information Security Theory and Practices: Smart Devices, Convergence and Next Generation Networks*, Springer, 2008, pp. 106–16.

[7] T. H.-J. Kim *et al.*, "Vanet Alert Endorsement Using Multi-Source Filters," *Proc. 7th ACM Int'l. Wksp. Vehicular Internetworking*, ACM, 2010, pp. 51–60.

[8] M. T. Garip *et al.*, "Congestion Attacks to Autonomous Cars Using Vehicular Botnets," 2015.

[9] M. Wolf and T. Gendrullis, "Design, Implementation, and Evaluation of a Vehicular Hardware Security Module," *Proc. Info. Security and Cryptology '11*, pp. 302–18.

[10] A. Hamieh, J. Ben-Othman, and L. Mokdad, "Detection of Radio Interference Attacks in Vanet," *Proc. IEEE GLOBECOM '09*, 2009, pp. 1–5.
[11] B. Rosenberg, *Handbook of Financial Cryptography and Security*, Ch. 10, "Hardware Security Modules," CRC Press, 2010.
[12] X. Lin *et al*., "GSIS: A Secure and Privacy-Preserving Protocol for Vehicular Communications," *IEEE Trans. Vehic. Tech.*, vol. 56, no. 6, 2007, pp. 3442–56.
[13] P. Papadimitratos *et al*., "Architecture for Secure and Private Vehicular Communications," *Proc. IEEE 7th Int'l. Conf. ITS Telecommun.*, 2007, pp. 1–6.
[14] R. Pueboobpaphan and B. van Arem, "Driver and Vehicle Characteristics And Platoon And Traffic Flow Stability," *Journal of the Transportation Research Board*, vol. 2189, no. 1, 2010, pp. 89–97.
[15] P. Golle, D. Greene, and J. Staddon, "Detecting and Correcting Malicious Data In Vanets," *Proc. 1st ACM Int'l. Wksp. Vehicular Ad Hoc Networks*, 2004, pp. 29–37.

Biographies

Mani Amoozadeh is currently a second-year Ph.D. student in the ECE Department at the University of California (UC), Davis. He received his M.S. degree in electrical engineering from KTH, Stockholm, Sweden, and his B.S. degree in computer engineering from Iran University of Science and Technology, Tehran. His research interests include vehicular ad hoc networks, wireless networks, network architecture, and simulation, with emphasis on intelligent transportation systems.

Arun Raghuramu is currently pursuing his Ph.D. in computer science at UC Davis. He received his B.E in electronics and communication engineering from R.V. College of Engineering, Bangalore, India. He was previously involved with research and development teams at McAfee, Guavus, and Sift Security Inc. His research interests include network security, security analytics, and IoT security and privacy issues.

Chen-Nee Chuah [F] is a professor of electrical and computer engineering at UC Davis. She received her Ph.D. in electrical engineering and computer sciences from the University of California, Berkeley. Her research interests include Internet measurements, network management, software defined networks, online social networks, and vehicular ad hoc networks. She is an ACM Distinguished Scientist. She received the NSF CAREER Award in 2003, and the Outstanding Junior Faculty Award from the UC Davis College of Engineering in 2004. In 2008, she was named a Chancellor's Fellow of UC Davis. She has served on the executive/technical program committee of several ACM and IEEE conferences. She was an Associate Editor for *IEEE/ACM Transactions on Networking* from 2008 to 2013.

Dipak Ghosal received his B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 1983, his M.S. degree in computer science and automation from the Indian Institute of Science, Bangalore, in 1985, and his Ph.D. degree in computer science from the University of Louisiana, Lafayette, in 1988. He is currently a professor with the Department of Computer Science, UC Davis. His main research interests include high-speed networks, wireless networks, vehicular ad hoc networks, next generation transport protocols, and parallel and distributed computing.

H. Michael Zhang received his B.S.C.E. degree from Tongji University, Shanghai, China, and his M.S. and Ph.D. degrees in engineering from the University of California, Irvine. He is currently a professor with the Department of Civil and Environmental Engineering, UC Davis. He is an Area Editor for the *Journal of Networks and Spatial Economics* and Associate Editor for *Transportation Research—Part B: Methodological*. His research interests include transportation systems analysis and operations.

Jeff Rowe obtained his Ph.D. from UC Davis in 1996 in particle physics. Since joining the Computer Security Laboratory, he has developed several algorithms for responding to network attacks. He was leader of the team testing and maintaining the GrIDS system. He also led the team subcontracted to produce the IDIP Discovery Coordinator for the Boeing Automated Response to Intrusions project. His research interests include architectures for very large-scale IDS systems, use of correlation techniques in intrusion detection, and automatic response methodologies. He conducts research with Prof. Karl Levitt in the Computer Security Laboratory.

Karl Levitt received his Ph.D. in electrical engineering from New York University in 1966. He has been a professor of computer science at UC Davis since 1986, and has previously been a program director for the National Science Foundation and director of the Computer Science Laboratory at SRI International. He conducts research in the areas of computer security, automated verification, and software engineering.

CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE

WIRELESS COMMUNICATIONS, NETWORKING, AND POSITIONING WITH UNMANNED AERIAL VEHICLES

BACKGROUND

Enabled by the advances in computing, communication, and sensing as well as the miniaturization of devices, unmanned aerial vehicles (UAVs) such as balloons, quadcopters, and gliders have been receiving significant attention in the research community. Indeed, UAVs have become an integral component in several critical applications such as border surveillance, disaster monitoring, traffic monitoring, remote sensing, and the transportation of goods, medicine, and first aid. More recently, new possibilities for commercial applications and public service for UAVs have begun to emerge, with the potential to dramatically change the way in which we lead our daily lives. For instance, in 2013, Amazon announced a research and development initiative focused on its next-generation Prime Air delivery service. The goal of this service is to deliver packages into customers' hands in 30 minutes or less using small UAVs, each with a payload of several pounds. 2014 was a pivotal year that witnessed an unprecedented proliferation of personal drones, such as the Phantom and Inspire from DJI, AR Drone and Bebop Drone from Parrot, and IRIS Drone from 3D Robotics.

Among the many technical challenges accompanying the aforementioned applications, leveraging the use of UAVs for delivering broadband connectivity plays a central role in next generation communication systems. Facebook and Google announced in 2014 that they will use a network of drones which circle in the stratosphere over specific population centers to deliver broadband connectivity. Such solar-powered drones are capable of flying several years without refueling. UAVs have also been proposed as an effective solution for delivering broadband data rates in emergency situations through low-altitude platforms. For example, the ABSOLUTE, ANCHORS, and AVIGLE projects in Europe have been investigating the use of aerial base stations to establish opportunistic links and ad hoc radio coverage during unexpected and temporary events. They can serve as a temporary, dynamic, and agile infrastructure for enabling broadband communications, and quickly localizing victims in case of disaster scenarios.

This proposed Feature Topic (FT) issue will gather articles from a wide range of perspectives in different industrial and research communities. The primary FT goals are to advance the understanding of the challenges faced in UAV communications, networking, and positioning over the next decade, and provide further awareness in the communications and networking communities of these challenges, thus fostering future research. Original research papers are solicited in topics including, but not limited to, the following themes on communications, networking, and positioning with UAVs.

- Existing and future communication architectures and technologies for small UAVs
- Delay-tolerant networking for cooperative UAV operations
- Design and evaluation of wireless UAV testbeds, prototypes, and platforms
- Multihop and device-to-device communications with UAVs
- Interfaces and cross-platform communication for UAVs
- QoS mechanisms and performance evaluation for UAV networks
- Game-theoretic and control-theoretic mechanisms for UAV communications
- Use of civilian networks for small UAV communications
- Integrating 4G and 5G wireless technologies into UAV communications, such as millimeter-wave communications, beamforming, moving networks, and machine type communications
- Use of UAVs for public safety and emergency communications, networking, and positioning
- Integration of software defined radio and cognitive radio techniques with UAVs
- Channel propagation measurements and modeling for UAV communication channels

SUBMISSIONS

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words (from introduction through conclusions). Figures and tables should be limited to a combined total of six. The number of references is recommended not to exceed 15. In some rare cases, more mathematical equations, figures, and tables may be allowed if well justified. In general, however, mathematics should be avoided; instead, references to papers containing the relevant mathematics should be provided. Complete guidelines for preparation of the manuscripts are posted at http://www.comsoc.org/commag/paper-submission-guidelines. Please send a pdf (preferred) or MSWORD formatted paper via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to Author Center. Follow the instructions there. Select "May 2016 / Wireless Communications, Networking and Positioning with UAVs" as the Feature Topic category for your submission.

SCHEDULE FOR SUBMISSIONS

- Submission Deadline: November 1, 2015
- Notification Due Date: January 15, 2016
- Final Version Due Date: March 1, 2016
- Feature Topic Publication Date: May 2016

GUEST EDITORS

Ismail Guvenc
Florida International Univ., USA
iguvenc@fiu.edu

Walid Saad
Virgina Tech, USA
walids@vt.edu

Mehdi Bennis
Univ. of Oulu, Finland
bennis@ee.oulu.fi

Christian Wietfeld
TU Dortmund Univ., Germany
christian.wietfeld@tu-dortmund.de

Ming Ding
NICTA, Australia
ming.ding@nicta.com.au

Lee Pike
Galois, Inc., USA
leepike@galois.com

ICOW: Internet Access in Public Transit Systems

Se Gi Hong, SungHoon Seo, Henning Schulzrinne, and Prabhakar Chitrapu

ABSTRACT

When public transportation stations have access points to provide Internet access to passengers, public transportation becomes a more attractive travel and commute option. However, the Internet connectivity is intermittent because passengers can access the Internet only when a transit vehicle is within the networking coverage of an access point at a stop. We propose ICOW, a system that provides a low-cost way for public transit operators to enhance quality of experience for passengers who access the Internet. Each public transit vehicle is equipped with a smart cache that serves popular content to passengers. The cache updates its content based on passenger requests when it is within range of Internet access points placed at stops, stations, or depots. We have developed a system architecture and built a prototype of the ICOW system. Our evaluation shows that ICOW is significantly more efficient than having passengers contact Internet access points individually and ensures continuous availability of content throughout the journey.

INTRODUCTION

Providing Internet access to passengers makes public transit more attractive as an alternative to commuting or traveling by car. Thus, many public transit organizations are trying to provide such access. There are several options:

- Cellular networks may become sufficiently ubiquitous that individual passengers can use their own devices, such as smartphones or tablets. However, when mobile devices (especially tablets and laptop computers) have only WLAN network capability and users do not want to use their cellular tethering service (i.e., hotspot service) because of data usage caps, they cannot use the cellular network service.
- Public transit organizations can install wireless routers on transit vehicles that combine a cellular network interface with local WLAN network interface. To make this system work ubiquitously, the cellular network connectivity should exist throughout the route of vehicles. However, major cities in America and Europe, such as New York City, London, and Washington, DC, do not offer cellular connectivity in subway systems. From an economic perspective, providing cellular connectivity throughout subway systems is not an easy project [1].
- Public transit organizations can install WiFi access points at stations, allowing connectivity for passengers. For example, New York City is installing public WiFi hotspots at subway stations to provide free WiFi services [2]. However, since such connectivity only lasts within limited network coverage of an access point, passengers experience intermittently connected network service. Despite this intermittent network service, the free or low-cost nature of the service encourages passengers to use this WiFi network.

We focus on the last option, developing a low-cost way for passengers in public transit systems to access web applications, especially in subway systems. When passengers carry a handheld mobile device that is equipped with a WLAN interface, they can freely use Internet service without activating a cellular data plan. This option is suitable for public transit systems in dense urban areas where the distance between stops is short. This option can, in particular, be used in some public transit systems where there is no cellular network connectivity, such as the New York City subway system.

These intermittently connected networks suffer from resource limitations, especially short network connection periods and bandwidth limitations. When requests for content are not handled during a network connection period because the stopping time is not long enough, the pending requests need to be queued until the next stop. This increases content retrieval delay. In addition, there is a flash crowd problem at the beginning of the network connection period because passengers try to access the network simultaneously when they get Internet connectivity. These simultaneous attempts cause network contention problems that degrade link throughput.

Web service enhancement mechanisms over delay-tolerant networking (DTN), Thedu [3], and HTTP-over-DTN [4] have been proposed to enhance interactive web search and reduce the delay of retrieving web pages given a limited connection time by sending all the objects of a web page in a bundle. However, these mechanisms are based on the Internet-side proxy, so users cannot use the services when there is no

Se Gi Hong is with Hughes Network System LLC.

SungHoon Seo is with Korea Telecom.

Henning Schulzrinne is with Columbia University.

Prabhakar Chitrapu is with AT&T.

This work was performed while Se Gi Hong and SungHoon Seo were at Columbia University and Prabhakar Chitrapu was at InterDigital Communications, LLC.

0163-6804/15/$25.00 © 2015 IEEE

Internet connectivity and cannot resolve the network contention problem. 7DS [5], an earlier project we developed in our lab, aims to provide systems for web access to retrieve information and message delivery in peer-to-peer mobile ad hoc networks. However, it does not address the resource limitation and network degradation problems that occur between a mobile device and an AP.

We propose Internet Cache on Wheels (ICOW) to enhance quality of experience (QoE) for passengers who use the Internet in public transit systems. ICOW is a cache-based system that can be deployed on a transit vehicle and avoids retrieving the same content from the Internet. However, ICOW differs from existing caching systems in the sense that it introduces new and practical approaches to increase bandwidth availability, increase system efficiency, and decrease content delivery delay. The ICOW system provides the following core features:

- *Increase bandwidth availability.* The ICOW system resolves the network contention problem and reduces bandwidth usage as it centrally controls and aggregates requests.
- *Increase system efficiency.* The ICOW system queues requests. This queuing mechanism allows passengers to send requests even in the absence of an Internet connection. ICOW also provides other cached web content to passengers when their requested content has not yet been cached. The asynchronous notification of content availability in the ICOW system makes the system easier to use since passengers do not need to check repeatedly whether the requested content is available.
- *Decrease content delivery delay.* The ICOW system schedules requested content retrieval. When the network connection time is not long enough to handle all requests, some requests might be handled during the next connection period. To decrease this content delivery delay, ICOW applies popularity-based scheduling and also prefetches web content.

We describe our design of the system architecture and implementation of the system. Our implementation shows that the queuing mechanism and automatic notification of availability of requested contents enhances QoE in intermittently connected networks. Our performance evaluation shows that ICOW obtains more content in a given time than a direct access system (where each passenger directly connects to an access point, AP). We analytically show that our caching-based system reduces the bandwidth usage during network connection periods and can reduce more bandwidth usage with a peer-to-peer mechanism over DTN. Our simulation results show that our popularity-based scheduling algorithm reduces average content delivery delay while simplifying implementation. More details about system implementation and performance evaluation can be found in [6].

The remainder of this article is organized as follows. We describe an overview of ICOW. We provide system architecture and implementation details. We evaluate performance. Finally, we discuss open issues.

System Design Overview

We introduce the ICOW system to increase content availability, system usability, and satisfaction of passengers by deploying ICOW nodes on public transit vehicles. This ICOW system is used to access web applications and content.

Centralized System

ICOW nodes actively cache content to reduce bandwidth usage by eliminating retrieval of duplicate web content. In addition, the ICOW node centrally controls traffic. It collects and handles all the requests, so the ICOW system can resolve the throughput degradation problem by reducing the number of competing nodes to one for the APs at stops. ICOW nodes have two network interface cards operating on different channels. One of them is used to communicate with passengers, and the other communicates with an AP. Hence, there is no single channel interference between the two wireless links by using different channels, so the overall throughput of the two links increases.

Queuing Requests

When requested contents are not in a local cache, and there is no Internet connectivity, existing HTTP caching proxies return a failure, indicating that it has a network connection problem. Hence, clients need to retransmit requests themselves when there is Internet connectivity. This paradigm causes inconvenience in the sense that clients have to manually poll for Internet connectivity every moment before sending requests.

To overcome this inefficiency of the manual system, we propose a queuing system that queues clients' requests until the system fetches the requested content from the Internet.

This system allows passengers to send requests even when the ICOW node has no Internet connectivity. The ICOW node queues the requests and retrieves the requested content when it has Internet connectivity so that passengers can access the content from the ICOW node. Thus, this ICOW node works as a proxy server.

The ICOW node creates multiple concurrent, parallel connections to web servers to retrieve contents when the Internet connectivity exists. If the ICOW node receives multiple requests for the same URL during a network disconnection period, it aggregates the requests, retrieves the content from the Internet once, and stores the content in the local cache. This aggregation can reduce bandwidth usage during network connection periods.

Cached Content Retrieval and Asynchronous Notification

The ICOW system provides other cached content to passengers when their requested content is not available (i.e., has not yet been cached). For example, when a passenger requests the *New York Times* web page but it is not cached, the passenger might be interested in other news websites that have already been cached, say, the CNN website. To provide cached contents, the

ICOW nodes actively cache content to reduce bandwidth usage by eliminating retrieval of duplicate web content. In addition, the ICOW node centrally controls traffic. It collects and handles all the requests, so the ICOW system can resolve the throughput degradation problem by reducing the number of competing nodes to one for the APs at stops.

ICOW node sends passengers a guide page which includes a list that it has cached. Figure 1a is a screenshot of a guide page a passenger receives from the ICOW node that we have implemented. Hence, the passenger can read the cached content (the CNN web page in Fig. 1b)

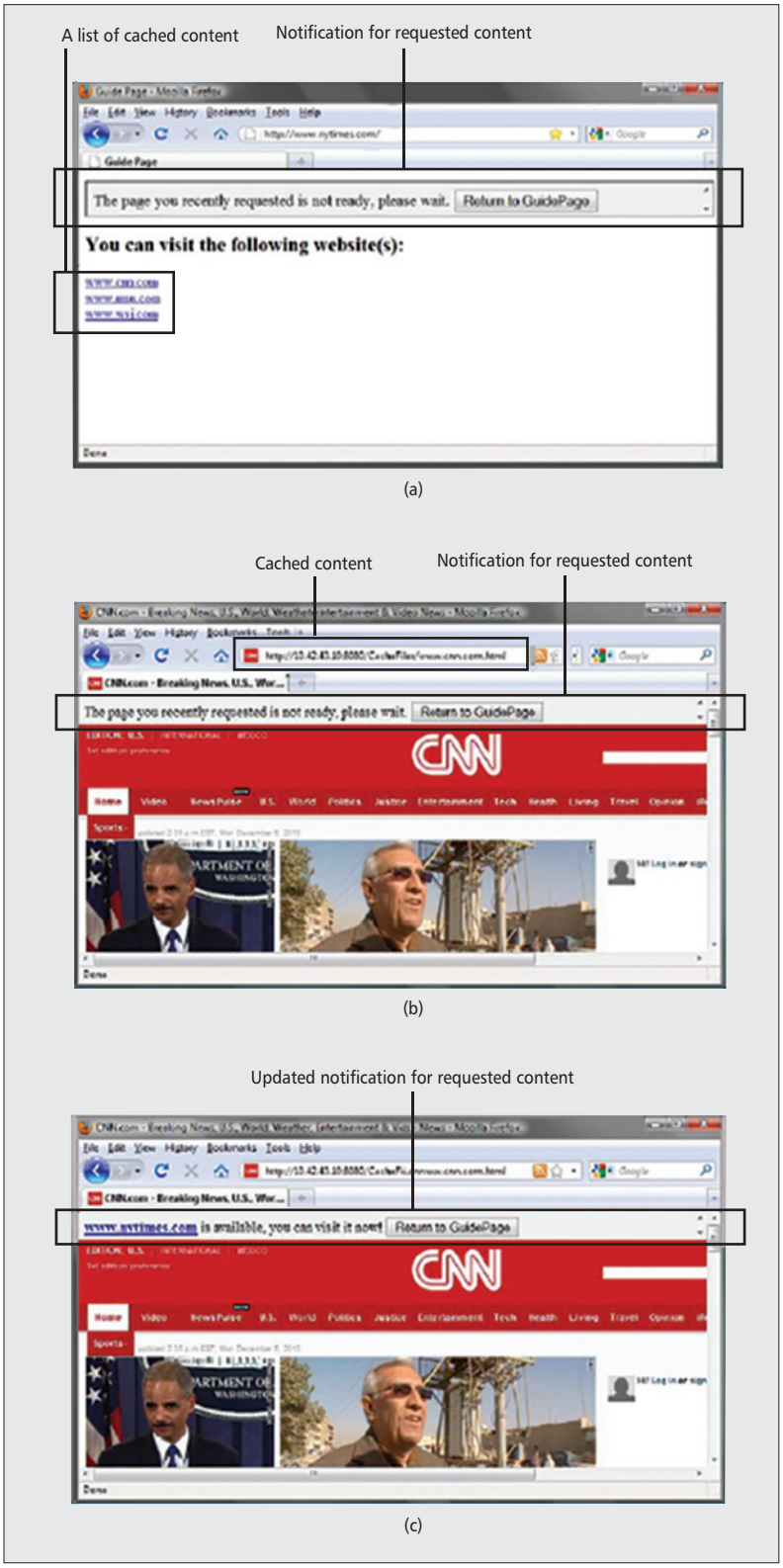


Figure 1. Screen shot of a web browser that shows a guide page and a cached page with notification: a) guide page; b) cached page; c) automatic, asynchronous notification.

while he or she is waiting for the *New York Times* web page.

However, passengers might still worry that they may miss their requested content, so they might frequently check whether their content has been cached. To make the system easier to use, the ICOW node notifies passengers when their content has been cached without having to manually check for updates. This automatic asynchronous notification is based on a server-initiated HTTP push mechanism (see the details of system architecture below). This push mechanism is compatible with web browsers, so passengers do not need to install a separate program. In Fig. 1c, if the ICOW node caches the requested page (the *New York Times* in this figure), the passenger can be known from the notification on the current web page.

Scheduling Algorithm

Popularity-Based Scheduling: When the network connection time is not long enough to handle all the pending requests, some passengers might experience a long delay in getting requested content or not receive the content at all before they get off. Hence, we need a scheduling mechanism to decrease content delivery delay. We use popularity-based scheduling in which the request with highest frequency has the highest priority. Because there might be several requests for the same content, popularity-based scheduling efficiently reduces bandwidth usage and thereby reduces average content delivery delay. To break a tie, we apply a normalized most-recent-first (MRF) algorithm in which the most recently requested content has the highest priority.

Combining the popularity and MRF algorithms, we can calculate the value of content that is used to determine the priority. The higher value the content has, the higher priority the content has. The value, $V_i(t)$, of content i at time t is

$$V_i(t) = N_i(t) + \frac{\sum_{j=1}^{N_i(t)} T_{ij}}{t \cdot N_i(t)}, \quad (1)$$

where $N_i(t)$ is the number of requests for content i at time t and T_{ij} is the requesting time of content i of passenger j ($T_{ij} < t$). Time t can be the sequence number of stops from a vehicle depot or the monotonically increasing time with a reference point that is the starting time at a depot. This scheduling algorithm is also used as a cache replacement algorithm when the cache is full.

Prefetching: Generally, web pages have hyperlinks that point to other web pages, and there is a high likelihood that passengers want to visit those linked pages from their current web pages. For example, after reading headlines of news web pages, passengers may read the articles in more detail, clicking the links of the articles. In public transit systems, however, when passengers want to read other articles after reading the current web page, the vehicle might have already left the stop. Thus, they have to wait until the next network connection period. If the traveling time of a vehicle to the next stop is long, the waiting time to obtain the article becomes high.

To reduce this waiting time, we use a prefetching mechanism. When the system obtains a web

page, if there are no more requests from passengers in a queue, it prefetches pages pointed to by hyperlinks (HTML <a> elements) on the received web page. Compared to requests from passengers, this prefetching has lower priority.

Even though prefetching has low priority, the system needs to reduce the number of prefetching web pages because the number of hyperlinks is significant on most web pages, especially front pages. Hence, we reduce the number of prefetching pages by applying rules: to eliminate advertisements, prefetched pages should be in the same domain as the requested pages, and to save bandwidth, prefetched pages should not be media elements (video or audio files).

Passenger Privacy

Cache Control: Since all requests and responses pass through the ICOW node, there might be a privacy issue. Indeed, passengers do not want their sensitive information to be cached and released to others. Hence, ICOW needs a mechanism to protect privacy.

To control caching, ICOW uses an explicit cache control mechanism provided by HTTP/1.1 [7]. There are two types of Cache-Control directives for privacy: private response directive and no-store request and response directive. By checking the Cache-Control header, ICOW decides whether it caches content or not, hence protecting passengers' privacy.

HTTP Cookie: HTTP cookies allow web servers to provide user-specific web pages by retrieving a user's history stored in web servers along with the cookies. However, a cookie causes user privacy issues because it allows first and third parties to track the activities and locations of clients. Especially, third-party HTTP cookies are the most common method to track clients. To protect user privacy, ICOW does not carry users' cookies. Thus, in the ICOW system, web pages that users receive are not user-specific pages.

HTTP over TLS: HTTP over TLS (HTTPS) provides channel security for sensitive information, so most e-commerce websites and web mail services use HTTPS. Since HTTPS provides cryptographic security between a client and a web server, the ICOW system does not cache encrypted web pages. Instead, ICOW provides HTTP CONNECT tunneling for HTTPS transactions. Thus, passengers directly access secure web pages through tunneling. The bandwidth usage for HTTPS traffic can be limited by setting up a different rate limited queue.

System Architecture and Implementation

An ICOW node consists of four software components: the network monitor, proxy server, cache, and local web server. Figure 2 shows the system architecture of an ICOW node.

Network Monitor

The main role of the network monitor is to determine whether the ICOW node can connect to an AP. The network monitor periodically checks whether network connectivity is available and notifies the proxy server component. If connectivity is available, the ICOW node tries to fetch content through the AP.

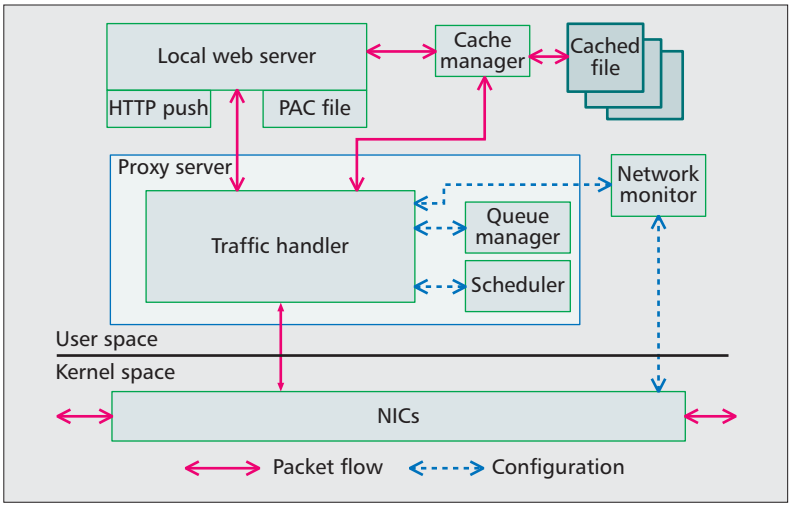


Figure 2. System architecture.

We have implemented the network monitor modules for two platforms, Windows and Linux, because these modules require a kernel-specific application programming iterface (API) to check for link connectivity. The Windows version of the network monitor module internally calls a driver level function call, DeviceIoControl, through the network driver interface specification (NDIS) and obtains the information of the network interface card (NIC). On the other hand, for Linux, we use a kernel-level system call, iwgetid, to check the status of NIC and the IP address that is bound to the MAC address of the NIC.

HTTP Proxy Server

The HTTP proxy server handles all incoming and outgoing packets. We use Muffin [8], an open source web proxy. We added three features to Muffin: traffic handler, queue manager, and scheduler. When the proxy server receives HTTP requests from passengers, it redirects the requests to the local web server if the local cache has the content. Otherwise, the requests are queued to be handled during network connection periods. Based on the popularity of requests, the priority of the requests is determined. The proxy server component triggers the cache manager, which downloads requested content in the queue when it has a network connection.

ICOW supports automatic discovery and configuration of the proxy server using Proxy Auto-Config (PAC) and Web Proxy Auto-Discovery Protocol (WPAD). A PAC file, which is in the local web server, is used to configure a proxy server, and the WPAD protocol is used to automatically find the location of the PAC file.

Cache Manager

The cache manager component downloads the requested web pages, including HTML files, images, CSS files, and JavaScript files. In addition, it prefetches other content pointed to by hyperlinks in the same domain of requested web pages. The list of cached content is accessible to the proxy server and the local web server.

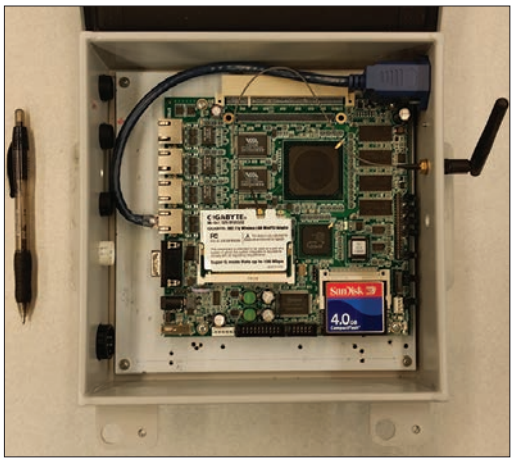


Figure 3. Soekris net5501 board. The board is in an outdoor enclosure. We open the cover to show the board.

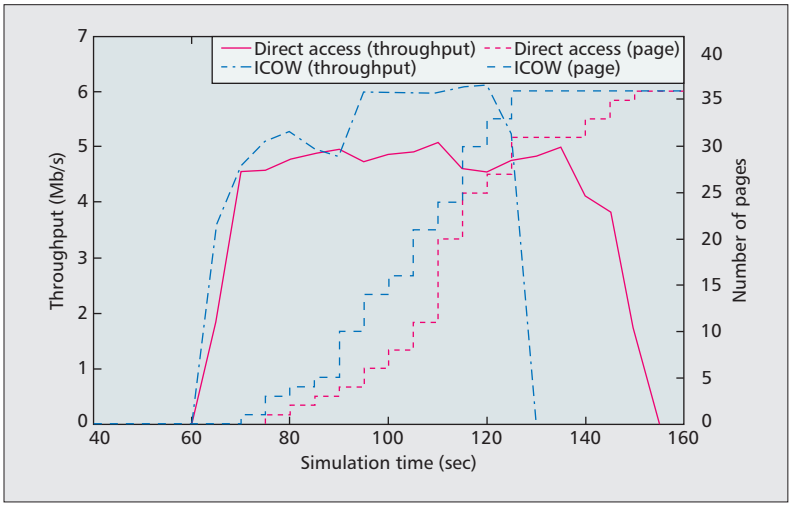


Figure 4. Downlink throughput. Channel rate is 11 Mb/s. The network disconnection period is from 0 to 60 s. The network connection starts at 60 s.

Local Web Server

The local web server provides the web-based user interface to passengers, informing them of the unavailability of the requested content and providing other cached content when the requested content is not available during a network disconnection period.

The local web server is built on Apache Tomcat, an open source software implementation of the Java Servlet and the JavaServer Page, with `Ajax Push`, a server-initiated push mechanism. The push mechanism asynchronously notifies passengers of updated information. An HTML iframe, along with delivered cached pages, is used to show the update in a passenger's browser. Visiting a guide page, passengers are able to see all the cached content.

Compact Communication Computer

We port the ICOW system to a low-cost low-power communication computer, Soekris net5501 [9], which has a 500 MHz AMD Geode LX processor, 512 MB memory, a 4 GB CompactFlash card, and two network interface cards. We run Ubuntu 10.04 on this board. Figure 3 shows the board. This low-cost low-power board can be installed at public transit vehicles to act as an ICOW node.

Performance Evaluation

We implemented the prototype of the ICOW system as shown in the previous section, but we have not deployed our system yet. Thus, we evaluate throughput, aggregation mechanism, and popularity scheduling algorithm using simulation and modeling.

Throughput

We use the OPNET modeler network simulator [10] to evaluate the throughput of the proposed ICOW system and compare it to the direct access system (in which passengers access APs directly). The objective of this simulation is to see how much ICOW can reduce network contention and utilize the available bandwidth, so we do not consider cached content in this simulation. We assume that each passenger retrieves a different web page from a different web server at randomly chosen [0–120] seconds. The number of requests in this simulation is 36. The network connection starts at 60 s, and [0–60] s is the network disconnection period. The generated HTTP requests during the network disconnection period are queued and transmitted as soon as the network connection is available, at the beginning of the network connection period (at 60 s of simulation time). We select the HTTP request size of 100 kB and HTTP response size of 1 MB for a web page, including all inline requests and responses for the inline links of the page.

Figure 4 plots the downlink throughput of HTTP responses from an AP to passengers. In the case of ICOW, at the beginning of the connection period (approximately between 60–90 s), the HTTPS requests queued during the disconnect period consume the channel resource, and there is network contention between uploading (requests) and downloading (responses). Thus, the downlink throughput is not that high compared to that between 90 and 125 s. Even in this period, however, the downlink throughput is slightly higher than that of the direct access system, and ICOW can download more pages. ICOW's throughput gain against the direct access system comes from the fact that ICOW performs fewer backoff processes than the nodes in the direct access system by reducing the number of competing nodes to one for the AP. To download all requested web pages, ICOW needs 65 s, and the direct access system needs 90 s. The result shows that ICOW allows a higher number of downloaded web pages than the direct access system within a given time.

Cooperation with DTN

DTN mechanisms focus on how to efficiently deliver contents from one node to specific nodes or how to distribute content to particular regions, while the ICOW system focuses on how to efficiently retrieve contents given limited bandwidth and time from the intermittently connected Internet. Some papers [11, 12] discuss scheduling algorithms to mitigate limited bandwidth and

time, but their algorithms are based on the likelihood to increase successful delivery ratio and decrease delivery delay using mobility patterns or social relationships with the destination. However, the ICOW system does not need the relationship with other peers or region information because it uses the intermittently connected Internet to serve web services. Thus, it might not be suitable to directly compare ICOW to other DTN mechanisms because their application scenarios and approaches are different. However, the ICOW system can be enhanced by working with those DTN mechanisms. During a network disconnection period, if an ICOW node encounters other ICOW nodes, it can retrieve contents from the ICOW nodes encountered. This peer-to-peer information sharing concept is proposed in DTN mechanisms [5]. This can reduce the bandwidth usage during the network connection period.

We evaluate how much the ICOW system and this peer-to- peer information sharing can reduce the amount of web page downloading during a network connection period compared to a direct access mechanism. The expected number of requests in a queue $E[N_R]$ is the number of requests at the Tth stop with N_c web pages when we use the request aggregation mechanism

$$E[N_R] = \sum_{k=1}^{N_c} \sum_{n=0}^{\infty} \Pr(N_r = n) P_Q(R_k, T, N_r), \quad (2)$$

where $P_Q(R_k, T, N_r)$ is the probability that a web page with rank k is in a queue when there are N_r requests at the Tth stop. We assume that the frequency of a web page with rank k is a Zipf random variable, and an independent and identically distributed random variable. The number of requests from passengers N_r is a Poisson random variable.

$$P_Q(R_k, T, N_r) =$$

$$\Pr(1_{Q_k}(N_r, T) = 1) \Pr(1_{R_k}(T) = 0), \quad (3)$$

$$\Pr(1_{Q_k}(N_r, T) = 1) \Pr(1_{R_k}(T) = 0) \Pr(1_{M_k}(T) = 0), \quad (4)$$

where Eq. 3 is for ICOW only system, and Eq. 4 is for ICOW with DTN. $1_{Q_k}(N_r, t)$ is the indicator function that shows whether the web page type k is requested at stop T when there are total N_r requests. $1_{R_k}(T)$ is the indicator function that shows whether the web page with rank k has cached by the Tth stop. $1_{M_k}(T)$ is the indicator function that shows whether the web page with rank k has cached from ICOW nodes encountered during a network disconnection period. We assume that the number of ICOW nodes encountered during a network disconnection period is a Poisson random variable. More details about these indicator functions are described in [6].

As shown in Fig. 5, both the ICOW only system and the ICOW with DTN system have smaller numbers of downloading web pages than a direct access mechanism. This means that the ICOW mechanism uses less bandwidth to handle the same requests from passengers. As the probability of encountering other ICOW nodes increases, the number of downloading web pages at stops decreases.

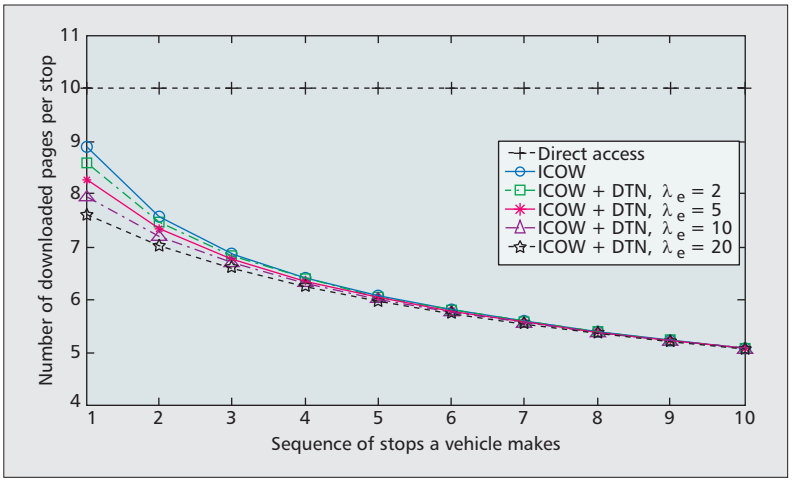


Figure 5. Expected number of downloaded web pages at each stop. The passenger arrival rate is 10 per stop. The number of web pages available N_c is 1000, and the exponent characterizing Zipf distribution is one. ICOW + DTN means the ICOW system with peer-to-peer information sharing. λ_e is the rate of ICOW nodes encountered during a network disconnection period.

Scheduling Algorithm

To evaluate scheduling algorithms, we applied two metrics: average content delivery delay and average aggregated utility. The utility function represents the value of users' satisfaction in completing jobs over time. In our system, the job is considered complete when passengers receive their requested content.

C. B. Lee *et al.* [13] conducted a survey to examine types of utility functions for real jobs on a real HPC system. They obtained three types of utility functions from the survey: step function, linear decay function, and exponential decay function. We applied these three types of utility functions to our evaluation.

The average aggregate utility, $A_i(t)$, of web page i is

$$A_i(t) = \frac{\sum_{j=1}^{N_i(t)} U_{ij}(t)}{N_i(t)}, \quad (5)$$

where $U_{ij}(t)$ is the utility of content i of passenger j at time t and $N_i(t)$ is the number of requests for content i at time t.

We use the Monte Carlo method in order to evaluate performance due to the complexity of analysis. Shown in Table 1, the parameters that we use in our analysis are passenger arrival rate, passenger travel time, successful rate of handling requests, number of web pages available, popularity of web pages, and utility function of web pages. We assume that each passenger requests one web page (including HTML files, images, CSS files, JavaScript files) while boarding the bus or train. We use a discrete time value, so the time value n means the nth stop from a vehicle departing from its depot. If passengers request web pages at the same stop, the generation times of those requests are the same.

We evaluate five scheduling algorithms: first come first served (FCFS), FCFS with popularity for tie-breaker (FCFS+popularity), popularity with FCFS for tie-breaker (Popularity+FCFS),

Parameters		Passenger arrival (passengers/stop)	Passenger travel time (stops)	Rate of handling requests at each stop	Popularity of web pages	Content types (utility function)
	Distribution	Poisson	Geometric	Uniform	Zipf	Discrete uniform
	Range	Mean: [10, 50]	Mean: [4, 10]	[0.5, 1.0]	Uniform distribution of *s*: [0, 1]	[step, linear decay, exponential decay]
Result	Algorithm	FCFS	FCFS + Popularity	Popularity + FCFS	Popularity + MRF	Max utility + MRF
	Average aggregate utility	0.67	0.67	0.68	0.77	0.82
	Average deliver delay (stops)	2.10	2.10	2.10	1.19	1.15
	Number of passengers who receive content	21.89	21.89	21.90	21.98	22.09

Table 1. Monte Carlo method simulation. The number of web pages is 100,000. The number of stops is 50. The number of simulations is 200. Main + Tie-breaker. MRF stands for most-recent-first.

popularity with most-recent-first (MRF) for tie-breaker (Popularity+MRF), and maximizing utility with MRF for tie-breaker (Max utility+MRF).

According to Table 1, Max utility+MRF has the best performance, as we expected. This scheduling algorithm is ideal because it reflects utility functions of all the requests to maximize satisfaction of passengers. However, it is based on the strong assumption that this algorithm knows the utility value of each request. In real implementation, however, it is almost impossible to satisfy this assumption.

Our goal is to propose an algorithm that has as close performance as Max utility+MRF while the proposed algorithm supports easy implementation. Table 1 shows that our proposed scheduling algorithm, Popularity+MRF, is close to the performance of Max utility+MRF in terms of average aggregate utility and delay. Because of the MRF function of Popularity+MRF, it has similar average aggregate utility value as Max utility+MRF. Because of MRF function, those two algorithms have shorter delivery delay than others. However, the popularity-based algorithm can easily be implemented by checking the number of requests at an ICOW node. Thus, considering the complexity of the implementation, Popularity+MRF might be suitable in terms of ease of implementation and performance.

Discussion

HTTP/2: HTTP/2 uses the existing semantics of HTTP/1.1 [14]. Thus, the core functionalities of the ICOW system, such as caching mechanism and asynchronous notification, are compatible with HTTP/2.

Deployment: Instead of having an AP, transportation stations might have a femtocell, a small cellular base station, to provide cellular network service. However, this femtocell has shorter range than WiFi, so passengers still experience intermittently connected network service. Thus, the ICOW system is feasible for enhancing passengers' QoE under this network environment. For the connection to the femtocell at stations, the ICOW system needs to have additional cellular network interface and needs a new network monitor module for the cellular network interface. However, the other core components of the ICOW system will remain the same.

Categories of Web Sites: The guide page of the ICOW system shows all the web sites it has cached. This might be a bit messy for passengers to find interesting pages. Thus, we might need to categorize the web pages so that passengers can easily find their preferred sites. Third-party databases such as OpenDNS domain tagging [15] can be used to categorize the web sites based on their domain names.

Starvation: Popularity-based scheduling algorithm might cause starvation for retrieving unpopular pages when the ICOW system does not have enough time to retrieve all the requested pages. We believe that the guide page, which shows the cached web content, of the ICOW system might be helpful for passengers in starvation to find other popular cached content from our ICOW system.

Conclusion

To maximize content availability and system usability in intermittently connected networks for public transit systems, we propose ICOW. This system enhances link throughput and resolves network resource limitations. Not only does this allow for maximizing network resource utility and minimizing bandwidth usage, but it also offers an easy-to-use system for passengers. Our evaluation shows that ICOW increases link throughput compared to a direct access system. Our caching-based mechanism allows more passengers to get more content with limited network resources. By using a popularity-based scheduling mechanism, ICOW provides higher satisfaction in obtaining content for passengers and decreases content delivery delay.

ACKNOWLEDGMENT

This work is supported by a grant from InterDigital Communications, LLC. The authors would like to thank the reviewers for their helpful and valuable comments and suggestions.

REFERENCES

[1] Metro's Cellphone Installation in Trouble; http://washingtonexaminer.com/metros-cellphone-installation-in-trouble/article/2520922
[2] Wireless Services at NYC Subway Stations; http://web.mta.info/nyct/service/WirelessServiceAtSubwayStations.htm
[3] A. Balasubramanian, B. N. Levine, and A. Venkataramani, "Enhancing Interactive Web Applications in Hybrid Networks," *Proc. ACM MobiCom*, San Francisco, CA, Sept. 2008.
[4] J. Ott and D. Kutscher, "Bundling the Web: HTTP over DTN," *Proc. Wksp. Networking in Public Transport*, Waterloo, Ontario, Canada, Aug. 2006.
[5] S. Srinivasan *et al.*, "7DS — Node Cooperation and Information Exchange in Mostly Disconnected Networks," *Proc. IEEE ICC*, Glasgow, Scotland, June 2007.
[6] S. Hong, *Mitigating Network Service Disruptions in High-bandwidth, Intermittently Connected, and Peer-to-Peer Networks*, Ph.D. dissertation, EE Dept., Columbia Univ., 2011.
[7] R. Fielding *et al.*, "Hypertext Transfer Protocol HTTP/1.1," IETF RFC 2616, June 1999.
[8] Muffin: World Wide Web Filtering System; http://muffin.doit.org/
[9] Soekris net5501; http://soekris.com/products/net5501.html
[10] OPNET modeler; http://www.opnet.com
[11] J. Burgess *et al.*, "MaxProp: Routing for Vehicle-Based Disruption-Tolerant Networks," *Proc. IEEE INFOCOM*, Barcelona, Spain, Apr. 2006.
[12] A. Balasubramanian, B. N. Levine, and A. Venkataramani, "DTN Routing as a Resource Allocation Problem," *Proc. ACM SIGCOMM*, Kyoto, Japan, Aug. 2007.
[13] C. B. Lee and A. E. Snavely, "Precise and Realistic Utility Functions for User-Centric Performance Analysis of Schedulers," *Proc. IEEE Int'l. Symp. High Performance Distrib. Comp.*, Monterey, CA, June 2007.
[14] M. Belshe, R. Peon, and M. Thomson, Eds., "Hypertext Transfer Protocol version 2," IETF Internet Draft, Aug. 2015, work in progress.
[15] OpenDNS Domain Tagging; https://community.opendns.com/domaintagging/

BIOGRAPHIES

SE GI HONG is a Senior Member of Technical Staff at Hughes Network Systems, LLC, Germantown, Maryland. He received his M.S. degree in electrical engineering from the University of Southern California (USC), Los Angeles, and his Ph.D. degree in electrical engineering from Columbia University, New York, New York. He was a member of the Internet Real-Time (IRT) Lab at Columbia University. His research interests include mitigating network service disruptions, heterogeneous/hybrid networks, software-defined networking (SDN), and secure network architecture.

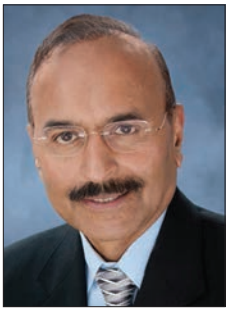
SUNGHOON SEO received M.S. and Ph.D. degrees in computer science from Yonsei University, Seoul, Korea, in 2002 and 2009, respectively. He worked at the IRT laboratory, Columbia University, as a postdoctoral research scientist during 2009 and 2011. Since 2011, he has been a principal researcher at Infra. Laboratory, Korea Telecom. His research interests include mobility and resource management, and network function virtualization on heterogeneous wireless networks.

HENNING SCHULZRINNE [F] is Julian Clarence Levi Professor of Computer Science at Columbia University. He received his Ph.D. from the University of Massachusetts in Amherst, Massachusetts. He was a member of technical staff at AT&T Bell Laboratories and associate department head at GMD-Fokus (Berlin), before joining the Computer Science and Electrical Engineering Departments at Columbia University. He served as chair of Computer Science from 2004 to 2009 and as chief technology officer of the Federal Communications Commission (FCC) from 2012 to 2014. Protocols co-developed by him, such as RTP, RTSP, and SIP, are now Internet standards, used by almost all Internet telephony and multimedia applications.

PRABHAKAR CHITRAPU, Ph.D., works at AT&T in the Small Cell Platforms team. He holds leadership positions at the Small Cell Forum, including Chairman of the Network Working Group. Earlier, he worked at InterDigital, Lockheed Martin, Dialogic (Intel), Bharat Electronics, Drexel University, Villanova University, and the University of Pennsylvania (as an adjunct professor). He authored a book, *Wideband TDD* (Wiley), published over 30 papers, and was awarded over 50 patents.

Our caching-based mechanism allows more passengers to get more content with limited network resources. By using a popularity-based scheduling mechanism, ICOW provides higher satisfaction in obtaining content for passengers and decreases content delivery delay.

RADIO COMMUNICATIONS: COMPONENTS, SYSTEMS, AND NETWORKS



Amitabh Mishra

Tom Alexander

Coding has long been fundamental to communication systems, even as far back as the advent of telegraphy in the 1840s: Samuel F. B. Morse created the Morse code as a means of efficiently encoding and transmitting messages in the English language. Channel coding has advanced considerably since the inauguration of information theory by Claude Shannon in 1948, followed by the seminal work by Richard Hamming on practical error correcting codes in 1949. Many crucial algorithms utilized in radio communications today, including convolutional and block codes, the Viterbi algorithm, BCH and Reed-Solomon codes, soft decoding, and so on, date back to the 1960s. Much of the interest in channel coding was driven by space research, to solve the formidable problem of communicating with space probes millions or even billions of kilometers from Earth.

Coding today has moved from being an esoteric branch of information theory, mostly of interest to space researchers and long-distance communication system designers, to being central to digital communications systems of all kinds. Specialized codes are used for error correction, compression, cryptography, modulation, and signal shaping. With the advent of ever higher data rates and more efficient use of spectrum, a great deal of research has been conducted on codes having different properties that make them suitable for different communications or information related applications. The low-density parity check codes in IEEE 802.11 PHYs come very close to the Shannon channel capacity limit.

A modern digital communications system utilizes various kinds of codes at different positions along the transmit chain. The very first step is usually to increase the entropy of the digital data using a pseudorandom scrambling code, which simplifies timing recovery and flattens the power spectral density of the transmitted signal. After this, a forward error correction (FEC) code is typically applied to allow bit errors induced during transmission to be corrected at the receiver without retransmission. Subsequently, a quadrature amplitude modulated (QAM) mapping code is applied to encode blocks of digital bits into symbols in a constellation prior to transmission, thus performing the process of modulation.

For a multiple-input multiple-output (MIMO) system, a space-time block code is then used to map the constellation points to different space-time streams that are transmitted simultaneously, allowing the effective channel capacity to be multiplied. If beamforming is required, a beamforming code is then utilized to improve the signal-to-noise ratio at the receiver. A spectral shaping code is usually resorted to, in order to make the transmitted signal conform to regulatory requirements and stay within the constraints of the analog signal path.

The above brief description alone has brought out six different points at which six different codes are applied in order to create a transmit signal; coding is obviously fundamental to the entire process. As communications systems attempt to squeeze out every last drop of efficiency from the radio channel, coding theory is even utilized to aid or improve techniques that are normally the domain of protocol designers. A classic example is code-division multiple access (CDMA), where the medium access itself is enabled by specialized codes. A more recent example is the hybrid automatic repeat request (HARQ) technique in LTE systems, which combines traditional FEC and ARQ for increased transmission efficiency. As another example, a great deal of work is ongoing in utilizing innovative codes to improve the efficiency of transmit power amplifiers, such as constant envelope coding or predistortion.

In this edition of the Radio Communications Series, we bring you two papers dealing with interesting aspects of applied coding theory, providing both a survey of the state of the art in the field as well as some recent analytical work.

The first article, "Coded Random Access: Applying Codes on Graphs to Design Random Access Protocols," deals with a novel way of viewing coding theory in the context of radio medium access, more specifically an adaptation of the well-known slotted ALOHA approach. The pioneering ALOHA radio system was developed in 1971,

and variations of the basic algorithm are still in use today. This article discusses a method of applying successive interference cancellation — a far more recent development — to this venerable medium access protocol, together with erasure coding and graph theory in order to increase the efficiency of the protocol. The intention is to recover information from collision slots, which would otherwise represent wasted channel capacity.

Our second article also involves coding, but in an entirely different dimension. The issue of controlling and shaping out-of-band emissions for orthogonal frequency-division multiplexing (OFDM) modulation is a significant challenge for digital transmission systems. The goal of reducing adjacent channel interference usually conflicts with the need to maximize spectral efficiency and increase power amplifier efficiency via a low peak-to-average-power ratio (PAPR). The article "Out-of-Band Emission Reduction and a Unified Framework for Precoded OFDM" surveys the existing approaches to reducing out-of-band emissions, and then demonstrates that many of these methods can be modeled by a simple generalized linear framework. The authors then utilize their framework to define a lower-complexity approach to out-of-band emissions, and provide some analytical results.

In future issues, we plan to put more such tutorial and research papers before you. For this to happen, of course, we need high-quality submissions; we therefore encourage our readership to contribute survey or tutorial papers concerning their areas of expertise or recent developments to the Radio Communications Series.

BIOGRAPHIES

THOMAS ALEXANDER [M] (talexander@ixiacom.com) is a senior architect at Ixia. Previously, he has worked at VeriWave Inc. (acquired by Ixia), PMC-Sierra Inc, and Bit Incorporated (acquired by PMC-Sierra), and prior to that was a research assistant professor at the University of Washington. He has been involved in various aspects of wired and wireless networking R&D since 1992, in the areas of ATM, SONET/SDH, Ethernet, and (since 2002) wireless LANs. He is also active in standards development, and has served as Editor of IEEE 802.3ae, Chief Editor of IEEE 802.17, and Technical Editor of IEEE 802.11. He received his Ph.D. degree from the University of Washington in 1990.

AMITABH MISHRA [SM] (amitabh@cs.jhu.edu) is a faculty member at the Information Security Institute of Johns Hopkins University, Baltimore, Maryland. His current research is in the area of cloud computing, data analytics, dynamic spectrum management, and data network security. In the past he has worked on the cross-layer design optimization of sensor networking protocols, media access control algorithms for cellular-ad hoc interworking, systems for critical infrastructure protection, and intrusion detection in mobile ad hoc networks. His research has been sponsored by NSA, DARPA, NSF, NASA, Raytheon, BAE, APL, and the U.S. Army. In the past, he was an associate professor of computer engineering at Virginia Tech and a member of technical staff with Bell Laboratories, working on the architecture and performance of communication applications running on the 5ESS switch. He received his Ph.D. in electrical engineering from McGill University. He is a member of ACM and SIAM. He has written 80 papers that have appeared in various journals and conference proceedings, and holds five patents. He is the author of a book, *Security and Quality of Service in Wireless Ad Hoc Networks* (Cambridge University Press, 2007), and a Technical Editor of *IEEE Communications Magazine*.

Coded Random Access: Applying Codes on Graphs to Design Random Access Protocols

Enrico Paolini, Čedomir Stefanović, Gianluigi Liva, and Petar Popovski

ABSTRACT

The rise of machine-to-machine communications has rekindled interest in random access protocols as a support for a massive number of uncoordinatedly transmitting devices. The legacy ALOHA approach is developed under a collision model, where slots containing collided packets are considered as waste. However, if the common receiver (e.g. base station) is able to store the collision slots and use them in a transmission recovery process based on successive interference cancellation, the design space for access protocols is radically expanded. We present the paradigm of coded random access, in which the structure of the access protocol can be mapped to a structure of an erasure-correcting code defined on a graph. This opens the possibility to use coding theory and tools for designing efficient random access protocols, offering markedly better performance than ALOHA. Several instances of coded random access protocols are described, as well as a case study on how to upgrade a legacy ALOHA system using the ideas of coded random access.

INTRODUCTION

We start with a deceptively simple question: When and why should we use random access? A concise answer would be: Whenever there is an uncertainty about the set of users that aim to transmit at a given instant. A canonical scenario falling in the above description is the one in which a set of uncoordinated devices aims to transmit over the shared wireless medium to the same receiver at approximately the same time, and the random access mechanisms are needed to break this "symmetry" and enable successful access. As such, random access is an essential component of any distributed wireless communication system, typically used for initial link establishment or distributed spectrum sharing among interfering networks, such as two collocated WiFi hotspots. Presently we are witnessing a revival of research interest in random access mechanisms, driven by the increasing presence of *machine to-machine* (M2M) communications in cellular and satellite networks. Efficient random access is instrumental in M2M scenarios, due to the fact that there is a massive and uncoordinated set of transmitting devices.

ALOHA [1] is a rather generic form of random access, typically operating under the assumption that collided packets are irrecoverably lost. Standard variants of the ALOHA protocol aim to maximize the number of collision-free transmissions within a given time interval, that is, to maximize the expected throughput. In slotted ALOHA (SA) [1], link time is divided into equal-duration slots, and the devices are slot-synchronized, contending for access on a slot basis with a predefined slot-access probability. A related solution is framed slotted ALOHA (FSA) [2], where slots are organized into frames, and the users transmit in a single, randomly chosen slot of the frame. In both variants, only the slots containing a single transmission (singleton slots) are useful and the corresponding transmission is successfully received, while the slots containing no transmission (idle slots) or multiple user transmissions (collision slots) are wasted. The throughput T, defined as the probability of successfully receiving a user transmission per slot, is equal to the probability that a singleton slot occurs. The maximal asymptotic throughput in both variants is a rather low $T_{\max} = 1/e \approx 0.37$.

Recently there has been a conceptual shift in the theory and practice of the slotted ALOHA protocol family, based on the use of successive interference cancellation (SIC) that enables "unlocking" of the collisions slots. Some of these advances apply SIC at the slot level, in order to separate the collided signals and allow multiple packets to be received within a single slot, c.f. [3], which may be regarded as an instance of multi-user detection (MUD). These access protocols, applied also to combat the hidden terminal problem in carrier sensing multiple access (CSMA) systems [4], still rely on an instantaneous feedback from the receiver, notifying the transmitters about unresolved collisions and initiating retransmissions. Other recent advances consist of combining SA with physical layer network coding [5].

Enrico Paolini is with the University of Bologna, Italy.

Čedomir Stefanović and Petar Popovski are with Aalborg University, Denmark.

Gianluigi Liva is with the Institute of Communications and Navigation, DLR, Germany.

0163-6804/15/$25.00 © 2015 IEEE

This work is dedicated to a conceptually different improvement, based on SIC across multiple slots [6]. The essence of these modifications is rather simple: active devices transmit replicas of the same packet in multiple slots, while SIC is used on the receiving side to remove replicas of already recovered transmissions from collision slots. Recovery and removal of replicas is performed in an iterative, that is, successive manner, where new iterations are propelled by the transmissions recovered in the previous round, as illustrated in Fig. 1. The exploitation of the collision slots boosts the throughput. In a basic scenario where active devices transmit two replicas in randomly selected slots of a frame [6], the asymptotic throughput increases to $T_{\max} \approx 0.55$.[1] The true potential of the SIC-enabled slotted ALOHA was revealed in [7], identifying analogies with modern channel coding based on sparse graphs and establishing the paradigm of *coded random access*.

The objective of this article is to introduce these new developments, identify the ways in which they can be beneficial for M2M applications, and highlight the important implementation issues. The outlined concepts are applicable in all systems that exploit slotted ALOHA, for example, in random access channels of the cellular access and of the next generation interactive satellite services.

Basics of Coded Random Access

Access Scheme Description

We start by considering coded slotted ALOHA (CSA) in which the access is organized in *contention periods*. Each contention period is a frame containing M slots of equal duration, where M is fixed. A set of N users uses contention periods to communicate with a base station (BS), which acts as a common receiver. We are interested in the regime where the user population is large with respect to the size of the contention period $N \gg M$, but only a subset N_a of the users is active in a given contention period. A simple model to create the uncertainty in the set of active users can be described as follows. At the beginning of a contention period each user independently generates a packet to be transmitted with *activation probability* p_a, where $p_a \ll 1$. The number of active users in a contention period N_a is then a binomially distributed random variable, with mean value $\bar{N}_a = p_a N$.

The CSA scheme works as follows. Each active user generates d packet replicas, where the repetition rate d is drawn randomly according to a pre-determined probability distribution, which is the same for all users. The repetition rate is picked by an active user independently of all other active users and independently of all his previous choices. The d replicas are then transmitted by the user over d slots picked uniformly at random among the M slots of the contention period. Following the example of Fig. 1, the users 2 and 3 picked a repetition rate $d = 2$, while user 1 did not replicate its packet, that is, its repetition rate is $d = 1$. A packet in a singleton slot is decoded correctly. Each packet is assumed to contain pointers to describe the positions of the other replicas in the contention period sent by the same user.[2] The packet is then re-encoded and re-modulated and the receiver removes its interference contribution from the $d - 1$ slots containing the replicas. The process proceeds iteratively, that is, recovered replicas may lead to solving other collisions, as illustrated in Fig. 1.

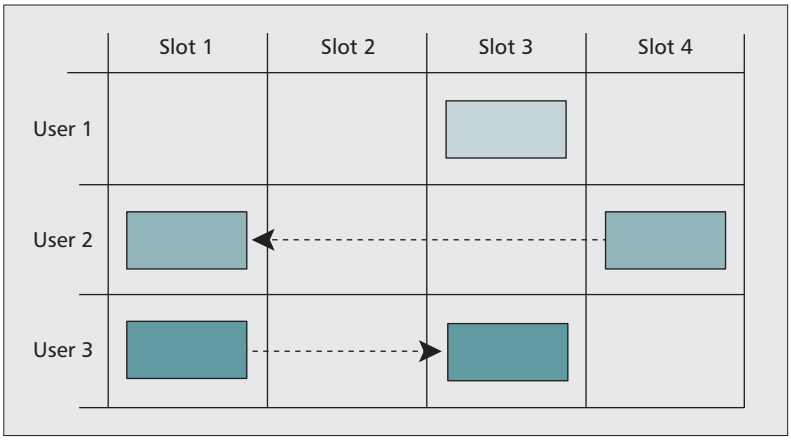


Figure 1. Successive interference cancellation in slotted ALOHA. Packet of user 2 is recovered in slot 4, enabling the recovery of packet of user 3 in slot 1, performed by subtracting the replica of user 2 packet in slot 1. In the same way, recovery of packet of user 3 enables the removal of its replica from slot 3, thus recovering packet of user 1. In this example, the use of SIC grants throughput of 0.75 packet/slot; without SIC, the throughput drops to 0.25 packet/slot.

The *rate* of the the CSA scheme is defined as

$$R = 1/\bar{d} \tag{1}$$

where $\bar{d}$ is the average number of replicas sent per user. Obviously, a lower rate implies a higher number of repetitions and the use of more energy per useful bit. The *logical* load of the channel is defined as the expected number of active users per slot

$$G = \frac{\bar{N}_a}{M} = p_a \frac{N}{M}, \tag{2}$$

that is, the logical load corresponds to the expected number of new packets generated during the contention period. The *physical* load of the channel, that is, the expected number of all transmitted replicas, is given by $G_{\text{phy}} = G \cdot \bar{d}$. In standard FSA there is only a single replica $d = 1$, thus the logical and physical loads coincide.

Bipartite Graph Representation and Asymptotic Analysis over a Collision Channel

Figure 2a shows the graph representation of the CSA scheme for the example in Fig. 1. Specifically, it is represented by a bipartite graph, consisting of a set N_a of *user nodes* (one for each active user), a set of M *slot nodes* (one for each slot), and a set of edges. An edge connects the ith user node (UN) and the jth slot node (SN) if and only if the user i sends a packet in the jth slot. The *degree* d of a given UN is equal to the number of edges connected to it, each edge corresponding to one of the replicas sent by the user. This graphical representation enables the

[1] One may argue that the comparison with standard FSA is unfair, as in FSA a user sends only one packet replica before receiving feedback on the contention outcome. However, it should be noted that in standard FSA a user may also transmit multiple replicas in order to get the data through, the difference is that the retransmission is initiated by the feedback.

[2] An efficient way to transport pointers is discussed later.

The decoder consists of initializing the status of all UNs to "unknown" and of repeating the following procedure until the status of all UNs has been updated to "known," in which case decoding terminates successfully, or until at some iteration the status of no UN is updated, when a failure is declared.

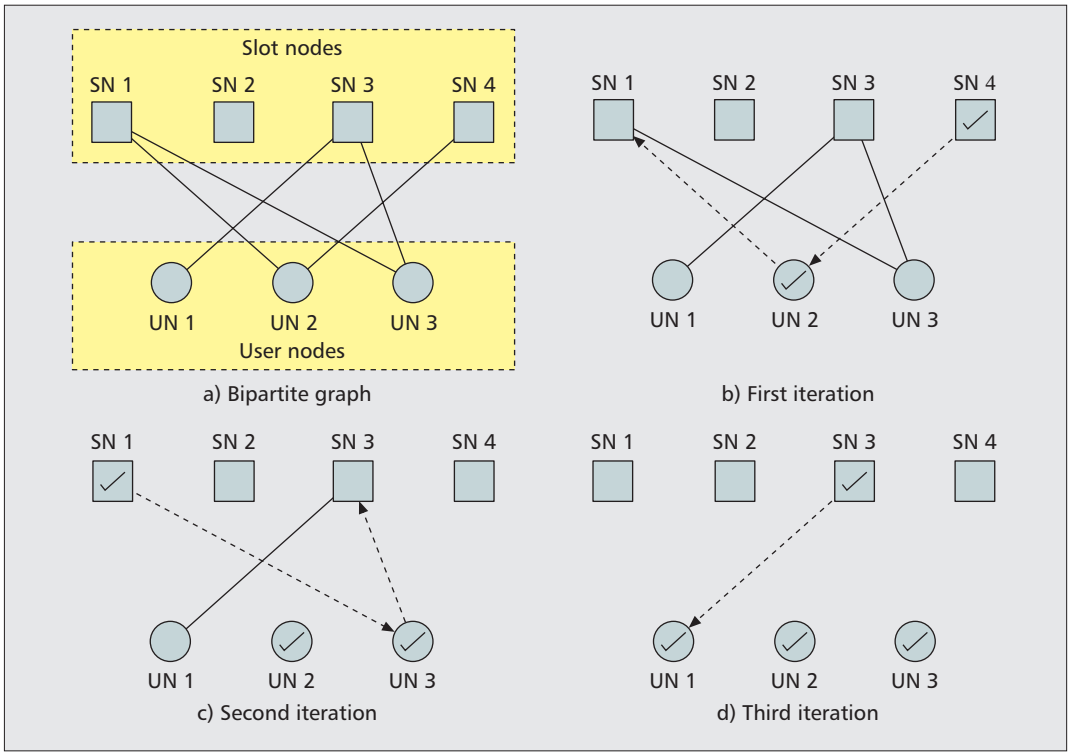


Figure 2. Bipartite graph representation of the access scheme of Fig. 1 and of the SIC process: a) bipartite graph; b) first iteration; c) second iteration; d) third iteration.

establishment of a connection between the SIC procedure and iterative decoding of channel codes based on sparse graphs. This connection is here illustrated under the following assumptions:

- For each slot, the receiver always discriminates between a "silence," singleton or a collision.
- When a packet is received in a singleton slot, data are always correctly decoded.
- Channel estimation and the interference cancellation are ideal.

The first two assumptions are typical for *collision channel* models. The channel estimation has to be performed to enable SIC and the third assumption simplifies the analysis without substantially affecting the obtained results, as shown in [6]. Further, error events due to fading and thermal noise may affect the performance; in this regard, the reader may refer to [6, 7]. We also outline the main difference to codes on graphs: the degree of a slot node cannot be controlled and it can even be equal to zero (idle slot). Clearly, if the BS could control the degree of each slot, we would not need random access at all, as a single user would be scheduled in each slot.

Under the above assumptions, the SIC procedure may be described as an instance of the iterative *peeling decoder* for codes constructed on sparse graphs and transmitted over a binary erasure channel (BEC) [8]. The decoder consists of initializing the status of all UNs to "unknown" and of repeating the following procedure until the status of all UNs has been updated to "known," in which case decoding terminates successfully, or until at some iteration the status of no UN is updated, when a failure is declared. The procedure is described as follows:

- For all SNs, if the SN has degree 1 then update to "known" the status of the unique UN connected to it.
- Remove all edges connected to the UN and update the degrees of the SNs accordingly.

The manner in which the SIC mimics the peeling decoder is illustrated in Fig. 2b–d.

The analogy between SIC for CSA and iterative decoding of codes on sparse graphs enables the use of techniques developed in the field of coding theory and applying them to random access. Accordingly, collisions are favored by CSA, in a statistically controlled manner. For example, the theory of codes on graphs makes it possible to properly design the probability distribution with which the users select their degrees to generate bipartite graphs on which SIC is successful with high probability. Judiciously designed probability distributions yield irregular graphs favoring the SIC procedure. Moreover, through the application of analytical tools from the theory of codes on graphs, such as density evolution or extrinsic information transfer (EXIT) charts, we can show the existence of a thresholding behavior of CSA under SIC. This happens when both the frame size M and the user population size N tend to infinity, but the ratio N/M remains constant. It turns out that there exists a threshold value G^*, such that when the logical load is $G \leq G^*$, the SIC procedure almost certainly terminates successfully, that is, each active user manages to send the packet to the BS within the contention period. Conversely, if $G > G^*$ then the opposite is true, that is, there is a fraction of users' packets that will certainly not be successfully delivered. It is possible to show that the threshold G^* depends

both on the selected user rates and on the probabilities with which these rates are selected. With a suitable selection of the repetition rates and their associated probability distribution, a threshold as large as $G^* = 1$ packet/slot can be achieved [9]. In other words, the throughput performance becomes equivalent to the perfectly scheduled access! The way the rate distribution is optimized follows the footsteps of the degree distribution optimization algorithms used in the design of low-density parity-check (LDPC) codes [10].

As both the threshold G^* and the rate R are functions of the repetition rates distribution, one may look for the maximum achievable threshold G^* for a given rate R. Note that when repetition coding is used, the rate is necessarily $0 < R \leq 1/2$, as there are at least two repetitions. Once R, as defined in Eq. 1, is fixed, it can be shown that the threshold G^* of a CSA scheme is upper bounded by the unique positive real solution of the equation

$$G = 1 - e^{(-G/R)}, \quad (3)$$

as shown in [10]. If the user invests more energy by increasing the number of repetitions, then R decreases and the right-hand side of Eq. 3 increases, also implying that the upper bound increases.

Variants of CSA

High-Rate CSA from Generic Component Codes

The upper bound resulting from Eq. 3 is valid for every rate R between 0 and 1. In order to achieve rates $R > 1/2$, [10] introduces a generalization of the CSA protocol that uses generic linear block codes instead of repetition codes. In this setting, a user that is active in a given contention period splits his packet into k *segments* of the same length. The k segments are then encoded using a linear block code and d segments are obtained as output. The linear block code is drawn randomly by the user from a set of component codes, according to pre-determined probability distribution. The information about the code used to encode the k segments may be conveyed in a header appended to each segment. The component codes may have different lengths d, but they all have the same dimension k. The rate of this generalized scheme is given by $R = k/\bar{d}$, where $\bar{d}$ is the expected length of the employed component code. This definition of the rate coincides with that given in Eq. 1 when repetition codes are used. With a judicious selection of k, of the lengths d of the component codes and of their probability distribution, any rate $0 < R < 1$ can be obtained. Note that the choice $k = 1$ reduces this generalized framework to the repetition-based case.

The d encoded segments, equipped with appropriate pointers in their headers, are transmitted over d slots picked uniformly at random within the contention period. The contention period is now organized into kM slots, each of the same time duration as that of a segment; the time duration of the contention period is thus the same as in the repetition-based case. The bipartite graph representing the access scheme is now composed of kM SNs and N_a UNs, where now each UN corresponds to k segments. On the receiver side SIC is performed similarly to the repetition-based case, the only difference being the execution of some form of erasure decoding at the generalized UNs at each iteration. In case simple codes are used, maximum a-posteriori (MAP) erasure decoding may performed. Similar to the case with repetition, a thresholding phenomenon is also observed for the high-rate CSA.

Spatially Coupled CSA

A variant of the CSA scheme is based on *spatial coupling*, a technique widely used in the field of modern error correcting codes. We present it in a simplified scenario in which all users exploit the same packet repetition rate d.

In the *spatially coupled CSA*, a user becoming active at the beginning of a contention period with M slots is allowed to transmit only one replica in that period, as opposed to the scheme described above in which all d replicas are transmitted in that contention period. Each of the other $d - 1$ replicas is transmitted by the user in one of the subsequent $d - 1$ periods. Assuming the average number of active users per contention period is $\bar{N}a = p_aN$, on average there are p_aN packet replicas in the first contention period (one per active user), $2p_aN$ packet replicas in the second contention period (one per user becoming active at the beginning of the first period and one per user becoming active at the beginning of the second period), and so on up to the d-th contention period in which we expect dp_aN packet replicas on average. The expected number of replicas in a contention period that comes after the d-th one "stabilizes" to dp_aN. Thus the expected physical load is $G_{phy,1} = G$ in the first contention period (see Eq. 2), then it is $G_{phy,2} = 2G$ in the second contention period, and so on, and stays $G_{phy,d} = dG$ from the d-th period and onward.

As shown in [11], the probability of a collision in a slot that belongs to a given contention period increases with the physical load imposed on that period. Due to the lighter physical load, the first contention period contains a lower number of collisions. The packets received in singleton slots of the first contention period may be used to remove the contribution of interference of their replicas in all $d - 1$ subsequent contention periods. Therefore, although a slightly higher number of collisions are expected in the second contention period, some of them are resolved by interference cancellation. The resolved collisions are exploited, together with the packets received in the singleton slots from the first and second periods, to resolve further collisions in the third period. This process, when iterated through the sequence of contention periods, determines a "chain reaction" that makes it possible to resolve more collisions than those resolved by the scheme above for the same repetition rates and probability distribution. Moreover, a thresholding phenomenon is again observed. Specifically, the G^* of the spatially coupled scheme reaches the theoretical, upper-bound threshold of the block scheme under optimal, MAP decoding on a-priori known graph![3]

Although a slightly higher number of collisions are expected in the second contention period, some of them are resolved by interference cancellation. The resolved collisions are exploited, together with the packets received in the singleton slots from the first and second periods, to resolve further collisions in the third period.

[3] We again stress the fact that in CSA the graph is not known a priori due to the randomness of the contention process.

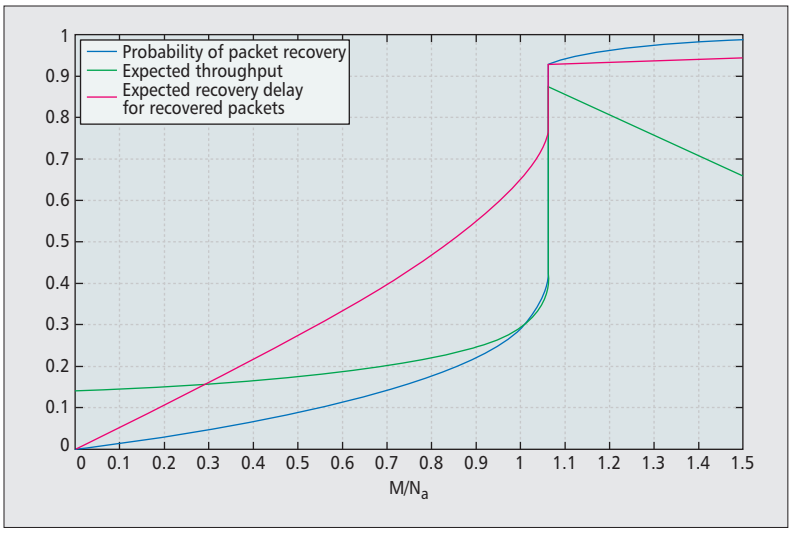


Figure 3. Asymptotic performance of frameless CSA.

Frameless CSA

Finally, we introduce *frameless* ALOHA [12], a variant of the CSA scheme inspired by the rateless codes [13]. Two essential differences with the previously described CSA protocols are:

- When the contention period starts, the active users decide whether or not to transmit on a slot basis, as the slots "appear" on the wireless medium.
- The contention period duration is not a-priori determined, but it is adaptive and tuned to the evolution of the contention/packet-recovery process.

In a general case, both the user access strategy (i.e. the choice of slot-access probabilities) and the contention termination criterion are subject to optimization. In [14] a simple version of the scheme is investigated, where the access strategy is "memoryless" and the slot-access probabilities are uniform both over users and slots. The scheme uses a heuristic termination criterion: the receiver monitors both the instantaneous throughput and the fraction of resolved user packets and, when either of them surpasses a predefined threshold, the contention is terminated through a suitable feedback signal. It was shown that although asymptotically suboptimal, this approach grants throughputs that are the highest in the reported literature for a low to moderate number of active users, that is, when N_a is in the range 50 – 1000.

Figure 3 illustrates the asymptotic performance of frameless ALOHA, showing the probability of packet recovery, expected throughput, and expected recovery delay of recovered packets, as functions of the number of elapsed slots vs. the number of active users M/N_a. The slot-access probability in the example is set to $3.1/N_a$, a value that maximizes the expected throughput [12]. It is seen that the probability of packet recovery at first increases slowly and then rises steeply for some critical value of M/N_a. The same behavior is also observed in iterative BP erasure-decoding of rateless codes. The critical M/N_a actually defines the (expected) asymptotically optimal length of the contention period with respect to the throughput maximization, also observed in Fig. 3. Finally, the expected recovery delay for recovered packets increases linearly until the critical M/N_a. Although this behavior seems favorable, one should take into account that most of the packets are actually not recovered and thus do not contribute to the calculation of the delay. After critical M/N_a, most of the packets become recovered and the delay saturates.

The principle of adaptive termination favors the "fortunate" instances of packet-recovery process, ending the contention as soon as the terminating conditions are met [14]. The adaptive termination also implies that the packet-recovery process can tune to the actual wireless link conditions and potential imperfect SIC instances, simply disregarding the affected slots and proceeding with the contention process. In other words, frameless CSA is inherently adaptable to the scenarios when the assumptions outlined above may not hold. The main drawback is that the moment when the users receive feedback that terminates the contention is not known a-priori. In scenarios where the uplink and downlink transmissions share the same spectrum, in frameless CSA the BS has to contend with the active users when transmitting the feedback, as analyzed in [14]. We conclude by noting that similar arguments apply when comparing the advantages/drawbacks of the block and rateless coding frameworks.

Practical Issues

One of the underpinning assumptions of CSA is that each replica is equipped with pointers to the slots containing other replicas transmitted by the same user. However, in practice it is not trivial to make the pointers, nor is the cost of sending many pointers negligible. A more elegant approach to address this issue is to embed in each replica a user-specific seed of a pseudorandom generator known both to the user and the BS. Once a replica is resolved, the BS can use the knowledge of the generator and the obtained seed to determine all the slots containing the other replicas.

Another important practical issue is the estimation of the number of active users in a contention period N_a, which is usually a-priori not known and may vary over time, and which is required both in the framed and frameless variants of CSA in order to attain the optimal performance. Specifically, in framed CSA the knowledge of N_a should be used to dynamically adapt the duration of the contention period size M, in order to guarantee a constant logical load and thus a constant throughput. In frameless CSA, both the optimal slot-access probabilities and the termination criterion depend on N_a [14]. An efficient estimation algorithm specifically tailored for a frameless version of the scheme is proposed in [15].

Case Study: Upgrading the Existing Slotted ALOHA Implementations

Coded random access protocols can be very useful in the context of M2M communications, both in cellular and satellite access. Specifically, the access reservation procedure in all cellular stan-

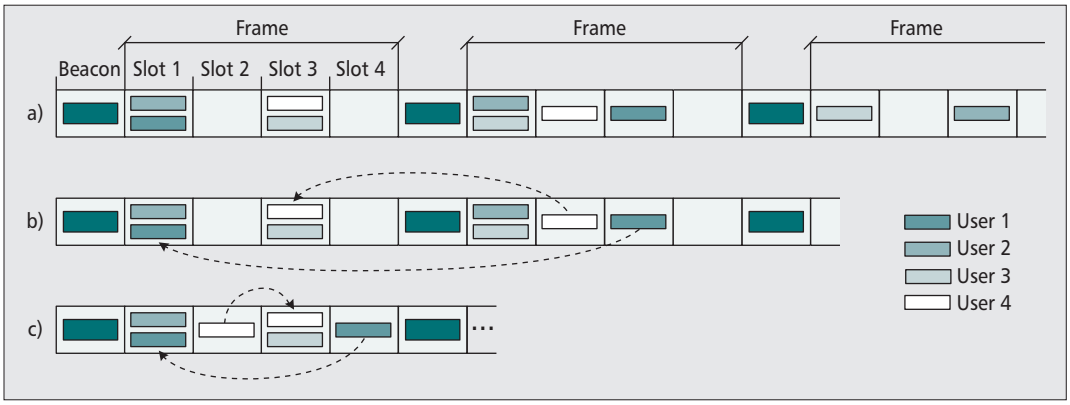


Figure 4. Example upgrade of framed slotted ALOHA.

dards, from GSM, over 3G, to LTE, is commonly based on framed slotted ALOHA, providing acceptable performance for human-oriented traffic. However, M2M traffic has fundamentally different requirements, primarily seen in the massive number of accessing terminals with short reporting deadlines, and the traditional ALOHA may create bottlenecks already in the access reservation.

We present a short study, describing how the contention phase of an existing cellular access reservation protocol can be upgraded to reap the advantages of coded random access while preserving the physical-layer behavior of the devices unchanged. The required modifications on the device side could be reduced to the implementation of the pseudorandom generators that will drive the selection of slots in which the access will be performed. This includes downlink signaling between the BS and the devices, in order to tune the pseudo-random generators, timers, back-off exponents, and other parameters of the actual FSA implementation, c.f. [16]. On the other hand, the BS stores the received uplink signals and uses SIC to process them, thereby absorbing the complexity of the upgrade, which is another highly desirable feature in practice.

Figure 4a presents an example of a generic framed slotted ALOHA. Active users transmit just once per frame, and only the transmissions occurring in singleton slots are successfully received and the corresponding devices are notified via the next beacon. The unsuccessful ones continue transmitting in the subsequent frames, choosing the slots where the repeated transmissions take place independently with respect to the choice made in the previous frames. In the example, the packets of user 1 and user 4 get through in the second frame, and the packets of user 2 and user 3 in the third frame.

In a simple upgrade, as shown in Fig. 4b, the active users also transmit once per frame, as in typical FSA. Nevertheless, the slot choice is dictated using the CSA approach, modified such that there can be only a single transmission within the subset of slots that belong to a frame. This effectively translates to a constraint imposed on the possible edge configurations in the bipartite graph. The slot choice is made locally at each user using a predefined function derived through the CSA graph-based design, whose inputs are the user ID and the information received from the beacons sent by the BS. Once a transmission is recovered, the BS retrieves the corresponding user ID, which enables the backtrack and cancellation of the replicas from the previous frames and potential resolution of other transmissions. In the example from Fig. 4b, the recovery of the packets of user 1 and user 4 in the second frame makes it possible to recover the packets of user 2 and user 3 from the first frame. For the sake of simplicity, we assumed that the choice of the slots is the same as in Fig. 4a.

Finally, the full upgrade that matches the standard CSA is presented in Fig. 4c. The users are allowed to repeat the same transmissions in multiple slots of the frame. The access strategies are again determined locally according to a predefined function, derived through the CSA approach and depending on the user ID and the information received from the BS. In this case, the BS removes the recovered packets both in the "forward" and "reverse" directions.

We conclude by noting that the application of the concepts described above could be made both in protocols that contend with data and protocols based on access reservation.

Conclusion

The legacy slotted ALOHA, although essentially inefficient, underpins the majority of the existing wireless random access protocols. The change of the perspective on the collision model through the application of successive interference cancellation has led to *coded random access*, an innovative approach superior to legacy SA. We have shown that coded random access is tightly related to codes on graphs and we have presented several protocol variants. Considering that the ALOHA approach dominated during the last four decades, we believe that the coded random access opens new grounds for designing communication systems that should embrace a massive number of M2M devices. Finally, we note that principles of the coded random access can be combined with any MUD technique, that is, they are not restricted to the simple chain of single-user detections' scenario assumed in the article.

The legacy slotted ALOHA, although essentially inefficient, underpins the majority of the existing wireless random access protocols. The change of the perspective on the collision model through the application of successive interference cancellation has led to coded random access, an innovative approach superior to legacy SA.

Acknowledgment

The work of E. Paolini was supported by the EU's Seventh Framework Programme (FP7/2007-2013) under grant agreement n. 288502.

Considering that the ALOHA approach dominated during the last four decades, we believe that the coded random access opens new grounds for designing communication systems that should embrace a massive number of M2M devices.

The work of Č. Stefanović was supported by the Danish Council for Independent Research, grant no. DFF-4005-00281. The work of P. Popovski was partially supported by the Danish Council for Independent Research, Sapere Aude Grant No. 11-105159.

References

[1] L. G. Roberts, "ALOHA Packet System with and Without Slots and Capture," *SIGCOMM Comput. Commun. Rev.*, vol. 5, no. 2, Apr. 1975, pp. 28–42.
[2] H. Okada, Y. Igarashi, and Y. Nakanishi, "Analysis and Application of Framed Aloha Channel in Satellite Packet Switching Networks — FADRA Method," *Electronics and Commun. in Japan*, vol. 60, Aug. 1977, pp. 60–72.
[3] A. Zanella and M. Zorzi, "Theoretical Analysis of the Capture Probability in Wireless Systems with Multiple Packet Reception Capabilities," *IEEE Trans. Commun.*, vol. 60, no. 4, Apr. 2012, pp. 1058–71.
[4] A. S. Tehrani, A. G. Dimakis, and M. J. Neely, "SigSag: Iterative Detection through Soft Message-Passing," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 8, Dec. 2011, pp. 1512–23.
[5] J. Goseling, M. Gastpar, and J. Weber, "Physical-Layer Network Coding on the Random Access Channel," *Proc. 2013 IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 2339–43.
[6] E. Casini, R. De Gaudenzi, and O. del Rio Herrero, "Contention Resolution Diversity Slotted ALOHA (CRDSA): An Enhanced Random Access Scheme for Satellite Access Packet Networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, Apr. 2007, pp. 1408–19.
[7] G. Liva, "Graph-Based Analysis and Optimization of Contention Resolution Diversity Slotted ALOHA," *IEEE Trans. Commun.*, vol. 59, no. 2, Feb. 2011, pp. 477–87.
[8] M. G. Luby, M. Mitzenmacher, and A. Shokrollahi, "Analysis of Random Processes via And-Or Tree Evaluation," *Proc. 9th ACM-SIAM SODA*, San Francisco, CA, USA, Jan. 1998.
[9] K. Narayanan and H. Pfister, "Iterative Collision Resolution for Slotted ALOHA: An Optimal Uncoordinated Transmission Policy," in *Proc. the 7th Int. Symp. Turbo Codes and Iterative Inf. Process.*, Gothenburg, Sweden, Aug. 2012, pp. 136–139.
[10] E. Paolini, G. Liva, and M. Chiani, "Graph-Based Random Access for the Collision Channel Without Feedback: Capacity Bound," *Proc. IEEE GLOBECOM 2011*, Houston, TX, USA, Dec. 2011.
[11] G. Liva, E. Paolini, M. Lentmaier, and M. Chiani, "Spatially-Coupled Random Access On Graphs," *Proc. IEEE ISIT 2012*, Boston, MA, USA, Jul. 2012.
[12] C. Stefanovic, P. Popovski, and D. Vukobratovic, "Frameless ALOHA Protocol For Wireless Networks," *IEEE Commun. Lett.*, vol. 16, no. 12, Dec. 2012, pp. 2087–90.
[13] J. Byers *et al.*, "A Digital Fountain Approach to Reliable Distribution of Bulk Data," *Proc. ACM SIGCOMM 1998*, Vancouver, BC, Canada, Sept. 1998.
[14] C. Stefanovic and P. Popovski, "ALOHA Random Access that Operates as a Rateless Code," *IEEE Trans. Commun.*, vol. 61, no. 11, Nov. 2013, pp. 4653–62.
[15] C. Stefanovic *et al.*, "Joint Estimation and Contention-Resolution Protocol for Wireless Random Access," *Proc. IEEE ICC 2013*, Budapest, Hungary, Jun. 2013.
[16] Medium Access Control (MAC) protocol specification, 3GPP TS 36.321.

Biographies

ENRICO PAOLINI (e.paolini@unibo.it) received the Dr. Ing. degree in telecommunications engineering in 2003 and the Ph.D. degree in electrical engineering in 2007, both from the University of Bologna, Italy. Currently he is a tenured assistant professor in the Department of Electrical, Electronic, and Information Engineering, University of Bologna. His research interests include error-control coding, random access techniques, and tracking algorithms. He is a frequent visitor at the German Aerospace Center, where he was a visiting scientist in 2012 and 2014. He is an editor for *IEEE Transactions on Communications* and *IEEE Communications Letters*. He has served as co-chair of the ICC'14 and ICC'15 Workshop on Massive Uncoordinated Access Protocols (MASSAP).

ČEDOMIR STEFANOVIĆ (cs@es.aau.dk) received Dipl.-Ing., Mr.-Ing. and Ph.D. degrees in electrical engineering from the University of Novi Sad, Novi Sad, Serbia. Since 2004 he has been affiliated with the Department of Power, Electronics and Communication Engineering, University of Novi Sad, where he holds the position of assistant professor. He is currently working as a postdoc researcher in the Department of Electronic Systems, Aalborg University, Denmark. His research interests are in the area of communication theory, including design and analysis of enhanced random access mechanisms, distributed algorithms for wireless ad-hoc and sensor networks and frame synchronization.

GIANLUIGI LIVA (Gianluigi.Liva@dlr.de) received M.S. and Ph.D. degrees in electrical engineering from the University of Bologna (Italy) in 2002 and 2006, respectively. His main research interests include satellite communication systems, random access techniques, and error control coding. He is currently leading the Information Transmission Group within the Institute of Communications and Navigation at the German Aerospace Center (DLR). He is/has been active in the DVB-SH, DVB-S2 and in the DVB-RCS standardization groups. In 2010 he was appointed guest lecturer for channel coding at the Institute for Communications Engineering (LNT) of the Technische Universitat Munchen (TUM).

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

Out-of-Band Emission Reduction and a Unified Framework for Precoded OFDM

Xiaojing Huang, Jian A. Zhang, and Y. Jay Guo

ABSTRACT

OFDM has been regarded as a promising candidate for use in cognitive radio systems with dynamic spectrum reuse capability. However, conventional OFDM has significant OOBE, which can cause severe interference to systems operating in adjacent frequency bands. In addition to conventional techniques such as spectral shaping filtering, guard band insertion, and time domain windowing, new OOBE reduction techniques, including cancellation carrier and spectral precoding, have been proposed in recent years. This article reviews various OOBE reduction techniques and proposes a generalized low-complexity OOBE reduction framework for discrete Fourier transform precoded OFDM. With the allocation of explicit frequency domain cancellation subcarriers and data domain cancellation symbols, the proposed framework enables various configurations to achieve significant OOBE reduction with low implementation complexity, and provides flexibility in balancing OOBE reduction and other performance metrics such as peak-to-average power ratio.

INTRODUCTION

Orthogonal frequency-division multiplexing (OFDM) has been widely used in various practical communication systems, such as wireless local area networks (WLANs) [1] and Long Term Evolution (LTE) mobile systems [2], and is also very promising for future multiband cognitive radio systems. However, conventional OFDM has significant out-of-band emission (OOBE). In a typical OFDM-based multiband cognitive radio system, the overall spectrum is divided into multiple subbands, and shared between licensed users (primary users) and unlicensed users (secondary users). Secondary systems need to ensure that their transmitted signals have very sharp spectrum rolloff to maximize the usable bandwidth and minimize interference to primary systems. How to effectively reduce OOBE for OFDM signals remains a major technical challenge.

Traditional existing OOBE reduction techniques include spectral shaping filtering, guard band insertion, and time domain windowing. They are less flexible for adaptive subband allocation or less spectrally efficient, and therefore not suitable for the cognitive radio applications mentioned above. Recently, more advanced OOBE reduction techniques using reserved in-band subcarriers and orthogonal spectral precoding have been proposed. These techniques provide more effective and dynamic OOBE reduction at the cost of higher computational complexity. For example, OOBE reduction by using cancellation carriers is formulated as a constrained optimization problem [3]. Spectral precoding methods [4] generally require matrix multiplications at both the transmitter and receiver sides. Hence, their complexity is proportional to M^2 in terms of the number of complex multiplications, where M is the number of total in-band subcarriers. A low-complexity sidelobe suppression with orthogonal projection (SSOP) scheme was proposed in [5] with complexity only proportional to M at both transmitter and receiver sides. However, its bit error rate (BER) performance is slightly degraded due to noise enhancement after data symbol recovery at the receiver.

OFDM with spectral precoding can be regarded as a form of precoded OFDM, that is, preprocessing the data symbols before inverse fast Fourier transform (IFFT), which converts data symbols from the data domain into the frequency domain. Mathematically, precoding is a matrix multiplication operation with the data symbol vector. In addition to OOBE reduction, precoding can also be used to improve other OFDM signal properties such as reducing the peak-to-average-power ratio (PAPR). For example, single-carrier frequency-division multiple access (SC-FDMA) [6] is one such special type of precoded OFDM, which has been used in LTE uplink with improved power efficiency and diversity performance. In SC-FDMA, the precoding matrix is a discrete Fourier transform (DFT) matrix; hence, precoding can be efficiently implemented via fast Fourier transform (FFT).

In this article, we review existing OOBE reduction techniques and propose a low-complexity OOBE reduction framework for DFT-based precoded OFDM systems. It generalizes a few well-known spectral precoding schemes and can be flexibly configured to generate new schemes according to specific requirements on the complexity and performance for OOBE

Xiajing Huang and Y. Jay Gio are with the University of Technology, Sydney.

Jian A. Zhang is with CSIRO Digital Productivity Flagship.

0163-6804/15/$25.00

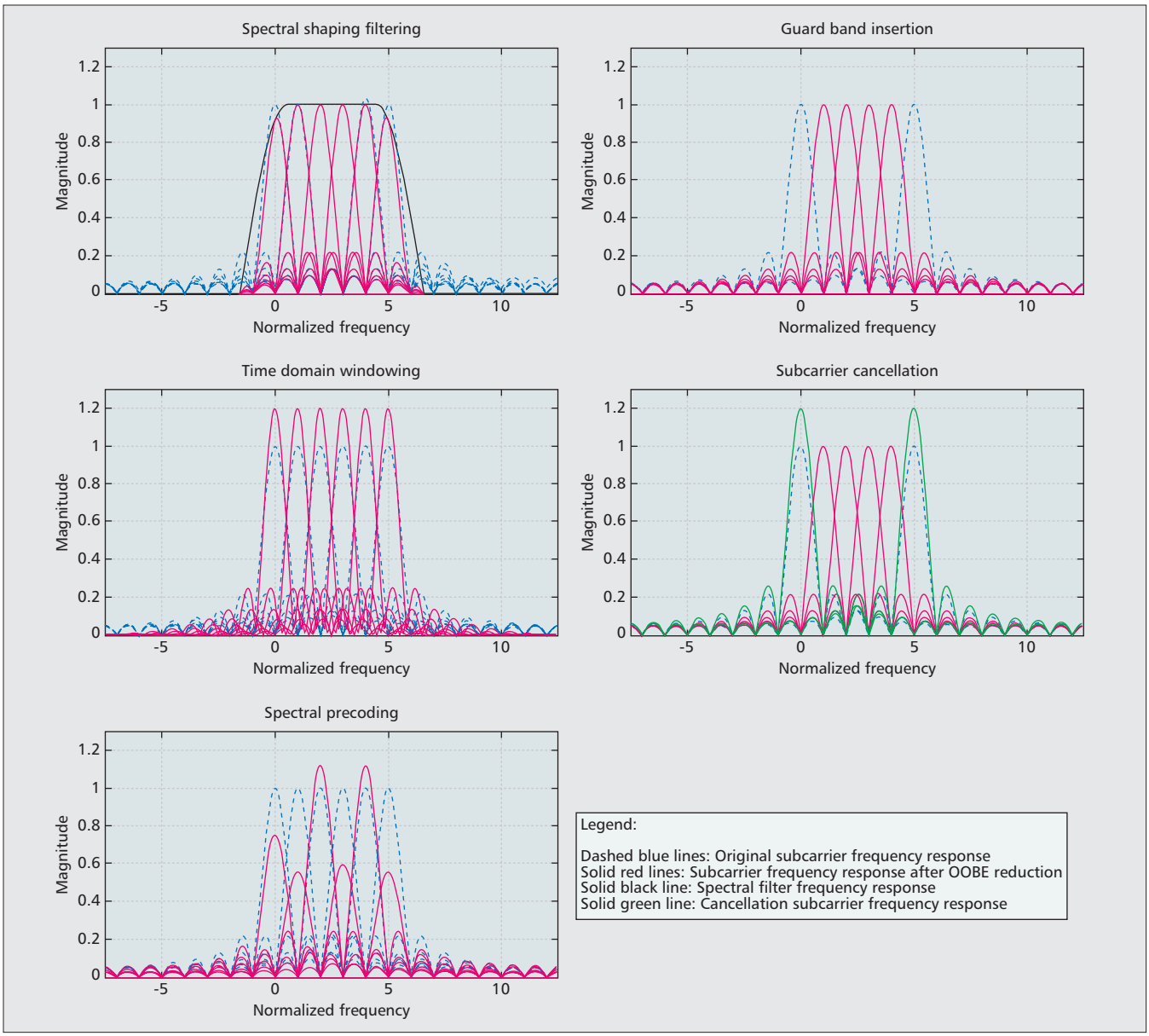


Figure 1. Illustration of different OOBE reduction techniques for an OFDM system.

reduction. OOBE reduction techniques, particularly precoding approaches, can generally increase PAPR due to signal superposition. With explicit frequency domain cancellation subcarriers (FDCSs) and data domain cancellation symbols (DDCSs), the proposed framework achieves significant OOBE reduction without increasing signal PAPR or degrading BER performance. The additional complexity required for OOBE reduction, which is proportional to the total number of in-band subcarriers, is only incurred at the transmitter without extra processing for subcarrier recovery at the receiver.

The rest of this article is organized as follows. First, we review existing OOBE reduction techniques for OFDM systems, and provide in-depth comparisons between them. We then describe the DFT precoded OFDM signal models, and illustrate how FDCSs and DDCSs are allocated. We further propose the generalized low-complexity OOBE reduction framework, and provide a special design by which explicit FDCSs and/or DDCSs are used for OOBE reduction. Finally, we show simulation results on the performance of the framework with various FDCS and DDCS configurations, and demonstrate that jointly using FDCSs and DDCSs not only achieves sufficient OOBE reduction, but also reduces PAPR of the transmitted OFDM signal.

Overview of Existing OOBE Reduction Techniques

Some typical OOBE reduction techniques and their effects are illustrated in Fig. 1. For convenience and clarity, we assume that there are six in-band subcarriers located at normalized frequencies (normalized to subcarrier frequency spacing) 0 to 5 in the studied OFDM systems. The average power spectral density is normalized to 1. The magnitude frequency response of

each subcarrier is plotted individually against the normalized frequency. A summary of the principles and the advantages and disadvantages of these techniques is given below.

Spectral Shaping Filtering

This is the most straightforward technique to suppress OOBE in the specified frequency band and is used in almost all narrowband systems in order to meet various transmit mask requirements [1, 2, 7]. The spectral shaping filter can be applied to the transmitted OFDM signal either before or after digital-to-analog conversion. To achieve high spectral efficiency, this spectral shaping filter should have a very sharp transition between pass and stop bands. However, digital implementations of such filtering have high processing complexity, and analog implementations can be less cost efficient or flexible for dynamic subband allocation. As illustrated in the spectral shaping filtering subplot of Fig. 1, where a root raised cosine filter with rolloff factor 0.35 is used as the spectral shaping filter, any imperfection in the filter frequency response such as passband droop and ripple will affect the original subcarriers in the frequency domain. In addition, the PAPR of the time domain OFDM signal will increase in general after spectral shaping.

Guard Band Insertion

This is commonly used in practical OFDM systems, such as IEEE 802.11a/g/n WLANs [1], digital video broadcasting (DVB) systems, and LTE mobile systems [2]. By inserting guard bands (i.e., null subcarriers) on the two edges of the subband, OOBE can be reduced in neighboring subbands to be protected thanks to the decay of the signal spectral sidelobe on the order of $1/f^2$ (assuming that the time domain window for an OFDM symbol is a rectangular function, and thus the frequency domain representation of a subcarrier is a sinc function), where f is the distance between the observing frequency and the mainlobe center frequency of the interfering subcarrier. Guard band insertion can be used alone or in conjunction with other OOBE reduction methods such as the preceding spectral shaping technique. As shown in the guard band insertion subplot of Fig. 1, insertion of guard bands will affect neither signals at other subcarriers in the frequency domain nor the PAPR of the time domain OFDM signal. Unfortunately, it reduces spectral efficiency and may not be able to provide sufficient OOBE reduction without using a large number of null subcarriers.

Time Domain Windowing

This technique applies windowing to the transmitted OFDM signal in the time domain, and has been widely used in modern broadband wireless systems such as WLANs [1] together with guard band insertion. By reducing the amplitude of the two edges of an OFDM symbol gradually toward zero and making the signal waveform smoothly transition between two OFDM symbols, the spectral sidelobes can be significantly suppressed. However, this leads to an extended OFDM symbol period with reduced spectrum efficiency and increased signal power. This can be seen from the time domain windowing subplot of Fig. 1, where the window function is selected as the convolution of a rectangular pulse with a half-sine pulse [8]. The magnitude of each subcarrier is increased after time domain windowing, and the orthogonality among original subcarriers is also broken. Hence, the time domain guard interval needs to be used together with windowing to avoid signal distortion and intersymbol interference. As a result, the spectrum efficiency is reduced. In addition, it is hard for the windowing method to achieve sufficient inter-subband emission reduction in cognitive radios where multicarrier modulation over non-contiguous subbands is frequently confronted [5].

Subcarrier Cancellation

This technique uses reserved cancellation subcarriers on the edges of a subband to cancel the OOBE [3]. Its application in existing systems has not been reported, but it is very promising for future cognitive systems. To achieve sufficient OOBE reduction, significant power at the cancellation carriers is required, resulting in reduced power efficiency. An extended active interference cancellation method is proposed in [9], by which the cancellation carriers are inserted directly in the frequency band where OOBE needs to be suppressed. However, this method may cause considerable self-interference to the in-band data subcarriers due to the non-orthogonality of the inserted cancellation carriers. To limit the total power of the cancellation subcarriers inserted in-band or the level of in-band self-interference caused by out-of-band cancellation carriers, subcarrier cancellation is generally formulated as a constrained optimization problem. As shown in the subcarrier cancellation subplot of Fig. 1, inserting cancellation carriers does not affect other subcarriers but increases the PAPR of the time domain OFDM signal in general.

Spectral Precoding

This technique uses a data-independent matrix to precode the data symbols to reduce OOBE [4, 10, 11]. In terms of mathematical operation, an OFDM system with spectral precoding can be regarded as a general precoded OFDM system. Because of its superior OOBE reduction performance and linear operation, spectral precoding has attracted considerable attention in recent years. The precoding matrix is the main differentiator between different spectral precoding methods. Optimal spectral precoding using orthogonal precoding matrices for both single-user and multiuser scenarios has been proposed [4, 11]. It has been shown that the achievable OOBE reduction is determined by the precoding redundancy (the difference between the number of total in-band subcarriers and the number of data subcarriers [4]). In general, the larger the precoding redundancy, the more OOBE reduction the spectral prcoding can achieve, at the cost of reduced spectral efficiency.

The design of the spectral precoding matrix typically relies on the expression of the frequency domain OOBE. An exception is N-continuous OFDM [12], where the precoding matrix is designed such that any two consecutive OFDM symbols are made continuous in their first N derivatives of the time domain signals. The pre-

In a precoded OFDM system, precoding converts the original data symbols from the data domain into precoded data symbols that are allocated to subcarriers in the frequency domain. The time domain OFDM signal is finally obtained after performing IFFT on the frequency domain signals.

DDCSs and FDCSs can also be allocated at a different symbol position and on a different subcarrier, respectively. However, the DDCS at the first symbol position of a precoding block and the FDCS on a subband edge can achieve more effective OOBE reduction.

coding and decoding are performed recursively at the transmitter and receiver, respectively, which not only increases complexity but also degrades BER performance. One major disadvantage of orthogonal spectral precoding is the high implementation complexity since both the transmitter and receiver require matrix multiplication operations. Considerable complexity reduction becomes possible when a non-orthogonal matrix is used. For example, the sidelobe suppression with orthogonal projection method [5] can be reformulated under the spectral precoding framework and achieves significant complexity reduction, at the cost of PAPR increase and/or BER degradation. Due to the precoding operation, signals at precoded subcarriers are different linear combinations of the original data symbols. In Fig. 1, the spectral precoding subplot shows one realization of the precoded subcarriers. The power spectral density of the precoded OFDM signal is the ensemble average of all the precoded subcarriers. Spectral precoding with SSOP has been used in the Ngara backhaul, which is a multiband multi-gigabit microwave point-to-point system [13].

Signal Models and DDCS/FDCS Allocation

We now describe the precoded OFDM signal models for a generalized DFT-based precoding framework, and describe how cancellation symbols and subcarriers can be allocated in the data domain and frequency domain, respectively, to reduce OOBE.

In a precoded OFDM system, precoding converts the original data symbols from the data domain into precoded data symbols that are allocated to subcarriers in the frequency domain. The time domain OFDM signal is finally obtained after performing IFFT on the frequency domain signals. The precoding matrix can be a single matrix, or consist of several sub-matrices. Each matrix can be orthogonal or non-orthogonal, but should be invertible such that the original data symbols can be recovered at the receiver. Without loss of generality, we consider the case where each sub-matrix is a DFT matrix. The proposed framework can be straightforwardly extended to other cases.

To achieve effective OOBE reduction, DDCSs and FDCSs are allocated before and after DFT precoding, respectively, as shown in Fig. 2a. A DDCS is a reserved data domain cancellation symbol, whereas an FDCS is a reserved frequency domain cancellation subcarrier. No data symbol is mapped on either a DDCS or an FDCS. OOBE reduction is then performed, followed by subcarrier mapping and IFFT. The information data bits are first encoded, interleaved, and modulated to produce data symbols which are then divided into one or more data symbol groups. A DDCS is inserted in each data symbol group to form a precoding block, and a precoding sub-matrix is applied to the block to generate a precoded block. Multiple precoded blocks are allocated to subcarriers in the transmission frequency band, called precoded subcarriers. FDCSs are then inserted among or on the edges of precoded subcarriers. The total number of DDCSs and FDCSs is the precoding redundancy, which represents the bandwidth sacrificed for the purpose of OOBE reduction. The locations of DDCS and FDCS are predefined, but the signals carried on them are generated dynamically, as discussed in the next section. After performing OOBE reduction, an IFFT is finally applied to all frequency domain subcarriers to generate an OFDM symbol, and a cyclic prefix (CP) or zero-padded (ZP) suffix is then appended.

The receiver structure is the same as that of a conventional precoded OFDM system, with the possible addition of a subcarrier recovery module for removing interference on subcarriers introduced by the OOBE reduction at the transmitter [5]. Whether or not subcarrier recovery is required depends on the specific OOBE reduction method, as discussed in next section.

An example of allocating four DDCSs and four FDCSs in one of the four subbands is illustrated in Fig. 2b. In the data domain, data symbols are grouped into four precoding blocks (only one precoding block is shown in the shaded area). In each precoding block, one DDCS is added and placed at the first symbol position. In the frequency domain, four FDCSs are inserted, two on each of the two ends of the subband. The DDCS and FDCSs can also be allocated at a different symbol position and on a different subcarrier, respectively. However, the DDCS at the first symbol position of a precoding block and the FDCS on a subband edge can achieve more effective OOBE reduction.

Note that pilots can be further inserted in the data domain and/or between precoded subcarrier blocks in the frequency domain for the general purpose of channel estimation and tracking, as well as compensating for impairments such as carrier frequency offset, sampling frequency offset, and phase noise. They are omitted here in order to focus on OOBE reduction.

Unified Framework for OOBE Reduction

The proposed framework only requires linear operation on the frequency domain signals to produce OOBE cancellation signals. Consider a subband with M subcarriers including FDCSs and precoded subcarriers. This processing can be simply expressed as $\mathbf{Y} = \mathbf{X} - \mathbf{DCX}$, where $\mathbf{X}$ and $\mathbf{Y}$ are $M \times 1$ column vectors representing the frequency domain signals before and after OOBE reduction, respectively, $\mathbf{C}$ is the *cancellation matrix* of dimension $L \times M$ ($L < M$), and $\mathbf{D}$ is the *distribution matrix* of dimension $M \times L$. Note that $\mathbf{X}$ is composed of precoded subcarriers after DFT precoding, and $\mathbf{Y}$ is a linearly distorted version of $\mathbf{X}$ after OOBE reduction operation. Recovery of $\mathbf{X}$ from $\mathbf{Y}$ at the receiver may be necessary depending on the selection of the $\mathbf{D}$ matrix. The cancellation matrix $\mathbf{C}$ is first applied to $\mathbf{X}$ to generate L *cancellation signals*, represented as an $L \times 1$ column vector $\mathbf{CX}$, whereas the distribution matrix $\mathbf{D}$ further weights and distributes the cancellation signals onto subcarriers in the transmission subband, so that $\mathbf{DCX}$ are

the frequency domain signals to be subtracted from the original signals **X** for OOBE reduction.

FDCSs and DDCSs are reflected in spectral procoding **X** – **DCX** through the selection of the **D** matrix. For example, if the element in the row of **D** corresponding to an FDCS is selected as one, and other elements in the columns of **D** are all zeros, the cancellation signals are only distributed to the FDCS. At the receiver, a signal mapped on the FDCS will be discarded. If the elements in the columns of **D** corresponding to the precoded subcarriers, which are the outputs of DFT to a precoding block, are all ones, the cancellation signals are only distributed to the DDCS. After IDFT at the receiver, a signal mapped on the DDCS will be discarded. If **D** is selected in such a way that any cancellation signal and any information symbol are mapped to the same useful data subcarrier, subcarrier recovery will be necessary, such as for the SSOP case.

The OOBE at Q specified *cancellation points* outside the transmission subband can be expressed as **AY**, where **A** is the *attenuation matrix* of dimension $Q \times M$. An element of **A** is the attenuation factor, which characterizes the strength of OOBE and can be represented as the inverse of the distance (normalized by subcarrier

For each subband, two subcarriers on each side of the subband are reserved as FDCSs. The remaining 64 subcarriers are divided into 2 or 4 blocks of precoded subcarriers, each having 32 or 16 subcarriers. For each precoding block of 32 or 16 data domain symbols, the first symbol is reserved as DDCS.

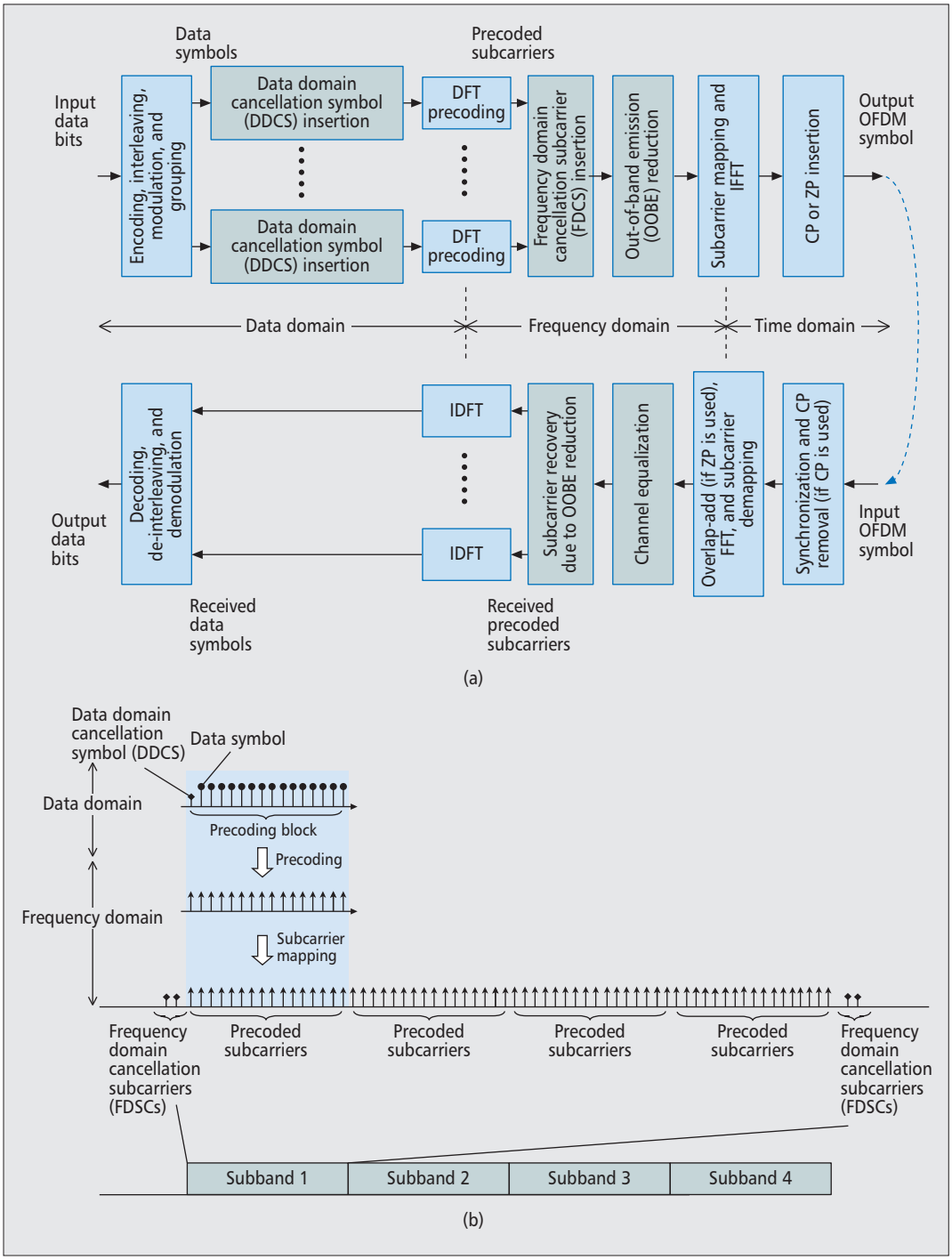


Figure 2. a) Transmitter and receiver signal models; b) cancellation symbol/subcarrier allocation.

Compared to FDCS only and DDCS only methods, jointly using FDCSs and DDCSs can not only achieve sufficient OOBE reduction, but also lead to lower PAPR. By adjusting the cancellation distance, a trade-off between the spectral rolloff sharpness and OOBE reduction level can easily be achieved to suit different application requirements.

spacing) between a cancellation point and an in-band subcarrier [5].

Given a distribution matrix **D**, the cancellation matrix **C** can be designed under various optimization criteria. The most commonly used criterion is to minimize the average power of **AY** (i.e., the OOBE power at the Q out-of-band cancellation points), and the solution is found to be $\mathbf{C} = (\mathbf{AD})^{+}\mathbf{A}$ where $(\cdot)^{+}$ denotes pseudo-inverse. Depending on different values of Q and L, $(\mathbf{AD})^{+}$ can be calculated as $(\mathbf{AD})^{-1}$, $(\mathbf{AD})^{T}(\mathbf{AD}(\mathbf{AD})^{T})^{-1}$, and $((\mathbf{AD})^{T}\mathbf{AD})^{-1}(\mathbf{AD})^{T}$ for $Q = L$, $Q < L$, and $Q > L$ respectively, where $(\cdot)^{T}$ denotes transposition.

The parameter L determines the precoding redundancy, which implies that L/M of the bandwidth is lost. The parameter Q specifies how many out-of-band subcarriers are selected for OOBE minimization in order to design the **C** matrix. In general, if $Q \le L$, the specified Q out-of-band subcarriers can be completely nulled (however, this does not mean that the total OOBE is zero since OOBE appears over all out-of-band frequencies). The **A** matrix determines the amount of OOBE at the Q cancellation points generated by all in-band subcarriers. The **D** matrix determines how the cancellation signals are distributed onto in-band subcarriers.

The above framework generalizes some existing OOBE reduction techniques, such as the unconstrained cancellation carrier method [3], self-cancellation for SC-FDMA [6], and SSOP [5]. These techniques can be reformulated under this framework by selecting an appropriate **D** matrix. For example, when **D** is selected as

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}^{T}$$

which implies $L = 2$, we obtain the cancellation carrier method (without power constraint) with one FDCS located at each side of the subband; when **D** is selected as $(1\ 1\ \cdots\ 1)^{T}$, which implies $L = 1$, we obtain the self-cancellation method for SC-FDMA, which equivalently uses one DDCS; and when **D** is selected as $\mathbf{A}^{T}$, which implies $L = Q$, we obtain the SSOP method. Note that for the cancellation carrier method and the self-cancellation method for SC-FDMA, L explicit FDCSs and DDCSs are used, respectively, whereas the number of cancellation points, Q, can be selected independently. Also, no subcarrier recovery module is necessary at the receiver since the FDCSs and DDCSs can be simply discarded in the frequency domain and data domain, respectively. However, for the SSOP method, the **D** matrix distributes the L cancellation signals onto all in-band subcarriers, and thus subcarrier recovery is necessary at the receiver. Detailed recoverability analysis can be found in [5].

OOBE Reduction with Explicit FDCSs and DDCSs

As discussed in the last section, using different distribution matrices in the proposed framework can lead to different OOBE reduction schemes with different performance. With explicit FDCS and DDCS allocation in the DFT precoded OFDM system, we are able to design a distribution matrix **D** that ensures low-complexity OOBE reduction, eliminates the need for subcarrier recovery at the receiver, and preserves superior PAPR performance.

The design rules for such a distribution matrix **D** are summarized as follows:

- If an FDCS is allocated at a frequency domain subcarrier, a cancellation signal can be directly distributed to it by setting the element in **D** at the row corresponding to the FDCS and the column corresponding to the cancellation signal to 1, and the remaining elements in the column to 0s. For example, for an FDCS located at the beginning of the subcarriers, the corresponding column in **D** is

$$\left(\underbrace{1\ \ 0\ \ \ldots\ \ 0}_{M \text{ elements}} \right)^{T}.$$

This ensures that a cancellation signal for an FDCS will not be distributed onto other subcarriers.

- If a DDCS is allocated at the first position of a data domain precoding block of size K, a cancellation signal can be distributed to a block of K precoded subcarriers (which are mapped from the size-K DFT output of the precoding block) by setting the K elements in the column of **D** corresponding to the cancellation signal at rows corresponding to the K precoded subcarriers to 1s. For example, for a DDCS in the precoding block that produces K precoded subcarriers starting from the third subcarriers in the subband, the corresponding column in **D** is

$$\left(\underbrace{0\ \ 0\ \ \underbrace{1\ \ \ldots\ \ 1}_{K \text{ ones}}\ \ 0\ \ \ldots\ \ 0}_{M \text{ elements}} \right)^{T}.$$

This arrangement will indirectly distribute the cancellation signal only to the DDCS in the data domain and hence will not cause interference to other data symbols.

- One cancellation signal will be distributed to one, and only one, FDCS or DDCS. This ensures that **D** will always be a full column-rank matrix.

From the above distribution matrix construction process, we see that L, the number of columns of **D**, equals the total number of FDCSs and DDCSs (i.e., the precoding redundancy).

Compared to existing OOBE reduction techniques, the proposed framework using explicit FDCSs and DDCSs has the following advantages. First, it has very low complexity since the number of multiplication operations is only LM for calculating **CX** at transmitter where **C** is a real-valued matrix. The data symbols are received immediately after IDFT without subcarrier recovery at the receiver. Second, despite the slight increase of transmitted signal power due to the insertion of FDCSs and DDCSs, the data symbols are not distorted by any FDCS or DDCS, so there will be no BER performance degradation under the same effective signal power carried on data symbols. Third, as demon-

strated in the next section, jointly using FDCSs and DDCSs for OOBE reduction also produces lower PAPR for the transmitted OFDM signal, whereas using only FDCSs or DDCSs for OOBE reduction may significantly increase PAPR.

Performance Evaluation

To evaluate the performance of the proposed low-complexity OOBE reduction framework, we now consider a multiband DFT precoded OFDM system with four subbands, each having 68 subcarriers. We assume that a secondary user wants to use the first, third, and fourth subbands that the primary user is not using, and our proposed OOBE reduction framework with explicit FDCSs and/or DDCSs is used to mitigate interference to the primary user.

For each subband, two subcarriers on each side of the subband are reserved as FDCSs. The remaining 64 subcarriers are divided into 2 or 4 blocks of precoded subcarriers, each having 32 or 16 subcarriers. For each precoding block of 32 or 16 data domain symbols, the first symbol is reserved as DDCS. The FFT/IFFT size is 512, and every OFDM symbol has a zero-padded suffix. Note that the proposed framework can also be applied to a CP-OFDM system by using a corresponding attenuation matrix. However, the OOBE reduction is not as significant as that for a ZP-OFDM system. Similar observations have been reported in the literature for other OOBE reduction techniques [3–5].

First, we compare the OOBE reduction performance using different combinations of FDCSs and DDCSs. The PAPRs of the resulting transmitted OFDM signals are also compared. Four different cases are considered:

1. Only the four FDCSs are used for OOBE reduction, and the remaining 64 subcarriers are used for transmitting data symbols with precoding redundancy $L = 4$.
2. Only the four DDCSs are used for OOBE reduction, and the remaining 60 subcarriers are used for transmitting data symbols with precoding redundancy $L = 4$ (the four FDCSs are nulled, and the precoding block size is 16).
3. Four FDCSs and four DDCSs are used for OOBE reduction, and the remaining 60 subcarriers are used for transmitting data symbols with precoding redundancy $L = 8$ (the precoding block size is also 16).
4. Two FDCSs and two DDCSs are used for OOBE reduction, so the total number of FDCSs and DDCSs is the same as that in case 1 or 2, that is, $L = 4$ (the remaining 62 subcarriers are used for transmitting data symbols and the precoding block size is 32).

As for the selection of cancellation points outside the transmission subband, we only consider the condition $Q > L$ for convenience and select the cancellation points as all out-of-band frequency subcarriers away from the two sides of the transmission subband with the same offset d, called *cancellation distance*.

Figure 3 shows the secondary user's signal spectra in subbands 1, 3, and 4 for the four cases with cancellation distance set to 1. Spectrum for conventional OFDM without OOBE reduction is also plotted for comparison. As we can see, the OOBE reduction using FDCSs only (case 1) reduces the OOBE to about –40 dB. However, using DDCSs only (case 2) and FDCS+DDCS with the same L value as in case 1 or 2 (case 4) can reduce the OOBE to about –60 dB. With L increasing to 8, OOBE can be further reduced to about –80 dB. From the inset plot of Fig. 3, we can also see that case 1 requires significantly higher power on FDCSs, which is undesirable as it will increase the total transmission power. Some power constraints have to be imposed on the FDCS-only method; however, it will reduce the OOBE reduction performance.

Figure 4 shows the complementary cumulative density function (CCDF) of PAPR, that is,

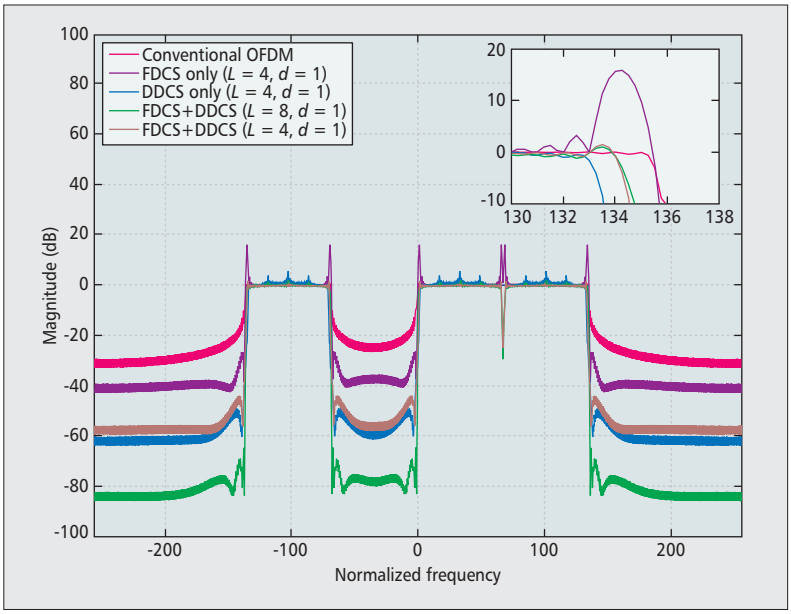


Figure 3. Power spectral densities of a transmitted secondary user's OFDM signal with various combinations of FDCSs and DDCSs.

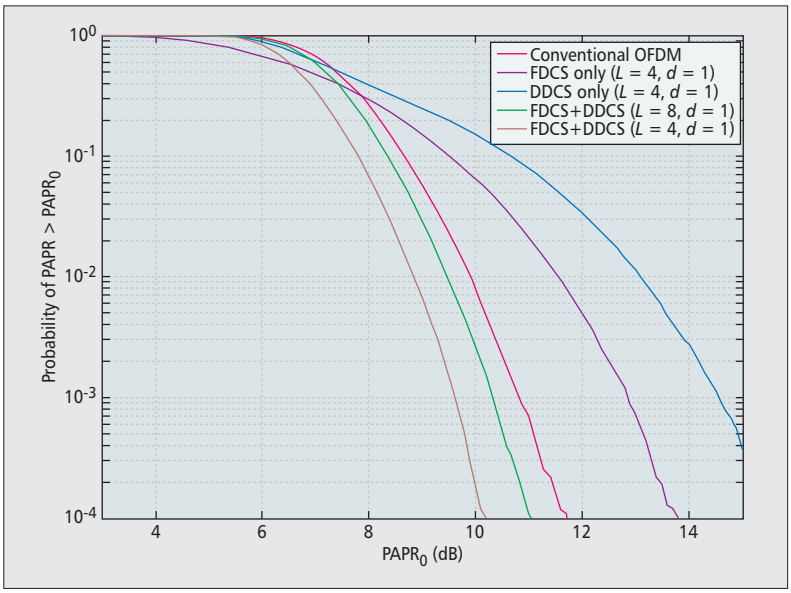


Figure 4. CCDF of PAPR of a transmitted secondary user's OFDM signal with various combinations of FDCSs and DDCSs.

the probability of PAPR greater than a specified threshold $PAPR_0$, of the transmitted signal for the four cases. We see that both cases 1 and 2 result in much higher PAPR than conventional OFDM and hence require much larger power backoff. The FDCS+DDCS method not only reduces OOBE but also produces smaller PAPR than conventional OFDM, although the PAPR improvement decreases with L increasing. The reason for the reduced PAPR is that when FDCS and DDCS are jointly used, signal power is more evenly distributed across the subband, while necessary correlations among all subcarriers are introduced for OOBE reduction.

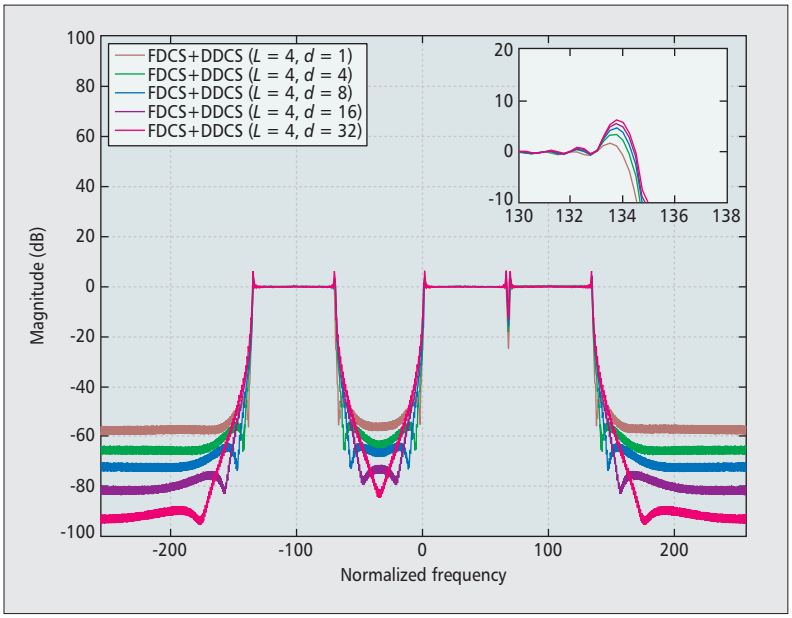


Figure 5. Power spectral densities of transmitted secondary user's OFDM signal with the same numbers of FDCSs and DDCSs but different cancellation distances.

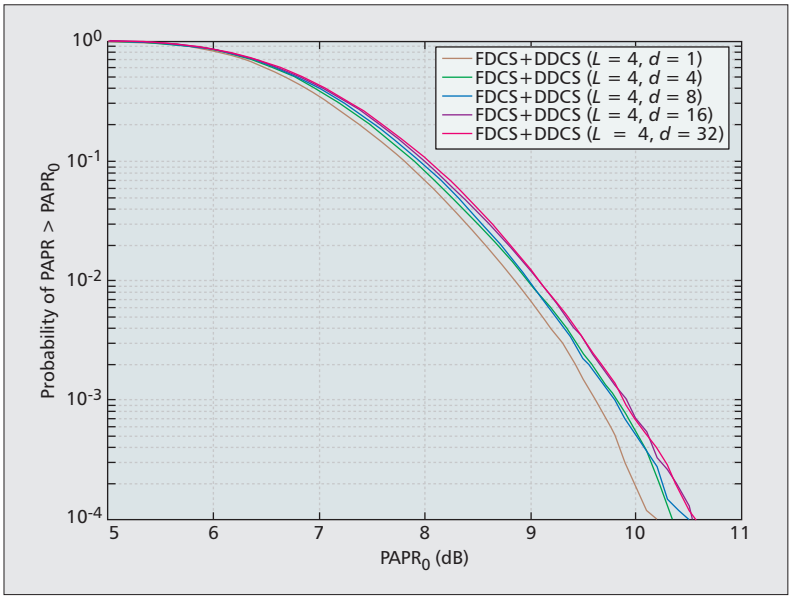


Figure 6. CCDF of PAPR of transmitted secondary user's OFDM signal with the same numbers of FDCSs and DDCSs but different cancellation distances.

The cancellation distance plays an important role in OOBE reduction, too. In general, if d is smaller, that is, the cancellation points are closer to the two edges of the subband, the reduced OOBE will roll off more quickly. If d is larger, the reduced OOBE will roll off more slowly, but more OOBE reduction can be achieved. This can clearly be observed from Fig. 5, where the spectra using the FDCS+DDCS method with the same L value but various d values are compared. As the cancellation distance increases from 1 to 32, the OOBE is approximately reduced from –60 dB to –90 dB, but the transition band becomes wider. Also note that when d is larger, more power on FDCSs is required, but the power increase is not as significant as that of the FDCS-only method, as can be seen from the inset plot.

Finally, Fig. 6 plots the PAPR performance for the FDCS+DDCS method with $L = 4$. It shows that the PAPR increases slightly as d increases, but is still lower than those shown in Fig. 4 for the FDCS+DDCS method with $L = 8$ and conventional OFDM.

Conclusions

Spectral precoding is an effective technique for reducing OOBE for OFDM cognitive radios. We have proposed a generalized framework using both frequency domain cancellation subcarriers and data domain cancellation symbols to reduce OOBE for DFT precoded OFDM systems with low implementation complexity. The framework enables flexible configurations to achieve various trade-offs between OOBE reduction effect and other performance metrics such as PAPR of the transmitted signal. Compared to FDCS-only and DDCS-only methods, jointly using FDCS and DDCS can not only achieve sufficient OOBE reduction but also lead to lower PAPR. By adjusting the cancellation distance, trade-off between the spectral rolloff sharpness and OOBE reduction level can easily be achieved to suit different application requirements.

References

[1] IEEE Std 802.11a/g/n, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," 1999, 2003, 2009.

[2] S. Parkvall *et al.*, "Evolution of LTE toward IMT-Advanced," *IEEE Commun. Mag.*, vol. 49, no. 2, Feb. 2011, pp. 84–91.

[3] S. Brandes, I. Cosovic, and M. Schnell, "Reduction of Out-of-Band Radiation in OFDM Systems by Insertion of Cancellation Carriers," *IEEE Commun. Lett.*, vol. 10, no. 6, June 2006, pp. 420–22.

[4] M. Ma *et al.*, "Optimal Orthogonal Precoding for Spectral Leakage Suppression in DFT-Based Systems," *IEEE Trans. Commun.*, vol. 59, no. 3, Mar. 2011, pp. 844–53.

[5] J. Zhang *et al.*, "Sidelobe Suppression with Orthogonal Projection for Multicarrier Systems," *IEEE Trans. Commun.*, vol. 60, no. 2, Feb. 2012, pp. 589–99.

[6] M. Ma, X. Huang, and Y. Jay Guo, "An Interference Self-Cancellation Technique for SC-FDMA Systems," *IEEE Commun. Lett.*, vol. 14, no. 6, pp. 512–14, June 2010.

[7] H. Bolcskei, P. Duhamel, and R. Hleiss, "Design of Pulse Shaping OFDM/OQAM Systems for High Data-Rate Transmission over Wireless Channels," *Proc. 1999 IEEE ICC*, vol. 1, pp. 559–64.

[8] B. Farhang-Boroujeny, "OFDM versus Filter Bank Multicarrier," *IEEE Signal Proc. Mag.*, vol. 28, no. 3, May 2011, pp. 92–112.

[9] D. Qu and Z. Wang, "Extended Active Interference Cancellation for Sidelobe Suppression in Cognitive Radio OFDM Systems with Cyclic Prefix," *IEEE Trans. Vehic. Tech*., vol. 59, 4, May 2010, pp. 1689–95.
[10] C.-D. Chung, "Spectral Precoding for Rectangularly Pulsed OFDM," *IEEE Trans. Commun*., vol. 56, no. 9, Sept. 2008, pp. 1498–1510.
[11] X. Zhou, G. Y. Li, and G. Sun, "Multiuser Spectral Precoding for OFDM-Based Cognitive Radio Systems," *IEEE JSAC*, vol. 31, no. 3, Mar. 2013, pp. 345–52.
[12] J. van de Beek and F. Berggren, "N-Continuous OFDM," *IEEE Commun. Lett*., vol. 13, no. 1, Jan. 2009, pp. 1–3.
[13] X. Huang *et al.*, "A Multi-Gigabit Microwave Backhaul," *IEEE Commun. Mag*., vol. 50, no. 3, Mar. 2012, pp. 122–29.

Biographies

Xiaojing Huang [SM'11] (Xiaojing.Huang@uts.edu.au) is a professor of information and communications technology in the School of Computing and Communications, University of Technology, Sydney (UTS), Australia. Before joining UTS, he was a principal research scientist at Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia, and the project leader of CSIRO microwave and mmWave backhaul projects since 2009, an associate professor at the University of Wollongong, Australia, since 2004, and a principal research engineer at Motorola Australian Research Centre since 1998.

Jian (Andrew) Zhang [SM] (Andrew.Zhang@csiro.au) received B.S., M.Sc., and Ph.D degrees from Xi'an Jiao Tong University, Nanjing University of Posts and Telecommunications, and the Australian National University, in 1996, 1999, and 2004, respectively. Currently, he is a senior scientist and team leader in the wireless and networks program of CSIRO. From 1999 to 2001, he was an engineer in ZTE Corp., China, and from 2004 to 2010, he was a researcher at NICTA, Australia. His research interests are in the areas of wireless and optical communications, sensor networks, and smart grid communications.

Y. Jay Guo [F'14] (Jay.Guo@uts.edu.au) is a Distinguished Professor and director of the Global Big Data Technologies Centre at UTS. He is a Fellow of the Australian Academy of Technological Sciences and Engineering (ATSE) and IET. He has over 25 years of experience in academia, CSIRO, and industry, leading a number of research programs including wireless and networking, broadband networks and services, and 3G.

Advertisers' Index

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

SIEMENS

E20001-F812-P820-X-7600

Industrial Wireless LAN – SCALANCE W

Making the most of air.

Wireless data transmission for warehouse logistics systems: Environments with limited space call for exceptionally intelligent solutions.

SCALANCE W770: optimal for easy wireless machine networking and for cases where the mechanical requirements are particularly demanding.

The Industrial Wireless LAN components from Siemens ensure greater flexibility and reliability for your high-bay warehouse, rack feeder, and shuttle or carrier system. Numerous options are designed to meet your needs – such as PROFINET communication and safety, seamless integration into an Industrial Wireless LAN, space-saving installation in the control cabinet, compliance with the IEEE 802.11n standard, and a comprehensive portfolio of antennas.

Enhance your possibilities!

siemens.com/iwlan

COMMUNICATIONS STANDARDS

A Supplement to IEEE Communications Magazine

JUNE 2015

www.comsoc.org

RESEARCH AND STANDARDS: ADVANCED CLOUD AND VIRTUALIZATION TECHNIQUES FOR 5G NETWORKS

IEEE COMMUNICATIONS SOCIETY

A Publication of the IEEE Communications Society

We're here to help you write your 5G future.

Unprecedented experience in wideband mmWave, 5G waveforms, and Massive MIMO design.

The world of wireless communications is about to change. Again. The fifth generation—5G—will mean "everything, everywhere, and always connected." If you're on the cutting edge of this emerging technology, we can help you. We have expertise in all areas of 5G research and development, including wideband mmWave, radio spectrum, ASIC, antenna technologies, and network architecture. So from design simulation and verification to wideband signal generation and analysis, from component characterization to optical solutions, we've got you covered.

HARDWARE + SOFTWARE + PEOPLE = 5G INSIGHTS

- Keysight engineers are active in the leading 5G forums and consortia
- Keysight engineers are keynote speakers at 5G conferences and key contributors in top technical journals
- Applications engineers are in more than 100 countries around the world

KEYSIGHT TECHNOLOGIES

Unlocking Measurement Insights

Download our white paper *Implementing a Flexible Testbed for 5G Waveform Generation and Analysis* at www.keysight.com/find/5G-Insight

USA: 800 829 4444 CAN: 877 894 4414

IEEE COMMUNICATIONS SOCIETY

COMMUNICATIONS STANDARDS

A Supplement to IEEE Communications Magazine

JUNE 2015

RESEARCH AND STANDARDS: ADVANCED CLOUD AND VIRTUALIZATION TECHNIQUES FOR 5G NETWORKS

RESEARCH AND STANDARDS

Glenn Parsons

The importance of standards to the work and careers of communications practitioners is the basis of this publication. It is a platform for presenting and discussing standards-related topics in the areas of communications, networking, research, and related disciplines.

The need for market driven standards has recently been attracting the attention of researchers in academia and industry. Standardization is the platform for global consensus and can greatly benefit from the engagement of researchers. Standards need to be clear and unambiguous in order to allow ubiquitous implementation in the market. However, in many cases this is only realized after many complex discussions and contributions involving various company proponents and academic researchers. It is a natural fit for research to fill this role, and reiterates the value of highlighting research in standards. In future issues we look forward to adding commentary from industry visionaries, as well as a new standards education section.

This issue contains the first feature topic for the supplement on Communications Standards. Tarik Taleb assembled an esteemed editorial team to provide the feature topic, "Research and Standards: Advanced Cloud & Virtualization Techniques for 5G Networks." This topic comprises most of the articles in this issue, and an editorial from the feature topic editors will introduce each article. Readers will notice the ongoing commentary from leaders in various standards bodies, as well as a standards news section with several SDOs offering current status and pointers to SDO material. I trust that the reader will find these informative and illustrative of the fundamental role standards play in the communications networking ecosystem.

The first article in this third supplement on Communications Standards, written by Dusan Schuster, introduces the 150-year-old ITU in historical context starting from when it became the UN's specialized agency for telecommunications. The common rules for use of radio spectrum, developed by ITU, have been one of its successes; regulations for telecommunications have been more challenging. The strategy and goals of ITU are embedded in its Constitution, which can be modified at plenipotentiary conferences. Dusan suggests in his article that reform is needed in order for ITU to evolve to meet the current market requirements for standardization and supporting regulation. However, a reform initiative will require a great degree of political maturity, courage, and motivation to ensure the future of ITU. Finally, he suggests several reform guidelines that can assist ITU and its members prepare for the next 150 years.

The remaining articles comprise the feature topic. Future standards supplements will also be "anchored" around a topic of current market relevance to drive focus, similar to the feature topics of *IEEE Communications Magazine*. The next issue will focus on the Internet of Things (IOT). Proposals for future standards feature topics are welcome.

BIOGRAPHY

GLENN PARSONS [SM] (glenn.parsons@ericsson.com) is an internationally known expert in mobile backhaul and Ethernet technology. He is a standards advisor with Ericsson Canada, where he coordinates standards strategy and policy for Ericsson, including network architecture for LTE mobile backhaul. Previously, he has held positions in development, product management and standards architecture in the ICT industry. Over the past number of years, he has held several management and editor positions in various standards activities including IETF, IEEE, and ITU-T. He has been an active participant in the IEEE-SA Board of Governors, Standards Board and its Committees since 2004. He is currently involved with mobile backhaul standardization in MEF, IEEE and ITU-T and is chair of IEEE 802.1. He is a Technical Editor for *IEEE Communications Magazine* and has been co-editor of several IEEE Communications Society Magazine feature topics. He graduated in 1992 with a B.Eng. degree in electrical engineering from Memorial University of Newfoundland.

CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE
WIRELESS TECHNOLOGIES FOR DEVELOPMENT

BACKGROUND

We live in a world in which there is a great disparity between the lives of the rich and the poor. Using information and communication technologies for the purpose of development (ICT4D) offers great promise in bridging this gap through its focus on connecting human capacity with computing and informational content. It is well known that Internet access has the capability of fostering development and growth by enabling access to information, education, and opportunities. Wireless technology is a promising solution to this problem of digital exclusion and can be instrumental in democratizing access to the Internet by unfettering developing communities from the encumbering constraints of infrastructure (traditionally associated with broadband Internet provisioning). The focus of the proposed Feature Topic is on leveraging wireless technologies for development (W4D) to increase the quality of life for a larger segment of human society by providing them opportunities to connect resources and capacity, especially by provisioning affordable universal Internet access. To reflect recent research advances in using W4D, this Feature Topic calls for original manuscripts with contributions in, but not limited to, the following areas:

- "Global access to the Internet for all" (GAIA) using wireless technologies
- Do-it-yourself (DIY) wireless networking (such as community wireless networks) for the developing world
- Cost-efficient wireless networked systems appropriate for use in underdeveloped areas
- Fault-tolerant resilient wireless networking technologies for the developing world
- Rural/remote area wireless solutions (that can work efficiently with resource constraints such as intermittent and unreliable access to power/ networking service)
- Simplified network management techniques (including support for heterogeneous service delivery through multiple solutions)
- Using cognitive radio technology and 5G standards (with possible native integration of satellites) for GAIA
- Techno-economic issues related to W4D (including development of flexible pricing and incentive structures as well as new spectrum access models for wireless)
- Techno-political and cultural issues related to using wireless communications for development
- Using emerging networking architectures and future Internet architectures (e.g., cloud computing, fog computing, network functions virtualization [NFV], information centric networking [CN], software defined networking [SDN], and delay tolerant networking [DTN]) with wireless technologies for development
- Using wireless access/ distribution technologies (such as the following) for development: TV white spaces (TVWS); satellite communications using advances in geostationary orbit (GEO) and low Earth orbit (LEO) satellites; low-cost community networks; cellular technologies (CDMA 450, the open-source OpenBTS, etc.); wireless mesh and sensor networks; Wi-Fi-based long-distance (WiLD) networks; and wireless based wireless regional access networks (WRANs).

Since our aim with this Feature Topic is to provide a balanced overview of the current state of the art of using wireless technologies for development, we solicit papers from both industry professionals and researchers, and we are interested in both reports of experience and new technical insights/ideas.

SUBMISSIONS

Articles should be tutorial in nature and written in a style comprehensible to readers outside the specialty of the article. Authors must follow *IEEE Communications Magazine*'s guidelines for preparation of the manuscript. Complete guidelines for prospective authors are found at: http://www.comsoc.org/commag/paper-submission-guidelines.

It is important to note that *IEEE Communications Magazine* strongly limits mathematical content, and the number of figures and tables. Paper length (introduction through conclusions) should not exceed 4500 words. All articles to be considered for publication must be submitted through IEEE Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee) by the deadline.

SCHEDULE FOR SUBMISSIONS

- Submission Deadline: December 1, 2015
- Notification Due Date: March 1, 2016
- Final Version Due Date: May 1, 2016
- Feature Topic Publication Date: July 1, 2016

GUEST EDITORS

Junaid Qadir
School of EE and CS (SEECS),
National University of Sciences and Technology (NUST), Pakistan
junaid.qadir@seecs.edu.pk

Marco Zennaro
The Abdus Salam International Centre for Theoretical Physics (ICTP), Italy
mzennaro@ictp.it

Saleem Bhatti
University of St Andrews
St Andrews, UK
saleem@st-andrews.ac.uk

Arjuna Sathiaseelan
Computer Laboratory,
University of Cambridge,
United Kingdom
arjuna.sathiaseelan@cl.cam.ac.uk

Adam Wolisz
Technische Universität Berlin and University of California, Berkeley, USA
awo@ieee.org

Kannan Govindan
Samsung Research, India
gkannan16@ieee.org

BRIEF NOTES ON THE EVOLUTION OF THE PRACTICE OF STANDARDS

BY BRUCE KRAEMER, PRESIDENT, IEEE STANDARDS ASSOCIATION

Standards are pervasive, and typically those that are long lived become absorbed into everyday living and hence virtually invisible and forgotten. The purpose of this article is to revitalize interest and awareness.

Let's take a quick look at when and where the practice of standardization began and how it has evolved. The oldest recorded engineering standard was the cubit, an Egyptian measure of length based on the distance from the elbow to the end of the middle finger. As has been frequently noted in the world of standards, there always seems to be room for one more, and the cubit finds itself competing for mind share with other units such as the foot and meter.

But as important as engineering standards and units of measure are, they were preceded by oral and graphic communication standards. The organizational structure, strategy, and tactics of Roman legions are well documented. Their recorded history tells us that 2000 years ago the Roman military used signa or standards. The soldiers of the legions, centuries, cohorts, etc., were expected to treat the signa (standards) with great respect. The symbols and colors of each standard became both rallying points for troops and easily recognized graphics for the troops.

The Roman standard bearer for each infantry legion carried a long pole containing several important icons. At the top was an animal figure. Around 100 BC the standard animal became the Aquila, or eagle. Below it was a banner called the Vexillum. Another addition to the standard was the bronze SPQR plaque indicating the troop was representing the Senatus Populus que Romanus (Senate and People of Rome)

But the Romans used descriptive terms that have not carried forward to English, so where does the modern term come from? The best available research suggests that the word standard can be traced back to a Frankish word standhard. (The Frankish Empire was a collection of Germanic tribes occupying a region of what is now west Germany, Belgium, and France during the period from 300 AD to 800 AD.) The term was used to describe battle field ethic to hold ground against an adversary during combat and was certainly influenced by Roman practices. And although the Latin signa militaria terms once used are not the same as those we now use in English the concepts remain unchanged even today. Similarly, the long flag carrying poles have been replaced by URLs and email for communication.

So is it safe to declare that the word standard that we use has a commonly held definition? Unfortunately, I must also note that our high-purposed use of the term is not the only one in modern vernacular. In the early 1960s the British introduced a term "BOG Standard." There is no clear source for the slang phrase nor agreement on exactly which kind of bog was being referred to, but the term was certainly meant to express an uninteresting or uninspired form of standard.

In future columns I will try to avoid writing anything that would be so labeled.

COLLABORATIVE SECURITY: INTERNET SECURITY, THE INTERNET WAY

BY OLAF KOLKMAN, CHIEF INTERNET TECHNOLOGY OFFICER, INTERNET SOCIETY

How do we enable people to trust in the security of their communication and connections across the Internet while ensuring that the Internet remains open and accessible and the ability for permissionless innovation carries forward? Given that the Internet is a global network of networks without any centralized control, there is no magic answer. There are no single solutions that can be prescribed by governments or just implemented by network operators. The reality is that comprehensive Internet security only comes through the efforts of many different people collaborating together to take action to help ensure the security, resilience, and stability of the global Internet. It is in a similar spirit that we developed the Open Standards that brought us the Internet.

The Open Standards paradigm has been captured by the Open Stand Principles (see open-stand.org). These principles center around cooperation, adherence to principles, collective empowerment, and voluntary adoption. In my view the collaborative peer review, a drive for quality, and a problem-solving mindset are at the heart of the agility that caused the Open Internet to flourish. Security solutions for the Internet work much the same way.

The Internet Society's Collaborative Security framework (see https://www.internetsociety.org/collaborativesecurity) documents our high-level approach, or mindset, toward Internet security. Much of the same qualities that make the Internet what it is today are found there too. That approach has five key elements:

Fostering Confidence and Protecting Opportunities: The objective of security is to foster confidence in the Internet and to ensure the continued success of the Internet as a driver for economic and social innovation.

Collective Responsibility: Internet participants share a responsibility toward the system as a whole.

Fundamental Properties and Values: Security solutions should be compatible with fundamental human rights and preserve the fundamental properties of the Internet: the Internet Invariants (https://www.internetsociety.org/internet-invariants-what-really-matters).

Evolution and Consensus: Effective security relies on agile evolutionary steps based on the expertise of a broad set of stakeholders.

Think Globally, act Locally: It is through voluntary bottom-up self-organization that the most impactful solutions are likely to be reached.

Many standards, Internet security standards in particular, are difficult to deploy because the security function may only be beneficial if a certain deployment size has been reached. It is for those functions where a sense of collective responsibility and collaboration are important.

An example of technology where there is more benefit to the system than the individual actors is DNSSEC, which provides integrity and authenticity for the domain name system. When widely deployed, DNSSEC would allow interesting security innovations. A second example is DKIM, SPF, and DMARC. The technologies behind that acronym soup allow for tighter checks on the source of email and is an important tool in the toolbox to fight spam and phishing. Yet another example is the deployment of measures to prevent IP address spoofing. The combination of the ability of hosts to spoof their address and widespread botnets make for cost-effective and devastating Denial of Service attacks. All of these technologies help increase confidence in the Inter-

net, but do not offer an immediate return on investment.

I believe these technologies are all building blocks for a sustainable future. While the "smart" future is built, we must collectively collaborate on making sure that future allows us to live confidently among the many automated things that surround us. That requires an outward-facing, global perspective, and proactive collaboration in which Open Standards play a major role. By fostering continued confidence in the Internet, that sort of collaborative security ensures the Internet's ongoing success as a driver for economic and social innovation.

BIOGRAPHY

As Chief Internet Technology Officer, Olaf has responsibility for leading Internet Society's Strategic Technical activities, particularly as they pertain to issues and opportunities for enhancing the Internet's evolution. He has been actively involved with Internet technologies since his astronomy studies during the early nineties. The Internet became his professional focus in 1996 when he joined the RIPE NCC to develop the first version of what has become a worldwide test-network. In 2007 he became the managing director of NLnet Labs. Under his responsibility NLnet Labs produced open-source products, performed research on technical issues with global impact, and contributed actively to the regional and global collaborative standard and governance bodies (e.g. ICANN, RIPE, IETF), and "pushed the needle" on the development and deployment of DNSSEC. He describes himself as an Internet generalist and evangineer, somebody with deep knowledge on some of the Internet's technical aspects who particularly enjoys bridging the technology-society-policy gaps.

ETSI' s Vision of a Connected World

By Dirk Weiler, Chairman of the Board, ETSI (European Telecommunications Standards Institute)

ETSI produces globally applicable standards for Information and Communications Technologies (ICT). The high quality of our work and our open approach to standardization has seen our influence extend from our European roots to impact every part of the world. In 2014 we published over 2 300 standards, specifications, reports, and guides, bringing the total published since our establishment in 1988 to almost 37 000. All of our standards and specifications are available free from our website.

ETSI is a not-for-profit organization with more than 800 member organizations worldwide, drawn from 64 countries. Although we are an officially recognized European Standardization Organization, our members include some of the world's leading companies and national administrations working alongside R&D organizations, smaller businesses, and innovative start-ups. ETSI is at the forefront of emerging technologies. We are building close relationships with research bodies and addressing the technical issues that will drive the economy of the future and improve life for the next generation.

Over the years we have built up a portfolio of partnership agreements around the world. Experience has shown that working with others is the best way to achieve alignment between our standards and those produced elsewhere, to avoid the duplication of effort and to ensure that our work is widely accepted and implemented. Co-operation reduces fragmentation in standardization and is a key factor when dealing with the convergence of technologies.

An excellent example of ETSI's success is the development of standards for Mobile Broadband Communications, which is the foundation of the most successful ecosystem worldwide. It started as an ETSI Technical Committee, and in 1998 was turned into the 3rd Generation Partnership Project (3GPP), together with our partner organizations in China, Japan, Korea, the U.S., and now India. The GSM/UMTS/LTE system enables billions of users and devices to communicate, and thousands of companies to participate and contribute products, technologies, and services. It is the foundation for some of the most valuable companies today. It creates huge opportunities for other sectors such as transport (Intelligent Transport Systems), energy (smart grids), medicine (Tele-medicine) or the Internet of Things (IoT). While LTE is now being deployed in many parts of the world, we have already started looking into 5G, enabling a whole new range of opportunities. Themes that can be found for 5G are "expanding the human possibilities of cellular technology" or "a scalable service experience everywhere and anytime where people and objects will obtain virtual zero latency gigabit experience when and where it matters."

The communication ecosystem is highly dependent on standardization. ETSI's members continue investing heavily in R&D and contributing their best technologies to standardization. This is possible as ETSI's IPR policy ensures a balance between the fair use of standards by implementers and adequate and fair reward for the use of innovators' IPRs. IPR owners are requested to grant licenses on fair, reasonable, and non-discriminatory ("FRAND") terms. FRAND is a very successful concept that has enabled countless IPR license agreements for Standards Essential Patents. This concept is the basis for the vast majority of international standards organizations' IPR policies, including the common patent policy of ITU, ISO, and IEC.

Further success factors are best-in-class processes and tools meeting the needs of our members and partners. As our environment is constantly changing, these processes need to evolve. Examples are a revised process for adopting Publicly Available Specifications (PAS) as ETSI deliverables, or the currently ongoing work to cooperate between standardization and open source software (OSS) projects.

The ETSI decision making process is based on consensus, both for standards making and for governance processes. This ensures that all materially affected stakeholder groups are taken into account.

In order to make our vision of a connected world happen we must address many new and exciting challenges based on our broad diversity of ICT standards. We address M2M and the Internet of Things, Cloud standards and Network Functions Virtualization, Intelligent Transport Systems, Cyber and Network Information Security, ICT Energy Efficiency and Environmental Sustainability, and many more. For a better understanding of all the areas ETSI is supporting today we have grouped our standardization activities into the following clusters: "Home & Office" (connecting devices for home, SOHO and SME environments); "Better Living with ICT" (technologies that improve people's lives and environment); "Content Delivery" (serving content users across different business areas); "Networks" (building networks that support users' communication needs); "Wireless Systems" (wireless systems and their regulatory environment); "Transportation" (systems for people on the move); "Connecting Things" (integrating objects to create new networked services); "Interoperability" (interconnecting in a multi-vendor, multi-network, multi-service environment); "Public Safety" (communication systems and services for public safety); and "Security" (standards for reliable and secure communications).

No single standards organization is able to deliver all the necessary standards in order to transform the vision of a connected world into reality. Cooperation of all stakeholders is necessary to make the benefits of this vision available to everybody.

BIOGRAPHY

Dirk Weiler is Chairman of the ETSI Board and the ETSI IPR Special Committee, and Head of Standards Management & Horizontal in the Networks Business of Nokia. With 30 years of technical and management experience in the telecoms industry he regularly speaks and writes about standardization, patents and the interplay of both. He holds an advanced degree in physics (Diplom-Physiker) from the University of Cologne.

STANDARDS ACTIVITIES FOR NEXT GENERATION INTERNET-OF-THINGS (IOT) ARCHITECTURES

JAESEUNG SONG
SEJONG UNIVERSITY, KOREA,
CONVENOR, TEST WG, ONEM2M,
IEEE COMSOC IOT ARCHITECTURES RG CHAIR

The Standards Activities Council of IEEE Communications Society held a one-day working meeting at IEEE Headquarters on September 30, 2014, in order to identify primary standards development opportunities in the Internet-of-Things (IoT) and related areas. About 20 IoT experts from around the world were invited to the meeting.

The expert group's main objective was to demonstrate and document the steps necessary to establish an early IEEE Standards Activities' presence in key areas of opportunity. In the meeting, the invited IoT experts introduced their ideas and suggested areas of standardization in various areas and topics, which led the group to an intense discussion on a rough gap analysis to determine standardization opportunities.

After this intensive discussion on various IoT related standards issues and gap analysis, the participants suggested establishing three official research groups (RGs) and one study group (SG) to discover potential IoT related areas that IEEE ComSoc can contribute. The IoT Architectures RG is one of these RGs, and has the following objectives:

- Document various IoT architectures that recognize a representative subset of use cases that effectively cover most applications of IoT. Because of the diversity of use cases, there may be more than one architecture.
- Capture attributes that are important for all use cases, such as ubiquity, scalability, security, etc.
- Recognize real world constraints and principles that drive the specific architecture such as fog vs. cloud, real time vs. non real time, etc.

To achieve the given objectives, the IoT Architectures RG has collected various IoT architecture research proposals to explore the possibility of future standards activities from broad areas. The IoT Architectures RG ended up with a total of 15 detailed architectural topics. The topics are further categorized into four categories based on their similarities: IoT architecture evolution, enhanced data gathering and delivering, advanced radio access network architectures for IoT, and secure and reliable IoT architectures. Detailed topics for each category are shown below.

Cisco	Ericsson	InterDigital	KETI
EURECOM	National Chiao Tung University	Sejong University	Yonsei University
Singapore University of Technology and Design	DGE Lab. of Princeton University	NextComm	Institute of Information Industry
National Taiwan University	University of Sussex	University of College London	Tsinghua University

Table 1. Organizations contributing IEEE ComSoc IoT architectures RG.

Category 1. IoT Architecture Evolution (federation and data reuse)

C1.1. Semantic IoT architecture.

C1.2. Architecture styles, e.g., RESTful, CoAP, SOAP.

C1.3. Horizontal IoT architecture platform.

C1.4. Software defined networking for IoT.

C1.5. Proximity oriented IoT architecture.

C1.6. Data centric networking.

Category 2. Enhanced Data Gathering and Delivering

C2.1. Fog IoT architecture.

C2.2. Smart filter embedded IoT gateway architecture.

C2.3. Reduction of latency delay and handover cost for HetNets for IoT applications.

Category 3. Advanced Radio Access Network Architectures for IoT

C3.1. Fog network-based Radio Access Network (F-RAN) versus Cloud-based Radio Access Network (C-RAN) in IoT services.

C3.2. Energy efficiency for small cell networks in IoT architecture.

C3.3. Device-centric network structure and M2M communication for IoT.

C3.4. Satellite and terrestrial wireless integration for IoT.

Category 4. Secure and Reliable IoT Architectures

C4.1. Cyber-physical security for IoT services and applications.

C4.2. Self-configuration architectures for IoT.

At the time of this writing, approximately 23 participants from 16 organizations (including leading global IT companies, universities, and research institutes (see Table 1)) are contributing to the IoT Architectures RG. These contributors are now actively working to develop a white paper that will contain the following items:

- Introduction to a proposed IoT architecture or technology.
- Relevant IoT standards that cover the proposed IoT architectures.
- Details of the proposed architecture (and technology), e.g., new features, new architecture style, interworking, etc.
- Evaluation and experiment results.
- Impact on standards, relevance to standards, and how the presented materials would evolve current standards.
- Suggestions to IEEE ComSoc for the feasibility of the proposed IoT architectures and derive recommendations for where and how to start a new project for standardizing the proposed topic.

The current tentative schedule for the white paper is to fill in the technical contents of each topic by the end of June. After doing further standards gap analysis on the proposed architectures, the RG plans to release its white paper to the public by 30 Septembr 2015.

NEW WORK IN IETF ROUTING

JOEL HALPERN, ERICSSON

In the last few months the IETF has chartered an interesting new working group, and made progress on other work of wide interest.

The work on Bit Index Explicit Routing (BIER) for multicast was proposed late last year, and has now been chartered as an IETF working group [1]. Current multicast mechanisms use signaling to enable construction of distribution trees that deliver multicast packets to the domain exits that need to receive them. However, this places a significant state and signaling load on all of the routers inside the domain. The goal of the BIER work is to reduce this load, while still preserving the effective delivery only to places that need the packets, under the assumption that the entry point to the domain knows where all the relevant exits are topologically.

The approach BIER uses for this is to build a covering tree, using MPLS, and then to encode in each packet the subset of the tree that the packet is to

follow. This subset is represented by a bit mask that starts out representing all the exits that need to receive the packet. As copies are sent in different directions at replication points in the tree, the different copies retain only the bits for exits to which that copy is to be delivered. So the copies are replicated and delivered to the right exits, without any risk of duplication. Details about this approach can be found in the architectural draft the working group is considering [2].

One of the challenges with this work is that it assumes hardware support for preforming the needed bit mask manipulations in the fast path of the forwarding plane. While the concept can be verified by software update and replication techniques, operation at scale assumes hardware support. A number of major IP operators have expressed interest in this for a number of different use cases, including EVPN and data center multicast uses.

Another area the IETF has been working on recently is data plane support for service function chaining. While the Service Function Chaining (SFC) working group is more than a year old, there were several significant steps forward in ITS work during the first part of 2015. The work continues to be focused on defining the data plane behavior needed to deliver subsets of traffic to the specific service functions that need to see those subsets. This enables behaviors such as those described by the ETSI NFV work in their discussion of forwarding graphs. The overall objective is described in RFC 7498 [3]

Two significant additional steps are the completion by the working group of the architectural description of the data plane behavior [4], and the adoption by the working group of an encapsulation [5] format to serve as the basis for agreement on the needed dataplane encapsulation. This encapsulation provides for carrying service function path identification and packet metadata in a fashion THAT is independent of the transport mechanisms used for carrying the packets across the service delivery environment.

GETTING READY FOR 5G

DINO FLORE
QUALCOMM TECHNOLOGIES INC.,
CHAIRMAN OF 3GPP RAN

Mobile communication is rapidly evolving. The success in connecting everything everywhere is posing formidable challenges to our industry. Mobile broadband demand continues to grow exponentially. At the same time, the proliferation of new types of devices and services is rapidly increasing the number and types of connected devices wireless systems must deal with. More generally, these new types of devices will be serving a variety of new applications that provide new experiences and demand new forms of communication.

To support the expanded connectivity needs of the next decade, 3GPP started making plans for the standardization of next generation cellular technology, also known as 5G. A tentative 3GPP timeline for 5G standardization was recently endorsed, including plans for a 3GPP technology submission to the IMT 2020 process triggered by ITU-R.

One key aspect of the discussion is the definition of next generation radio technology, as this will be a key element to address the expanded connectivity needs of the future. This is what 3GPP RAN will start to discuss at the workshop that will take place on 17–18 September in Phoenix, AZ, USA. In December, 3GPP RAN will then start a new study to define the exact scope and requirements of next generation radio technology, so that the evaluation of technology solutions can start in 2016. As it is already clear that the work will include operation above 6 GHz, in September 3GPP RAN will also start a new project on channel modelling for higher frequencies so that the new channel model is available when the group starts to evaluate the different radio technologies.

Besides meeting future mobile broadband demand, the next generation 3GPP platform needs to enable a broad range of new services and connectivity paradigms. 3GPP SA has started an effort to articulate the service and connectivity vision by defining the service requirements of the next generation platform.

THE 2016 ITU WORLD TELECOMMUNICATION STANDARDIZATION ASSEMBLY (WTSA): CHALLENGES AND OPPORTUNITIES

BRUCE GRACIE
AFFILIATION??

As the 2016 WTSA approaches, the Telecommunication Standardization Sector (ITU-T) is facing the prospect of either addressing the need to adapt its structure and working procedures to the realities of a rapidly changing environment, or facing the inherent risk endemic to all organizations that rely on familiar traditions and conventions that tended to serve them well in facing challenges to their legitimacy and relevance. The ITU-T is no exception to this trend; however, there are important developments in its recent history that could serve to mitigate pressures to maintain the status quo, which is particularly important as the standardization landscape grows in complexity, includes the existence of competing regional and international organizations, forums, and consortia.

To illustrate the point, the ITU has now adopted at its 2014 Plenipotentiary Conference (PP-14) a viable results-based framework (strategic plan) which, through a series of well-defined strategic objectives, can link expected results with concrete deliverables. Second, based on decisions made at recent WTSAs, there is a clear recognition of the importance of a) the role of the private sector in contributing to the sustainability of ITU in the field of standardization; b) the need for the Union to provide and demonstrate value in pursuing topics within its core competencies and mandate; c) the development of strategies to strengthen collaboration and cooperation with other entities engaged in similar, yet complementary, fields of endeavour.

With regard to the first point, regular consultations between chief technology officers (CTOs) have been held since 2009 for the purpose of discussing, inter alia, which types of standards are needed in an increasingly complex and fragmented environment; to identify the different standards-related development organizations, along with their roles and capabilities; and to implement improvements in the current standards ecosystem. Topics for recent discussion have included IoT, ITS, next generation video technologies and standards, SDN and network function virtualization.

As part of the process of enhancing the engagement of the private sector in the standardization work of ITU, PP-14 also formally adopted an initiative to review the current methodologies for the participation of sector members, associates and academia in the activities of the Union. The non-member state constituencies will now have the opportunity to comment directly on their future engagement in ITU in terms of their participatory rights and obligations.

Regarding the value-added dimension of ITU's involvement in standards development, efforts are underway to address this critical point, including the

prospective establishment of a new "standards strategy" function intended to identify the main technological trends, as well as market, economic, and policy needs in the ITU-T's fields of activity through, for example, a series of well-planned and comprehensive consultations with industry. This is part of a wider effort underway in ITU-T to undertake a thorough strategic and structural review initiated by a decision of WTSA-12 in a so-called "review committee".

As part of the process of examining the wider standardization landscape and its many issues, a Global Standards Symposium was organized just prior to WTSA-12, included a key session on global standards collaboration. Among its principal conclusions was that a collaborative mechanism should be developed between related standards bodies to identify new work areas at an early stage, and agree on a common approach so as to bring the relative skills of the different bodies together in a cooperative manner in order to develop common international standards or suites of standards. By decision of WTSA-12 itself, a new initiative under the Telecommunication Standardization Advisory Group (TSAG) has been launched to pursue and develop strategies to strengthen collaboration, the results of which will form an important part of the deliberations at WTSA-16.

In conclusion, while activities are underway in ITU-T to enhance its status and legitimacy in an ever changing standards environment, the decisions of WTSA-16 will be crucial in ensuring that the momentum gained in these initiatives are carried over into the work programme for the following four-year period and beyond.

References

[1] IETF Charter for the BIER working group, available: http://datatracker.ietf.org/wg/bier/charter
[2] I.J.Wijnands *et al.*, "Multicast using Bit Index Explicit Replication," available: http://datatracker.ietf.org/doc/draft-wijnands-bier-archtiecture
[3] P. Quinn and T. Nadeau, RFC 7498, "Problem Statement for Service Function Chaining," available: http://www.rfc-editor.org/rfc/rfc7498.txt
[4] J. Halpern and C. Pignataro, "Service Function Chaining (SFC) Archtiecture," available: http://datatracker.ietf.org/doc/draft-ietf-sfc-architecture
[5] P. Quinn and U. Elzur, "Network Service Header," available: http://datatracker.ietf.org/doc/draft-ietf-sfc-nsh

The Open Networking Foundation

Mike McBride, Ericsson
Market Deputy Area Director, ONF

The Open Networking Foundation (ONF) is a user-driven organization dedicated to the promotion and adoption of SDN through open standards development. The signature accomplishment to date is the OpenFlow Standard, which enables remote programming of the forwarding plane. ONF's technical communities are organized into areas, councils, and groups. Areas handle specific issues related to SDN, and collaborate on SDN and the OpenFlow Standard regarding SDN concepts, frameworks, architecture, software, standards, and certifications. Councils provide overall leadership with respect to strategy, operational execution, and technical direction of the organization. Groups provide guidance and advise ONF on activities to help accomplish the organization's goals.

Table Type Patterns (TTP) is a major new work coming out of the ONF. TTPs are an optional enhancement to OpenFlow that allow controllers and switches to agree in advance on the forwarding pipeline details. The pipeline details outline specific groups of OpenFlow rules that will be supported by each table in the pipeline. By creating TTPs, multi-vendor interoperability becomes simpler to achieve between switches and controllers sourced from multiple vendors. A number of vendors, such as Broadcom and Corsa, are investing in TTPs. OpenDaylight (ODL) and Open Networking Operating System (ONOS) are adding support for TTPs.

ONF collaborates with a variety of standards organizations. ONF is working with IEEE on their 802.16r proposed protocol efforts within ONF's wireless and mobility working group. ONF has created formal liaisons with various subgroups of ITU, including the Q12/15, and Q14/15 question areas. The ONF and the IETF are currently in discussions regarding mobile and wireless standards and SDN. ETSI NFV and ONF are drafting a comprehensive collaboration agreement with several liaisons across various areas of each organization. ONF and OpenStack are collaborating on a number of technical efforts, including software development, and a liaison was established between the OIF and ONF's Optical WG to initially work on security related use cases.

The OpenFlow driver is the first open-source code developed by the SDN community to create an interoperable open-source implementation of the latest OpenFlow standard. This driver is intended for use by the larger SDN community, including network equipment vendors, ISVs, and operators for easy transitions between different implementations of the OpenFlow protocol.

The SampleTap application is an educational tool based on OpenFlow 1.0 and designed to run on an OpenDaylight controller. The SampleTap application allows any OpenFlow 1.0-compliant switch to be used as a tap-aggregation device. The SampleTap app gives the end user the ability to program a series of match/action requests on the incoming tap port, add port-chains for transformation (e.g. truncation, decryption etc.), and distribute captured packets to multiple capture devices simultaneously.

ONF recently launched the eighth project of OpenSourceSDN (the Open Source SDN software community sponsored by ONF), called Project FLORENCE. Started by members of the ONF Security WG, Project FLORENCE will solicit use cases for SDN security, and will look at the creation of specialized assessment tools for network security in SDN networks.

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.

International Telecommunication Union —150 Years of History: Adaptation to Change and the Opportunity for Reform

By the end of the 1980s the technology evolution and fresh regulatory dynamism announced global changes in the national and international telecommunication environment and consequently in the role of the ITU Constituencies. The author argues that for different reasons, the Union has not adapted itself with sufficient speed to these processes, and the performance of the ITU mission has been suffering due to its increasingly weaker leadership and inefficient governing bodies.

Dusan B. Schuster

Abstract

Through its history the International Telecommunication Union (ITU) has witnessed great inventions and has played an active role in the technology development that has been shaping the world of telecommunications. The ITU milestones are numerous. One of them is of particular importance because it symbolizes better than any other the need for harmonious worldwide development and desire for universal cooperation. It was established at the Atlantic City conference in 1947 when the entry of ITU into the United Nations (UN) family as a Specialized Agency for Telecommunications was decided.

By the end of the 1980s the technology evolution and fresh regulatory dynamism announced global changes in the national and international telecommunication environment and consequently in the role of the ITU Constituencies. For different reasons, the Union has not adapted itself with sufficient speed to these processes and over the years the technical spirit of the Atlantic City conference has been weakening and political debates have been installed instead. Performance of the ITU mission has been suffering due to its increasingly weaker leadership and inefficient governing bodies.

The most recent ITU Plenipotentiary Conference, held in Busan in 2014, was facing the challenge to address the need for reforms in order to ensure the sustainability of the Union into the future and reinforce its status as an advocate of global technical excellence. Unfortunately, the need for reform was not specifically recognized and reflected in the Objectives enunciated in the Strategic Plan for the Union for the period 2016-2019, and any reorganization of the structure and functioning of the Union, currently recognized as being archaic and conservative, was left in abeyance.

The analysis made and the conclusions reached in this paper may assist future high-level discussions on the subject matter and may facilitate an early promotion of the ITU Reform Policies. The reform guidelines offered herein may also further reinforce the views of those Member States that are sincerely concerned and interested in creating a new paradigm for a better future of their Union.

COMMUNICATIONS STANDARDS

Historical Context

The ITU Atlantic City conference was the first conference after WWII searching to redefine the purposes of the Union and find the means of achieving them. To maintain the ITU's role at the center of international cooperation on matters of telecommunications, an independent body to regulate the use of radio-frequencies (International Frequency Registration Board (IFRB) was created. The major decisions taken may be summarized as follows.

The role of the ITU conferences was amended and the Plenipotentiary Conference (PP) remained the main governing body. The time and date of these conferences would be set at the preceding conference, but could be changed if a new location was proposed by at least 20 Member States. The Administrative Conferences, responsible for revising the Regulations, would meet at the same time and place as the PPs. Furthermore, an Administrative Council (AC) was created to act on behalf of the plenipotentiaries in the period between the two conferences. The Council would have eighteen (18) members elected by the Plenipotentiary Conference every five years. The selection procedure for the Council members would follow the same guidelines as electing the members to the IFRB.[1]

The Plenipotentiary Conference determined that a Secretary-General, Assistant Secretary-Generals, and staff would need to be appointed by the AC to administer the General Secretariat of the Union. Geneva was chosen as the ITU's permanent seat.

An open-ended working group, the Provisional Frequency Board (PFB), was created to draw up the procedures for IFRB, including its duties. The Board would have eleven (11) members representing five administrative Regions: Americas, Western Europe, Eastern Europe, Africa, North Asia, and the rest of the world. The representatives for the IFRB were elected and they held their first meeting in September 1947.

The Atlantic City conference was well aware of the importance of the international order in the use of the radio-frequency spectrum. The IFRB members were elected as recognized experts and considered as "custodians of an international public trust" and not as representatives of their respective countries or regions. They were supposed to act independently and represent themselves only. To support their independence, they were paid from the ITU budget, and to maintain their objectivity they were expected not to request or receive any instructions, and the Member States were asked to refrain from attempting to influence them in performing their duties. The IFRB decisions were final and could be changed only by the

Dusan B. Schuster is an independent international consultant on telecommunication regulatory policies. He worked for the ITU for 25 years before retiring in 2007.

[1] The International Frequency Registration Board (IFRB) was created at the International Radio Conference and approved by the International Telecommunication Conference. The creation of the Board introduced a new concept in the regulatory Policy of the Union.

 0163-6804/15/$25.00 © 2015 IEEE

majority at the PP conference or at a World Administrative Radio Conference (WARC).

The decision to create the Board implied that the IFRB members must be "thoroughly qualified in the field of radio, possessing practical experience in the assignment and utilization of frequencies" and must be elected by the PP conference from the candidates proposed by the Member States.

Even with its limited powers, IFRB, acting on its founding values, proved useful for many years to come. It served the membership well and impartially, helping to ensure that the RR were interpreted coherently and strictly observed.

ITU Regulatory Mission

The use of the radio-frequency spectrum is a matter of Member State sovereignty. However, by ratifying the Treaty texts, the Member States agree to respect the common rules for using the spectrum, the goal being the efficient use and equitable access to it. The RR provide for the relative rights of the Member States if/or when inappropriate uses of the frequency spectrum give rise to a harmful interference.

For the majority of the Member States, application of the numerous Articles, Appendices, and Procedures proves to be uneasy, difficult, or even complicated. In addition, the properties of the current updating and ratification procedure, making the amended RR applicable only to a limited number of the Member States, represent certain legal anachronisms that one of the future Treaty conferences will need to deal with in order to agree on an alternative and acceptable solution.

The part-time Radio Regulatory Board (RRB) is supposed to act as a watchdog to monitor the compliance of the Member States with the RR. Its members are elected by the plenipotentiaries and all Member States may vote for the candidates from each Region, as voting is not restricted to the Members States of the Region concerned. The creation of the RRB (as a replacement for the permanent IFRB) in 1992 was perceived more as a structure of "transitional character" but after 20 years it continues to operate and no Member State has the courage to propose its suppression or, at least the modification of either its mission or its prerogatives and responsibilities!

The RRB proved in practice its limited juridical capacity to deal with complex technical questions/problems. Its inability to make bold decisions is disturbing. Such an approach is effectively harming the credibility of this body, making it "de facto" a politically inspired ad-hoc group with no responsibility for its actions that one may understand as not being in the service of justice and even less in the service of Member States!

The IFRB members in 1947 were supposed to be recognized experts specialized in radiocommunication matter, whereas the RRB members (12) today are "political figures" and representatives of their administrative regions, elected by the plenipotentiaries! Such evolution of the Board membership's profile is contrary to the Atlantic City conference's basic criteria as being high level and independent professionals that are acting in a neutral capacity.

The International Telecommunication Regulations (ITR) set up the Rules applicable to Administrations or Recognized Operating Agencies (ROAs), which were adopted in Melbourne at the World Administrative Telephone and Telegraph Conference (WATTC-88). They were amended or replaced (according to CS and CV) by the recent Treaty Conference, i.e. WCIT,[2] that took place in December 2012 (Dubai).

While the adoption of ITR in 1988 marked the beginning of the liberalization process in international telecommunications, the WCIT-12 conference left behind an unusual legislative practice. The embarrassment came from the misleading conference preparations that sparked an overall politicization of the subject matter that finally resulted in no consensus. The ambitions of some Member States to modify the ITR content together with a possible redefinition of the ITU's role, its legal jurisdiction over new services, while retaining the preeminent role on the subject matter, were simply refused.

The conference concluded in disagreement where, for the first time, a substantial number of the Member States present and having the right to vote refused to sign the Final Acts. This event is regrettable in particular due to the inability of the ITU leadership at the time to suggest compromises. The WCIT-12 conference is as a political fiasco and technical failure, the consequences of which may take quite some time to be absorbed by the membership.

Constitutional Evolution

The way toward better functioning and reform has proved to be difficult due to the diverging interests and different priorities among the membership. It would appear that the fundamental changes in the national telecommunication policies by the end of the 1980s (deregulation, liberalization, etc.) were not detected in time to have them translated into the ITU environment. The partial operational reforms, launched in the 1990s, led to more bureaucracy, thus not satisfying either the majority nor the minority of the Members and, in particular, the Sector Members (m). It was expected that each new PP cycle would bring new opportunities for intellectual reflection inspiring more reforms — new repartition of responsibilities — but this was not the case.

With the number of the Members increasing, the sovereign governments, although serving one Union, started grouping and creating smaller interest groups or "sub-unions." Closed ad-hoc groups, led by various interests, became an instrument for partial solutions over a wide range of issues. These groups became influential and even harmful to the federal Union, depending on the subject and degree of egoism, commonly interpreted or referred to as "a national interest."

With deregulation in process, the traditional monopolies have disappeared and industry, users, service providers, and network operators

The way toward better functioning and reform has proved to be difficult due to the diverging interests and different priorities among the membership. It would appear that the fundamental changes in the national telecommunication policies by the end of the 1980s were not detected in time to have them translated into the ITU environment.

[2] World Conference on International Telecommunications (new name for ex-WATTC).

Along with their interests in telecommunication applications and service offering, the governments have less and less to contribute, and many of them have nothing to contribute! This negative element is giving rise to politicization of any debate, thus introducing the anachronisms in the system.

have become the ITU's natural partners, Sector Members (m). Many administrations have handed off their operational functions to these partners that have suitable human and financial resources to perform, thus making the advancement in telecommunications feasible for the benefit of the respective economies.

In certain areas the Administrations of today are seen as sclerotic components of that process that has largely bypassed them. Along with their interests in telecommunication applications and service offering, the governments have less and less to contribute, and many of them have nothing to contribute! This negative element is giving rise to politicization of any debate, thus introducing the anachronisms in the system. The need for substantial reforms is obvious since the technology-inspired and market-driven initiatives have triggered the movement that has profoundly changed the national understanding of telecommunications and corresponding policies.

At the Nice PP-89, the much needed realism did not prevail except for a superficial refreshment of the management team. A High Level Committee (HLC) was created and over two years it accomplished its mission; however, it left behind many grey areas. At the Geneva APP-92, the governments were still not ready to admit the Union's increasing weaknesses, but for political reasons, they decided to abolish IFRB and suspended its members' function until the Kyoto PP-94.

The idea of a part-time Radio Regulations Board (RRB) was not accompanied with any technical argument; however, there was not enough force to reject it and even less political courage to invent a better solution. Replacement of IFRB by a part-time RRB was proven to be an error of great strategic importance.

The HLC, inter-alia, recommended the simplification of the RR through the Voluntary Group of Experts (VGE) comprised of representatives of ITU Member States. After two years, the task was completed in 1994; however, the regulatory procedures became even more complex and non-transparent, thus undermining the ability of Members to both understand and apply them. While the HLC recommendations concerning the new WRC format and cycle of world conferences were implemented, no serious analysis was undertaken on whether such an approach would be sustainable in an increasingly complex radio environment.

In 1994, the first World Telecommunication Development Conference (WTDC) took place in Buenos Aires where the Bureau for Development of Telecommunications (BDT) was created. The two Study Groups that were set up in the Telecommunication Development Sector brought into the ITU operational system a very unusual organizational feature. The functioning of BDT became on longer term a symbol for inefficiency and political arithmetic.

The Kyoto Plenipotentiary Conference in the same year confirmed that the Union was still looking for more reforms and appropriate status for Sector Members (m). But at the same time it was just too early to start another process that might appear to be as HLC bis! However, the Minneapolis PP-98 recognized the need and appreciated the aspirations for reforms by creating an open ended Working Group (ITU-2000) with the objective "to prepare for a modern ITU"! Unfortunately, even its high level and pragmatic suggestions were not sufficient to inspire the Administrations with new and original ideas. And over that period, the Sector Members (m) continued to demonstrate their creative force in different forums outside ITU.

The Union was facing a major obstacle, i.e. non-imagination coupled with system inertia. The short-term changes for the daily environment instead of a long-term strategy became part of the Union's weaknesses. More inspiration and political courage and more cooperation through full transparency were missing.

Creative Environment for Sector Members

The Radiocommunication Sector with its Bureau (BR) has a very distinctive role in the application of the ITU regulatory policy, complemented with scientific activities. To a large extent, it has an essentially different mission from the Telecommunication Standardization Sector with its Bureau (TSB), and none of them is comparable to the Telecommunication Development Sector and its Bureau (BDT). The technology convergence in ex-CCIR and ex-CCITT did call for an early merging, which indeed was done based on the HLC recommendations, but in the wrong direction; CCIR merged with IFRB although they were in essence two bodies with different missions (regulatory on the one hand and academic research on the other). The evolution of the environment proved it was not appropriate; in fact, it was wrong.

The existence of the regional standardization organizations (ETSI and others), with significant resources available, raised the question of the ITU's future in its standardization mission. If the ITU's structures do not allow for liberty in the creation of standards, and if the "standardization for the market and users" can be accomplished effectively outside it, why would the experts of the Sector Members (m) still be coming to ITU and having their products approved by an "external circle", these being the Government administrations?

Operators, industry, and telecommunication service providers, as the driving force in the ITU's standardization processes, had ambitions and important potentials, but no assurances of sovereignty over their decisions to be made in the area of exclusive competence. The academic studies in the ITU-R and ITU-T sectors are driven largely by industry where the developing countries are practically absent.

It is therefore essential to create an environment where the Sector Members (m) would be given an adequate status and sovereignty over their technical decisions. They should assume primarily the responsibility for the work to be accomplished in their area of competence and for budgeting of the associated activities. Non-governmental entities should find their responsibilities properly labeled and should enjoy full autonomy in drafting their

policy guidelines as well as programming and budgeting.

WRC: Concept to Be Improved

The conference Agendas are saturated with numerous items, some of them being complex, so no WRC can be managed reasonably and satisfactorily. The resulting anachronisms are weakening the ITU's capacity to make its regulatory regime stable, equitable, and just. Furthermore, the practice to postpone most of the issues for further study and to deal with certain "hot" issues through the WRC Resolution mechanism is considered inappropriate.

The concerned Administrations and Sector Members (m) aspire to operate with an instrument applicable in the same manner to all members, with equal rights to all. Currently, the majority of the membership is misled by the WRC decisions that regularly leave behind a "new updated" version(s) of the RR that may never apply to all parties simultaneously. It is therefore urgent to give thought to different types of intergovernmental multilateral meetings/conferences to deal with administrative issues based on a transparent platform, prepared for a technical discussion in advance.

The major negative performances of the current system to be rectified are:

- The WRC regularity is not necessary and is even counter-productive. The conference preparatory rules and procedures are archaic and not adapted to the needs of the organization, thus contributing to the the WRC's inefficiency.
- The WRC has become a "political gathering" instead of remaining an ITU technical Forum in service of its mission, i.e. addressing the administrative and regulatory issues in conformity with the ITU regulatory policy, as defined in its CS, CV, and RR.
- Each WRC agenda appears to be "a wish list" saturated with complex items (four years in advance), whereas the important and burning issues are usually dealt with in a non-transparent way among the administrations concerned, often outside the Conference's official meetings.
- Making coordinated and harmonized decisions on many items has become difficult/impossible due to late contributions and complex subjects, assisted by the chaotic considerations at the conference. Such an environment is making acceptable solutions unachievable.
- The "Resolution type solutions" that have become modern after WRC 2000 (Istanbul) are making the core RR provisions "de facto" an auxiliary and degraded legal instrument, pending ratification of the modified Treaty texts, which makes RR a political platform for satisfying interests that are not necessarily universal and transparent.
- A just application of the updated RR provisions by the Bureau cannot be equally applied to all parties with equal rights, due to the phenomenon of late (or none) ratification of the RR Treaty. This has necessarily direct and negative implications on the work of RRB on the one hand, and of Member States on the other if a disagreement/dispute arises among parties concerned.

Contrary to the ITU tradition, the current WRC are used to generate an unnecessary number of politically sponsored Resolutions that are affecting the character of the RR in a way that makes their implementation difficult. They are diluting the regulatory character and deviate considerably from the RR's original intent, as established in Atlantic City.

It is therefore desirable to review the whole concept and put a balance in the RR updating process by creating a simple and transparent environment, where each of three elements in the regulatory process, i.e. the RR establishment, RR implementation, and RR amendments, should be thoroughly re-examined and put into a clear and workable perspective.

Pursuing this understanding and taking account of the digital technology applications in most radiocommunication services, the question is whether the "Intergovernmental Conference", as a multilateral coordination meeting (MCM), is still an appropriate concept for a dynamic Frequency Spectrum Management or should ITU agree on and introduce a different practice, with the ITU secretariat having a new and an important role?

A Platform for the Constitutional Change

ITU needs "a strategic vision" that the Member States may wish to adopt and to apply its new mantra. While the Plenipotentiary Conference in 2014 was successful in adopting a results-based Strategic Plan that focuses on linking strategic, financial, and operational planning, the ITU "vision, mission and values" need to be more clearly integrated into its strategic objectives and program priorities. The bold decisions by the plenipotentiaries must become a priority instead of being concentrated on electoral campaigns (five members of the executive management team leaders, 12 part-time RRB members, and 48 members of the Council) every four years.

Returning to the Atlantic City wisdom, the Organization needs a compact leadership in terms of having three elected officials, i.e. the Director General (DG) and two deputy-Director Generals (DDG). They should be elected by the Plenipotentiary Conference for a period of either two four-year terms or one six (6) year term and thus may serve only one mandate with no possibility of renewal for the post in question. Moreover, the elected officials after the natural expiration of their mandate should not compete for any other elected post.

A political consensus on these basic values and modern policy orientations is therefore necessary. Egoistic behavior of the traditional constituents looking only after their immediate interests is no longer valid/tolerable and governments must acknowledge that they are no longer the (only) creator of the modern telecommunication environment and ICT future. The Member States need to recognize this reality by agreeing on a **new repartition of roles among their governments and the private sector**. The Reform should preserve the ITU integrity and reestablish its credibility with new properties, making it:

While the Plenipotentiary Conference in 2014 was successful in adopting a results-based Strategic Plan that focuses on linking strategic, financial, and operational planning, the ITU " vision, mission and values" need to be more clearly integrated into its strategic objectives and program priorities.

Without compromising the inter-governmental profile of the Union, it is essential to create new and favorable conditions where a non-governmental ITU role would be given a respectable status, where creative Partners would be treated on equal footing.

- Universal for all the Members States and Sector members.
- Revitalized in structure to be functionally manageable and efficient.
- Innovative and productive, becoming the world's first reference for the ICT regulatory matters.

Without compromising the inter-governmental profile of the Union, it is essential to create new and favorable conditions where a non-governmental ITU role would be given a respectable status, where creative Partners would be treated on equal footing. These Partners should find their responsibilities properly labeled in order to enjoy full autonomy in drafting the respective policy guidelines as well as programming and budgeting their activities in the field of their competence.

These are the Sector Members (m), and no longer the Governments, that possess the necessary creativity and dynamism, both in the field of finances and human resources. With the creation of a new Entity within ITU, the Partners would get a responsive operational structure that would provide an attractive and catalyst environment for all parties to work together and be considered as equal. The members of this new Entity would carry on academic and research studies related to the telecommunication standardization in all ICT segments, including radiocommunications.

This new Entity should become a facilitator for collaboration and co-operation of different regionally established ICT standardization bodies, competent industry forums, and regional institutions related to the telecommunications.

Conclusions

Any ITU Reform, and in particular a constitutional one, is a complex matter and will require a consensus among the 193 Member States that may have different interests and priorities as well as very different strategic perspectives on what may be best for their Union. It is evident that the Member States have their "national interest" on behalf of which they may accept or reject any (good or better) initiative for the constitutional reform. History has proven that such divergent interests have often prevented implementation of well-founded initiatives, either for improvement of the Union's operational performance or those for enhancement of the working efficiency of the Union.

Therefore, the Reform initiative will require a great degree of political maturity, courage, and motivation within the scope of shared vision and shared responsibility for the future of the ICT and ITU. It will need to involve all Member States through a fully transparent process, thus permitting actual Sector Members (m)/future Partners, to make useful and respected contributions.

The following Reform guidelines (Conclusions/Recommendations (C/R)) are offered in this view to stimulate the understanding and reflection toward a rational and substantial review process that should be well prepared, well understood, approved and implemented over a longer period of time. After reaching a consensus on these guidelines, concrete proposals for objectives may be gradually modulated and complemented, as appropriate.

C/R 1. The federal structure with three Sectors must be remodelled to create comprehensive conditions to be able to deal with technical, operational, and policy matters. The content of the two traditional ITU missions, i.e. **the international regulatory issues related to the Frequency Spectrum Management** (FSM), and **adoption of the global standards for telecommunications**, needs to be adjusted. The emphasis should be given to the policy matters where an open discussion and coordination may facilitate harmonization of the national policies (including issues related to the Internet).

C/R 2. The potential of the public/private sector partnership and resulting synergy should adopt a new image and new content to reflect the primary role of the private sector in the modern, competitive telecommunications environment. The emphasis should be placed on the regional telecommunication organizations (APT, CITEL, CEPT, PATU, etc.) and ITU regional offices in order to de-centralize the current development work, to eliminate politicization, to avoid bureaucracy, and to allow the membership to demonstrate their creativity.

C/R 3. With the understanding that ITU type federalism is an outdated format, the future constitutional architecture should follow a pragmatic business concept. The starting PP period (2014–2018) would appear to be an appropriate timing to raise awareness of the fundamental issues. Electing a new leadership in 2014 is a positive sign for better understanding the need for reforms.

C/R 4. The Governments should concentrate and work on regulatory matters related to FSM, associated primarily with the intergovernmental meetings as appropriate and when needed (WAC, WTC, or WRC). The Partners should enjoy complete autonomy within the new Entity and should be given the responsibility for all the convergence/standardization matters on a global scale.

C/R 5. The ITU operational culture must be reviewed and revised in order to attract Partners and innovative thinkers to contribute substantially in the new areas of relevance (example: software-defined networking and virtualization). It is essential to reestablish ITU technological relevance as its driving force, while fully developing its ability to innovate technologically.

C/R 6. A wide participation of the non-governmental entities from industry and private sector (that may be sitting in other "fora" and setting de-facto standards) should be encouraged by promoting their new constitutional rights, allowing them to be the major source of the ITU's creative potential.

C/R 7. Accepting the new features, based on proposed guidelines, the Member States may wish to reaffirm their sense of pragmatism to continue playing an important role in global telecommunications. This way, the Union would preserve the necessary capacity and would build upon a sufficient capability to be responsive to the challenges, being already

high on the agenda in the modern telecommunication environment.

C/R 8. A thorough examination of financing the Union, together with proposals for a different model, the rules and procedures related to the funding of the Union need to be reviewed and possibly a different model is to be adopted.

C/R 9. The rational use of working languages at various levels of working processes (currently six) is an important issue within the scope of the Reform process, in financial and operational terms.

Provided the ideas were successfully promoted and complemented along the road, and if the preparatory work by the Member States is performed well, then a gradual implementation of the Reform process would be feasible within the period 2018-2022. This would require that concrete proposals be advanced and agreed to by the next Plenipotentiary Conference in 2018.

If these aims were to be accepted as guidelines for the Union's long term Policy, the Governments should demonstrate a sufficient political wisdom over the next two PP periods, in order to make the implementation plan feasible.

References

[1] ITU WARC and RARC Conferences (1970–1992)
1974 and 1975 LF/MF Region1 and Region3 (in two Sessions)
1977 Broadcasting Satellite Conference (BC)
1981 and 1988 MF Region 2 (in two Sessions)
1982 and 1984 FM Region 1 (in two Sessions)
1983 BC Sat Region 2
1984 and 1987 HFBC (in two Sessions)
1985 EMA/MM
1985 and 1988 ORB-85 and ORB-88
1986 and 1988 MF Region 2 (in two Sessions)
1986 and 1989 AFR-TV (in two Sessions)
1987 MOB-87
1992 WARC (general administrative radio conference)
[2] ITU Master international Frequency Register — MIFR
[3] ITU Plenipotentiary Conference, Atlantic City, May/Oct. 1947.
[4] ITU Radio Regulation Board RRB.
[5] ITU WRC-12, Geneva 2012.
[6] ITU WICT-12, Dubai 2012.
[7] ITU Radio Regulations, RR 2012.
[8] ITU International Telecommunication Regulations, ITR 2012.
[9] ITU Plenipotentiary Conference, Busan 20.10 — 07.11.2014.
[10] D. B. Schuster, "International Telecommunication Union — Challenges for the Plenipotentiary 2014 — Time for Change," *IEEE Commun. Mag.*, Feb. 2014.

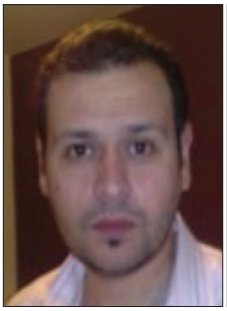
Biography

Dusan B. Schuster (dusan.schuster@ties.itu.int) is an independent international consultant on telecommunication regulatory policies. He retired from ITU in 2007 after 25 years of working in high level positions, most of this period as a counselor to the Secretary General and other elected officials on ITU policies and radiocommunication regulatory issues.

RESEARCH & STANDARDS: ADVANCED CLOUD & VIRTUALIZATION TECHNIQUES FOR 5G NETWORKS


Kan Zheng


Tarik Taleb


Adlen Ksentini


Chih-Lin I


Thomas Magedanz


Mehmet Ulema

The evolution of mobile network architecture is an essential part in the development process of the fifth generation (5G) of cellular mobile systems, and that is through the incorporation of advanced cloud technologies and network function virtualization techniques. The new network architecture needs to support a wide range of high data rate applications and services, offering capacities of up to multiple gigabits per second, yet meeting extremely stringent latency and reliability requirements under a diverse variety of scenarios. Thus the automation of network control and overall system management to achieve such an ambitious set of performance targets became crucial. Significant global effort for the necessary new technologies has been initiated.

Emerging paradigms, such as and not limited to, Software-Defined Networking (SDN) and Network Function Virtualization (NFV) represent a first concrete step towards this direction, catalyzing the idea of decoupling software defined control plane from hardware driven data plane and thus the virtualization of network functions on general purpose hardware. There is an increasing trend towards implementing more and more functions of mobile communications systems in software, e.g., for signal and protocol processing. This will significantly influence the future developments of 5G technologies and architectures. In addition to savings in both operational and capital expenditure, the introduction of logically centralized controllers enables the employment of various intelligent control algorithms. However, how to leverage the benefits of NFV in both the Radio Access Network (RAN) and the mobile Core Network (CN) is yet to be investigated fully.

The aim of this special issue is to highlight the 5G requirements and how they can be met through novel mobile network architecture designs. The special issue mainly focuses on new technologies being researched that has great potential to impact standards. Original contributions were solicited on topics of relevance to the evolution of mobile network architectures, and particularly on SDN, NFV and cloud computing. Out of a total of 35 high quality submissions, the following eight papers have been selected for publication.

To cope with the growing mobile traffic and the associated computation demand, the authors of "On the Computation Offloading at Ad Hoc Cloudlet: Architecture and Service Modes" propose offloading computation at ad-hoc cloudlets. They categorize the computation offloading techniques into three modes, i.e., remote cloud service (RCS), connected ad hoc cloudlet service (CCS), and opportunistic ad hoc cloudlet service (OCS). The authors propose OCS mode for the first time, and conduct a detailed analytic study for the three modes. In order to provide insights for facilitating the utilization of the newly proposed mode, the authors build up a general and novel mathematical model, based on which optimal problems are formulated and solved.

In "Software Defined and Virtualized Wireless Access in Future Wireless Networks: Scenarios and Standards," F. Granelli *et al.* provide provide an overview on the different perspectives of using SDN in future wireless access networks, including 5G and Wireless Local Area Networks (WLANs). Several typical scenarios for mobile SDN are given and related standardization activities are discussed. A framework for the cellular network to be software-defined is presented in the third article, "Cellular Software Defined Network — a Framework," by Bradai, *et al.* In this proposed network architecture, the authors exploit the capability of the mobile edge networks to gather information related to the network as well as to the users.

Different from academia, the industry has different views on how to apply new technologies into future wireless communication systems. As one of the largest operators in the world, China Unicom proposes a new network structure, i.e., i-Net, to better support localized services with SDN and cloud computing techniques. i-Net is introduced in the fourth article, "i-Net: New Network Architecture for 5G Networks" written by Wang, *et al.* i-Net presents an evolution path from existing networks while still providing backward compatibility and allowing direct inter-BS communications through newly defined interfaces. Field trials are carried out to demonstrate the effectiveness of i-Net and a detailed roadmap from LTE-R10 to the future i-Net is also discussed.

Meanwhile, another new way to implement the radio access network (RAN), called as Virtual Radio Access Technology (RAT), is also proposed in the fifth article, "Virtual RATs and a Flexible and Tailored Radio Access Network Evolving to 5G" by Chen et. al., presents a new way to implement virtual RANs. The authors present the overall architecture and protocol stack of Virtual RAT. Two essential features of flexible C/U plane separation and coordination between Virtual RATs are studied as well. Examples to flexibly implement the RAN by using Virtual RAT are also provided. In addition, the feasibility in hardware to support implementation of Virtual RATs is analyzed. Besides, the sixth article, "Cloud Assisted HetNets Toward 5G Wireless Networks" by Zhang *et al.,* proposes using cloud to ease and simplify the

operation and management of the heterogeneous network (HetNet). This article highlights possible approaches toward 5G and discusses the challenges of operation and management.

Mobile social networking has been gained great attention recently. The seventh article, "Content Distribution over Content Centric Mobile Social Networks in 5G" by Su et al, presents a novel framework for content centric mobile social networking, which can well support content distribution in 5G. This article presents the design and highlights the strengths and potentials of content centric mechanisms for mobile social networks over 5G. More specifically, the authors propose a content centric mobile social network architecture, which consists of mobile users, communities, small cells and macro cells to work with each other cooperatively.

The eighth and last article of the feature topic, "Software-Defined Networking Security: Pros and Cons," by Dabbagh *et al.*, presents security aspects of SDN networks, which is very important for 5G systems. The authors pointed out that there are three key characteristics, which let SDN be security over traditional networks. On the other hand, new threats and attacks targeting the forwarding plane, the control plane, or the links connecting the two planes make SDN more vulnerable. Then, several preventive and mitigation techniques are described as well in this article. It is expected that SDN networks can exploit further the paradigm's security advantages in the future.

As Guest Editors, we would like to thank all the authors for their submissions to this Feature Topic. The interest and quality of submissions were beyond our imagination. We are also grateful to the reviewers for the timely responses and their valuable comments to improve the quality of the articles. We appreciate the support from both Mr. Glenn Parsons, current Editor-in-Chief of IEEE Communications Magazine Supplement on Communications Standards, and Dr. Osman S. Gebizlioglu, Editor-in Chief of IEEE Communications Magazine. We also appreciate the help of Joseph Milizzo, Jennifer Porcello, and Charis Scoggins throughout the publication process. Finally, our hope is that the readers of IEEE Communications Magazine Supplement on Communications Standards enjoy the articles of this Feature Topic, and would consider contributing to future editions.

BIOGRAPHIES

KAN ZHENG [SM] (zkan@bupt.edu.cn) is currently a professor in Beijing University of Posts &Telecommunications (BUPT), China. He received the B.S., M.S. and Ph.D degree from BUPT, China, in 1996, 2000 and 2005, respectively. He has the rich industry experiences on the standardization of the new emerging technologies. He is the author of more than 200 journal articles and conference papers in the field of resource optimization in wireless networks, M2M networks, VANET and so on. He holds editorial board positions for several international journals. He has organized several special issues in famous journals including IEEE Communications Surveys & Tutorials, Transactions on Emerging Telecommunications Technologies (ETT). He was the general Vice-Chair of Mobiquitous, 2012, and workshop co-chair on QoE in Energy-Efficient Wireless Networks in IEEE ISCIT' 2012. Also, he was the TPC Co-Chair of IEEE PIMRC 2013, and TPC Track Chair of IEEE WiMob, 2015 and IEEE SmartGridComm, 2015. He has served as TPC member of IEEE conferences including INFOCOM, ICC, Globecom, and VTC.

TARIK TALEB (Tarik.Taleb@neclab.eu) is an IEEE Communications Society (ComSoc) Distinguished Lecturer and a senior member of IEEE. He is currently a Professor at the School of Engineering, Aalto University, Finland. Prior to his current position, he was working as Senior Researcher and 3GPP Standards Expert at NEC Europe Ltd, Heidelberg, Germany. He was then leading the NEC Europe Labs Team working on R&D projects on carrier cloud platforms. Before his work at NEC and till Mar. 2009, he worked as assistant professor at the Graduate School of Information Sciences, Tohoku University, Japan, in a lab fully funded by KDDI, the second largest network operator in Japan. From Oct. 2005 till Mar. 2006, he was working as research fellow with the Intelligent Cosmos Research Institute, Sendai, Japan. He received his B. E degree in Information Engineering with distinction, M.Sc. and Ph.D. degrees in Information Sciences from GSIS, Tohoku Univ., in 2001, 2003, and 2005, respectively. His research interests lie in the field of architectural enhancements to mobile core networks (particularly 3GPP's), mobile cloud networking, mobile multimedia streaming, congestion control protocols, handoff and mobility management, inter-vehicular communications, and social media networking. He has been also directly engaged in the development and standardization of the Evolved Packet System as a member of 3GPP's System Architecture working group. He is a board member of the IEEE Communications Society Standardization Program Development Board. As an attempt to bridge the gap between academia and industry, he has founded and has been the general chair of the "IEEE Workshop on Telecommunications Standards: from Research to Standards", a successful event that got awarded "best workshop award" by IEEE Communication Society (ComSoC). Based on the success of this workshop, he has also founded and has been the steering committee chair of the IEEE Conference on Standards for Communications and Networking (IEEE CSCN). He is/was on the editorial board of the IEEE Transactions on Wireless Communications, IEEE Wireless Communications Magazine, IEEE Transactions on Vehicular Technology, IEEE Communications Surveys & Tutorials, and a number of Wiley journals. He is serving as chair of the Wireless Communications Technical Committee, the largest in IEEE ComSoC. He also served as Secretary and then as Vice Chair of the Satellite and Space Communications Technical Committee of IEEE ComSoc (2006–2010). He has been on the technical program committee of different IEEE conferences, including Globecom, ICC, and WCNC, and chaired some of their symposia. He is the recipient of the 2009 IEEE ComSoc Asia-Pacific Best Young Researcher award (Jun. 2009), the 2008 TELECOM System Technology Award from the Telecommunications Advancement Foundation (Mar. 2008), the 2007 Funai Foundation Science Promotion Award (Apr. 2007), the 2006 IEEE Computer Society Japan Chapter Young Author Award (Dec. 2006), the Niwa Yasujirou Memorial Award (Feb. 2005), and the Young Researcher's Encouragement Award from the Japan chapter of the IEEE Vehicular Technology Society (VTS) (Oct. 2003). Some of His research work has been also awarded best paper awards at prestigious conferences.

ADLEN KSENTINI [SM] (adlen.ksentini@irisa.fr) is an Associate Professor at the University of Rennes 1, France. He is a member of the INRIA Rennes team Dionysos. He received an M.Sc. in telecommunication and multimedia networking from the University of Versailles. He obtained his Ph.D. degree in computer science from the University of Cergy-Pontoise in 2005, with a dissertation on QoS provisioning in IEEE 802.11-based networks. His other interests include: future Internet networks, mobile networks, QoS, QoE, performance evaluation and multimedia transmission. He is involved in several national and European projects on QoS and QoE support in Future wireless and mobile Networks. Dr. Ksentini is a co-author of over 60 technical journal and international conference papers. He received Best Paper Award from IEEE ICC 2012, and ACM MSWiM 2005. He is TPC Chair of the IEEE third Workshop on Standards on telecommunication (collocated with Globecom 2014), and workshop chair of the ACM/IEEE QShine 2014. Hei is guest editor for IEEE Wireless Communication Magazine, SI on Research & Standards: Leading the Evolution of Telecom Network Architectures. He has been in the technical program committee of major IEEE ComSoc conferences, ICC/Globecom, ICME, WCNC, PIMRC.

CHIH-LIN I (icl@chinamobile.com) is the China Mobile Chief Scientist of Wireless Technologies, in charge of advanced wireless communication R&D effort of 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; C-RAN and soft base station. He received her Ph.D. degree in Electrical Engineering from Stanford University, has almost 30 years experience in wireless communication area. She has worked in various world-class companies and research institutes, including wireless communication fundamental research department of AT&T Bell Labs; Headquarter 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 itsVP and the Founding GD of its Communications Technology Domain. He received the Trans. COM Stephen Rice Best Paper Award, and is a winner of CCCP "National 1000 talent" program. She was an elected Board Member of IEEE ComSoc, Chair of ComSoc Meeting and Conference Board, and the Founding Chair of IEEE WCNC Steering Committee. She is currently the Chiar of FuTURE Forum 5G SIG,a Steering Board Member of WWRF, an Executive Board Member of GreenTouch, and a Network Operator Council Member of ETSI NFV.

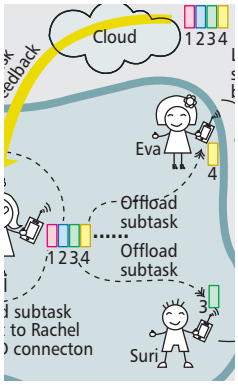
THOMAS MAGEDANZ (thomas.magedanz@fokus.fraunhofer.de) Ph.D. is professor of the chair for Next Generation Networks (AV — Architektur der Vermittlungsknoten in German) in the electrical engineering and computer sciences faculty of the Technische Universität Berlin, Germany, where he is educating Master and PhD students in the converging fields of SDN-based control platforms for multimedia and M2M/IOT applications on top of converging fixed and mobile broadband networks. In addition, he leads the Next Generation Network Infrastructures (NGNI) Competence Center at Fraunhofer Institute FOKUS in Berlin, Germany, where he is responsible for the performance of major international R&D co-operations and related academic and industry projects in the context of future seamless communication infrastructure prototyping. In this context he is a globally recognized pioneer for the development and delivery of advanced network and service technology software tools, known as the OpenXXX toolkits, and related testbeds, known as the FOKUS playgrounds, in the fields of Mobile Next Generation Networks. Well known examples include the OpenIMSCore, OpenEPC OpenMTC and the new Open5GCore, as well as the Open IMS Playground, the FUSECO-Playground and the new Open 5G Playground being part of the 5GBerlin testbed.

MEHMET ULEMA (mehmet.ulema@manhattan.edu) is a professor at the Computer Information Systems Department at Manhattan College, New York. Previously, he held management and technical positions in Daewoo Telecom, Bellcore (now called Telcordia), AT&T Bell Laboratories, and Hazeltine Corporations. He is an active member of IEEE. Currently, he is a ComSoc Director of Standards Development. He served as the chair and co-founder of the IEEE Communications Society's Information Infrastructure Technical Committee. He is involved in numerous IEEE conferences. He was a General co-chair of IEEE BlackseaCom 2014. He was the Technical Program chair for IEEE Global Communications (Globecom) conference) in 2009. He was the General co-chair of IEEE Network Operations and Management Symposium (NOMS) in 2008, the program chair of IEEE International Communications Conference (ICC) in 2006, IEEE Consumer Communications and Networking Conference in 2004. He received MS & Ph.D. in Computer Science at Polytechnic University (now called Polytechnic Institute of New York University), Brooklyn, New York. U.S.A. He also received BS & MS degrees at Istanbul Technical University, Turkey.

On the Computation Offloading at Ad Hoc Cloudlet: Architecture and Service Modes

With the development of mobile devices in terms of increased memory and computational capability, a novel peer-to-peer communication model for mobile cloud computing is proposed to interconnect nearby mobile devices through various short range radio communication technologies to form mobile cloudlets, where every mobile device works as either a computational service provider or a client of service requester.

Min Chen, Yixue Hao, Yong Li, Chin-Feng Lai, and Di Wu



Abstract

As mobile devices are equipped with more memory and computational capability, a novel peer-to-peer communication model for mobile cloud computing is proposed to interconnect nearby mobile devices through various short range radio communication technologies to form mobile cloudlets, where every mobile device works as either a computational service provider or a client of a service requester. Though this kind of computation offloading benefits compute-intensive applications, the corresponding service models and analytics tools are remaining open issues. In this paper we categorize computation offloading into three modes: remote cloud service mode, connected ad hoc cloudlet service mode, and opportunistic ad hoc cloudlet service mode. We also conduct a detailed analytic study for the proposed three modes of computation offloading at ad hoc cloudlet.

Introduction

With an ever increasing number of mobile devices and the resulting explosive growth of mobile traffic, 5G networks call for various technology advances to transmit traffic more effectively while changing the world by interconnecting a tremendous number of mobile devices. However, mobile devices have limited communication and computation capabilities in terms of computation power, memory, storage, and energy. In addition to the broadband bandwidth support from 5G, cloud computing needs to be utilized to enable mobile devices to obtain virtually unlimited dynamic resources for computation, storage, and service provisioning that will overcome the constraints in smart mobile devices [1]. Thus, the combination of 5G and cloud computing technology is paving the way for more attractive applications that involve compute-intensive task execution at "small" devices carried by mobile users who enjoy "large" capabilities enabled through the new technologies.

With the support of mobile cloud computing (MCC) [2], a mobile user basically has one more option to execute the computation of their applications, i.e., offloading the computation to the cloud [3]. Thus, one principal problem is: under what conditions should the mobile user offload the required computation to the cloud [4]. An illustrative scenario of computation offloading to a remote cloud is shown in Fig. 1a, where the user is covered by WiFi. Since the terminal device at the user end has limited resources, i.e., hardware, energy, bandwidth, etc., the cellphone itself is infeasible to finish some compute-intensive task. Instead, the data related to the computation task can be offloaded to the remote cloud via WiFi which is free of charge. After the execution of the computation task in the remote cloud, the computation result will be sent back to the user's cellphone. We call this conventional computation offloading mode "remote cloud service" (RCS), as shown in Fig. 1. Typically, WiFi is used in RCS and user mobility is strictly limited to areas covered by WiFi. In the case that a mobile user moves into an outdoor environment without WiFi coverage, the computation task must be offloaded to the remote cloud via a cellular network, which results in high communication cost. Compared with traffic offloading [5, 6], applications involving computation offloading usually are more delay-sensitive. Thus, how to design a cost-effective computation offloading service mode while achieving good quality of experience (QoE) becomes an important problem to be solved in this work.

COMMUNICATIONS STANDARDS

As mobile devices are equipped with increased memory and computational capability, a novel peer-to-peer communication model for MCC is proposed to interconnect nearby mobile devices through various short-range radio communication technologies (e.g., WiFi, Bluetooth, etc.) to form mobile cloudlets, where a mobile device can work as either a computational service provider or a client of a service requester [7]. In recent years, direct short range communications between cellular devices, known as device-to-device communications (D2D), has been widely studied, with a comprehensive discussion of related usage cases and business models given in [8]. In [9], a fundamental question while utilizing a cloudlet is proposed as follows: whether and under what conditions a mobile cloudlet is feasible for providing mobile application services. To answer the question, cloudlet size, a cloudlet node's lifetime and reachable time are defined. However, computation offloading is only allowed when a computation node (i.e., a cloudlet service requester) stays in contact with a service node (i.e., a cloudlet service provider) [9]. In this paper, we denote the previous computation offloading mode for cloudlet assisted service as "connected ad hoc cloudlet service" (CCS).

Under CCS mode, the computation task is offloaded to another local cloudlet node when there are available D2D connectivities during the whole procedure, including computation offloading, computation execution, and computation feedback. Though local D2D connectivity might be unavailable due to network dynamics,

Min Chen and Yixue Hao are with Huazhong University of Science and Technology.

Yong Li is with Tsinghua University.

Chin-Feng Lai is with National Chung Cheng University.

Di Wu is with Sun Yat-sen University.

 0163-6804/15/$25.00 © 2015 IEEE

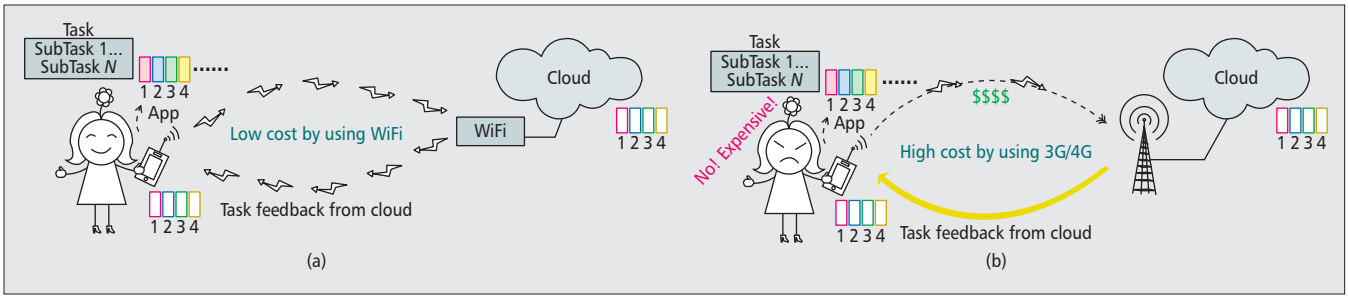


Figure 1. Illustration of computation offloading through remote cloud service mode.

one advantage of employing CCS is the lower communication cost and short transmission delay compared to the case when the computation task is offloaded to a remote cloud.

In this paper, we articulate the features of RCS and CCS modes as follows:

RCS mode: With the support of stable 3G/4G, computation nodes offload their computation tasks to a remote cloud at any time. The advantage of this mode includes high reliability during the provisioning of on-demand services. The disadvantage is the unpleasant cost incurred by using cellular network resources, as shown in Fig. 1b.

CCS mode: After available D2D connectivity is built up between a computation node and a service node, the energy cost is economical since local wireless (e.g., WiFi) can be utilized for content delivery and whatever direct communications can be involved to complete the execution of a computational task. The limitation of CCS mode is the strict requirement on the contact duration between a computation node and a service node to guarantee enough processing time for the offloaded computational task. Once a computation node and a service node are disconnected due to mobility or other network dynamics while the offloaded computational task is not finished, the computation execution is failed.

Thus, the major focus of previous work is to minimize the cost by finding an optimal tradeoff between CCS mode and RCS mode. As shown in Table 1, either RCS mode or CCS mode has the disadvantage of high cost or limited mobility. In order to solve this issue, this paper proposes "opportunistic ad hoc cloudlet service" (OCS) modes, which are further classified into three categories: OCS (back & forth), OCS (one way-3G/4G), and OCS (one way-WiFi). Table 1 compares the features of various computation offloading service modes in terms of different performance metrics such as cost, scalability, mobility support, freedom of service node, and computation duration.

In summary, the contributions of this paper include:

- We summarize the existing categories of computation offloading via a cloudlet or the cloud as CCS mode or RCS mode.
- The features of existing computation offloading modes are analyzed, and a novel computation offloading mode named OCS is proposed.
- We differentiate the two cases where subtasks can be cloned or not in OCS and propose the optimization problem for scheduling sub-tasks.

The remainder of this paper is organized as follows. We first present the OCS architecture. Next, we analyze the issues of the OCS architecture and describe optimization problems in OCS. We then conclude the paper.

System Overview and Architecture

In this paper we propose a novel service mode for cloudlet-assisted computing by considering the following realistic scenario. The typical contact duration might be too short to guarantee a valid computation offloading, execution, and result feedback under CCS mode. However, it is reasonable to presume that the contact duration is enough to transmit the content associated with the computation to the service node via D2D

Structure	Service mode	Cost	Scalability	Mobility support	Freedom of service node	Computation duration
Remote cloud	RCS (3G/4G)	High	Coarse	High	N/A	Medium
	RCS (WiFi)	Low	Coarse	Low	N/A	Medium
Ad hoc cloudlet	CCS	Low	Coarse	Low	Low	Low
	OCS (back & forth)	Low	Medium	Medium	Medium	High
	OCS (one way-3G/4G)	Medium	Fine	High	High	High
	OCS (one way-WiFi)	Low	Fine	High	High	High

Table 1. A comparison of service modes for task offloading.

connectivity. After the connection between the computation node and the service node is over, the computation is still processed in the service node for a certain amount of time until the sub-task execution is finished. We call this new service mode "opportunistic ad hoc cloudlet service" (OCS). The basic idea of OCS is the utilization of the opportunistic contacts among a computation node and service nodes while not limiting the mobility of the user. It is assumed that each computation task has a certain deadline, by the end of which the computation result should be sent back from the service nodes to the computation node. Based on the locations of the service node, there are three possible scenarios:

- Meeting a computation node again.
- Losing D2D connectivity with a computation node while seeking help from 3G/4G.
- Losing connectivity with a computation node while WiFi is available.

Corresponding to the above three scenarios, we divide OCS service modes into the following three categories.

Terminology	Definition
Computation task	Workflow with a certain amount of data associated with a computation
Computation node	A node that has a computation task to be executed, it can also be called task node
Service node	A node that is available to provide service for a computation node to handle an allocated sub-task
Sub-task	Multiple sub-tasks consist of a computation task
Sub-task result	The execution result of a sub-task by a service node
RCS	A computation offloading service mode, where the computation task is uploaded to a remote cloud, then the computation result is sent back to a computation node
CCS	A computation offloading service mode, which requires a computation node always keep connectivity with a service node
OCS	A computation offloading service mode, which mainly utilizes the opportunistic contacts among a computation node and service nodes
Back & forth	Service node meets twice with a computation node, which enables the submission of the sub-task result in the second meeting
One way-3G/4G	When the sub-task is finished, a service node is out of connectivity with a computation node while the cellular network is available
One way-WiFi	When the sub-task is finished, a service node is out of connectivity with a computation node while WiFi is available
Computation allocation	The method by which a computation node allocates multiple sub-tasks
Computation classification	The mechanism to classify the computation task based on the feature of the sub-task
D2D	Device to device communication method
eNB	Evolved node base station

Table 2. Definition of terminologies.

OCS (back & forth): In [10], Li *et al.* proposed a mobility-assisted computation offloading scheme, which calculates the probability of meeting twice between a computation node-service node pair. To calculate the probability, the statistics of node mobility are used. Before the computation task deadline, once a service node meets a computation node again while the execution of the allocated sub-task is finished, the sub-task result can be successfully sent to the computation node. We call this computation offloading service mode via ad hoc cloudlet as "back-and-forth service in cloudlet." However, user mobility under this mode is typically limited within a certain area, in order to guarantee the second meeting between the computation node and the service node. The mobility support of OCS (back & forth) mode should be higher than that of CCS and RCS (WiFi). Thus, the rank of mobility support is marked as "medium" in Table 1.

OCS (one way-3G/4G): It is challenging to achieve cost-effective computation offloading without sacrificing mobility support and the mobile nodes' freedom, i.e., a service node might roam to another cell. For the sake of generality, let us consider the scenario without WiFi coverage, where a service node needs to upload the sub-task result to the cloud via 3G/4G. Typically, the data size of the sub-task result ($S_{sub\text{-}tk}^{result}$) is smaller than the size of the original sub-task that a service node receives (i.e., $S_{sub\text{-}tk}^{recv}$). Let r denote the ratio of $S_{sub\text{-}tk}^{result}$ and $S_{sub\text{-}tk}^{recv}$. The lower r is, the better the performance of OCS (one way 3G/4G) mode will be.

OCS (one way-WiFi): In the case that the service node roams to a different cell that is covered by WiFi, e.g., the mobile user goes back home, the sub-task result can be uploaded to the cloud via WiFi. For most practical values of r, the communication cost under this service mode is between RCS (WiFi) and RCS (3G/4G).

Figure 2 shows illustrative examples to explain the above three OCS service modes. Rachel gets a compute-intensive task, which is infeasible to be executed in a timely manner by her own mobile phone. Within the range of D2D connectivity, Rachel has four friends named Bob, Eva, Cindy, and Suri, whose mobile phones are in idle status. Thus, Rachel divides the computation task into four sub-tasks, and forwards the corresponding contents to their four mobile phones via D2D links, respectively. Cindy does not move much, and keeps connectivity with Rachel. After execution, Cindy's sub-task result is sent to Rachel directly under either CCS or OCS (back & forth) service mode. In comparison, Bob and Suri move to another cell before the end of the sub-task execution. Thus, they use OCS (one way-3G/4G) service mode to upload the sub-task result to the cloud. As for Eva, let us assume she comes back to her home with WiFi support, thus utilizing OCS (one way-WiFi) service mode.

Since OCS does not require that both the computation node and the service nodes should keep in contact or locate in a certain area, it has higher scalability. In fact, OCS is especially useful in some applications where the size of the data content associated with a task is large while

the size of the result data is relatively small. Given the application of image segmentation as an example, the size of pictures generated by an image sensor of the mobile device can be large. Thus, transmitting the whole picture to the cloud via a 3G/4G link consumes a large amount of valuable bandwidth and energy. Under OCS (one way-3G/4G) mode, the computation node first delivers the whole picture along with image segmentation code to a service node via D2D links. After performing the segmentation code for the whole picture, only a certain region of interest (ROI) in the picture is obtained as the computation result. Uploading the small size of ROI to the cloud via 3G/4G link leads to the reduction of communication cost compared to RCS mode, while OCS offers more freedom for the computation node and the service node without the requirement of a strict contact duration compared to CCS mode. Moreover, RCS mode is usually not viable in some emergent situations such as disaster recovery and medical emergency handling in outdoor environments. By comparison, OCS mode is more flexible to overcome this problem. Therefore, OCS mode can be considered as a novel compromise between CCS mode and RCS mode, which achieves more flexibility and cost effectiveness to enable a more energy-efficient and intelligent strategy for computation offloading by the use of a cloudlet. To our best knowledge, this paper is the first to propose OCS mode. In order to provide insight when using this new computation offloading mode, we build a mathematical model and present solutions to a few optimal problems. In the proposed OCS mode, we make the following assumptions:

- A computation task can be divided into multiple sub-tasks.
- Depending on specific applications, the sub-tasks are different from each other or identical.
- The content delivery for a sub-task can be finished during a short contact period between the computation node and the service node, while the computation execution incurs relatively longer delay.
- Each service node can only be utilized for one sub-task.
- We do not consider packet loss and the communication cost to be decided by the amount of data transferred in the network.

Analysis for the OCS Mode

Before the analysis for OCS mode, we present some assumptions, based on which the following issues are considered in terms of computation allocation and classification.

Computation Allocation

Before offloading the computation task to the ad hoc cloudlet, the principal problem for the computation node is how to divide the computation task into a certain number of sub-tasks. This problem is related to the number of service nodes (N_{sn}), as well as their processing capabilities. Intuitively, the larger N_{sn} is, the higher accumulated capacity of computation that the cloudlet possesses, and the computation duration will be shorter.

However, it is critical to decide an optimal number of sub-tasks, and the following strategies can be considered:

- If we assume that the number of nodes within the same cell is stable in a certain period (e.g. a day), the node number can be estimated by the number counted in the last period. The number of sub-tasks can be the same as the number of service nodes.
- Based on social networks, the relationships among users can be extracted. Usually, the service nodes that have strong connections with the computation node are more likely to accept the sub-task assignments.

Once we know the number of sub-tasks, the remaining problem is to decide how to divide the computation task. There are several ways to allocate the computation sub-tasks:

Static Allocation: For general cases, the computation node does not know the computational capability of the service nodes. Thus, for the sake of simplicity, we assume the capability of all service nodes is similar. The computation node assigns sub-tasks equally among service nodes. However, the weak point of this strategy is that the computation duration is calculated based on the worst delay by a bottleneck service node.

Dynamic Allocation: More realistically, the capabilities of the service nodes should be different. When the information of the service nodes in terms of processing capability can be obtained, we can achieve the assignment of sub-tasks more intelligently. Typically, if the capability is higher, sub-tasks with heavier computation load will be allocated.

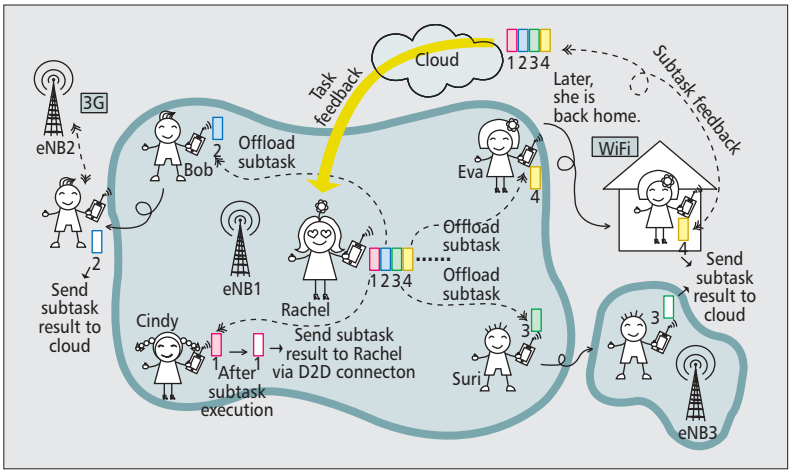


Figure 2. Illustration of the computation offloading at ad hoc cloudlet.

Computation Classification

For computation classification, we further divide the sub-tasks based on the following two situations: the sub-tasks are different from each other, or the sub-tasks can be cloned to each other.

Opportunistic ad hoc cloudlet service without computation task clone: Given an example scenario as shown in Fig. 3a, Rachel has 10 pictures, each of which contains a special ROI. If the dynamic method is used, the computation node (corresponding to Rachel) can divide the computation task (i.e., handling the 10 pictures)

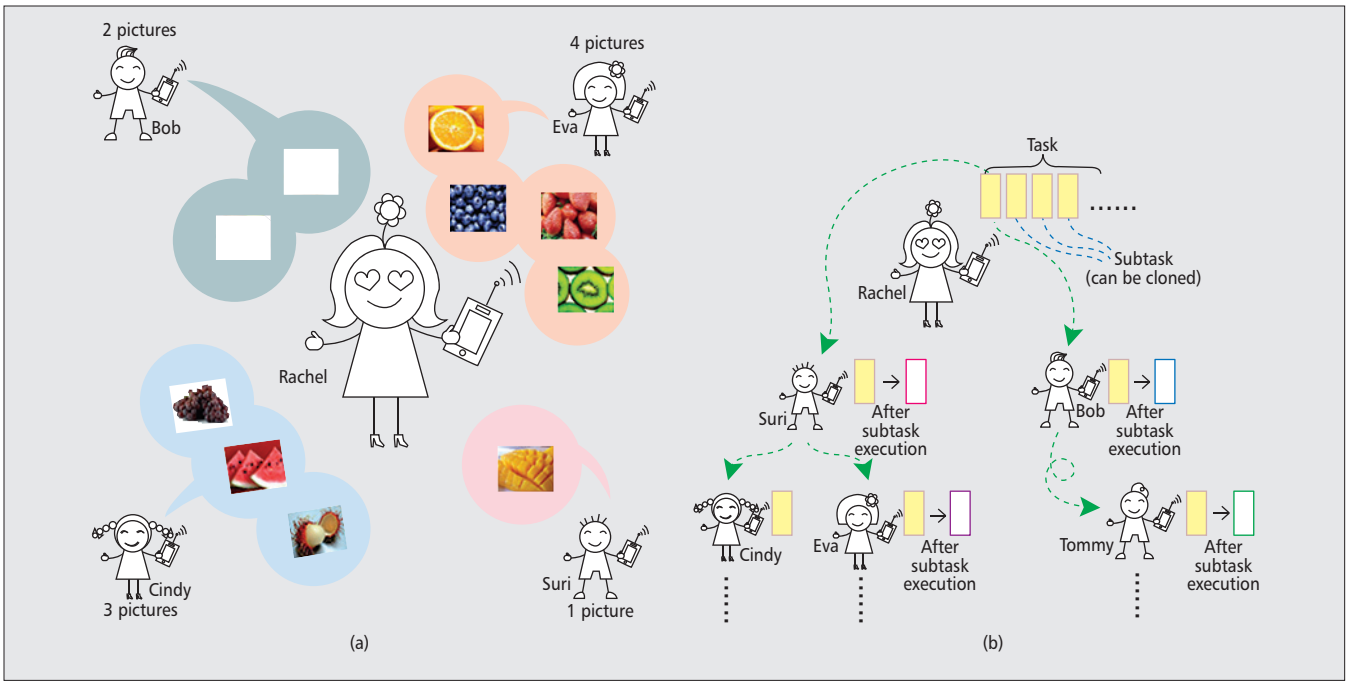


Figure 3. Illustration of sub-task offloading in opportunistic ad hoc cloudlet service: (a)without computation task clone; (b) with computation task clone

into four sub-tasks with various set of pictures which are assigned to Cindy, Bob, Suri, and Eva, respectively. Each person has a different smart phone with specific computational capability. As an example, Cindy and Bob get three and two pictures, while Suri and Eva are allocated one and four pictures, respectively. Obviously, the four sub-tasks are different in two ways. First, the number of pictures that each sub-task contains is different; second, each picture is different. After ROIs are segmented, they will be sent back to the computation node.

Opportunistic ad hoc cloudlet service with computation task clone: In some applications, a computation task can be divided into numerical equivalent sub-tasks. For this case, after a service node receives a sub-task, the sub-task content can be cloned and offloaded to another service node, which is similar to the epidemic model in online social network. As shown in Fig. 3b, the sub-task containing the same content is duplicated to nearby service nodes, in order to accelerate the execution of the computation task. Regarding energy consumption, the cost will be decreased since D2D is utilized during the flooding of the sub-tasks. There are two important parameters for building the model in this situation: the initial number of computation nodes, and the current number of service nodes that have the sub-tasks.

Optimization Problem in OCS

As shown in Table 1, different service modes have both advantages and disadvantages. Trade-offs arise when we need to provide users with high QoE, while saving communication cost and maintaining a degree of scalability to enable a wide range of intelligent applications.

In this section cost under these different service modes will be analyzed. Let us consider the scenario where M nodes exist in the cell and WiFi is not available by default (if the situation has WiFi, we use WiFi first). For the sake of simplicity, assume there is only one computation node, which has a computation task with a total size of computation load Q. The task can be divided into n sub-tasks. Assume each service node can process x_i workload in dynamic allocation, then $\Sigma_{i=1}^{n} x_i = Q$. The energy cost includes three parts, i.e., computation offloading, computation execution, and computation feedback.

RCS: Let $E_{n\to c}^{cell}$ denote the per unit communication cost from the computation node to the cloud; let $E_{c\to n}^{cell}$ denote the per unit communication cost for cloud-based result feedback; let E_{proc}^{cloud} denote the per unit energy cost for processing the computation task in the cloud. Then the total cost in RCS can be calculated as:

$$C_{RCS} = \sum_{i=1}^{n} (E_{n\to c}^{cell} x_i + E_{proc}^{cloud} x_i + r E_{c\to n}^{cell} x_i) = Q(E_{n\to c}^{cell} + E_{proc}^{cloud} + r E_{c\to n}^{cell}) \quad (1)$$

CCS: The major energy consumption is caused by D2D communications and the energy required to periodically probe the surrounding nodes. Let E_{D2D} denote the per unit communication cost from the computation node to the service node; let E_{proc}^{node} denote the per unit energy cost for the service node to process a sub-task locally; let ρ denote the probing cost per time unit; let t^* represent the task duration for a successful computation offloading. Please note that t^* is related to the average meeting rate of two nodes in the cell, which is denoted as λ. Then,

$$C_{CCS} = \sum_{i=1}^{n} (E_{D2D} x_i + E_{proc}^{node} x_i + r E_{D2D} x_i) + M \rho t^*$$
$$= Q(E_{D2D} + E_{proc}^{node} + r E_{D2D}) + M \rho t^* \quad (2)$$

OCS: If we consider a typical scenario where WiFi is not available and a service node roams to another cell, there are two cases exist: OCS (back & forth) and OCS (one-way-3G/4G). In the case of OCS (back & forth), we need to consider the probability (P, $0 \le P \le 1$) of the service node meeting the computation node twice, where D2D canbe utilized to deliver the sub-task result. Otherwise, the cellular network is the only option to deliver the sub-task result in OCS (one-way-3G/4G). Then,

$$C_{OCS} = Q(E_{D2D} + E_{proc}^{node}) + rPQE_{D2D} + r(1-P)$$
$$= Q(E_{n\to c}^{cell} + E_{c\to n}^{cell}) + M \rho t^* \quad (3)$$

Typically, the cost for a service node to offload computation task to the cloud or for the cloud to send back the computation result to the computation node via 3G/4G is larger than the D2D cost, i.e., $E_{n\to c}^{cell}$, $E_{c\to n}^{cell} > E_{D2D}$, and the processing cost in a service node is larger than in the cloud, i.e., $E_{proc}^{node} > E_{proc}^{cloud}$. Thus, considering energy and delay under various scenarios, it is expected that flexible trade-offs should be achieved according to specific application requirements. Figure 4a shows the comparison of energy cost under RCS and OCS modes. The cost of RCS is represented by the solid blue curve, while the other lines represent the cost of offloading by the use of OCS mode with various r. As $E_{n\to c}^{cell}$, $E_{c\to n}^{cell} > E_{D2D}$, when r is less than 1, OCS mode always has lower cost than RCS. However, when r is larger than 1, the cost of OCS increases with the increase of r, and the rate of growth is larger than the rate of growth with RCS. Furthermore, when E_{D2D} increases, the cost of OCS becomes larger. In summary, OCS outperforms RCS in the following cases, as shown in Fig. 4a:

1) $E_{D2D} = 0.5, r < 2$.
2) $E_{D2D} = 1, r < 1.15$.
3) $E_{D2D} = 1.5, r < 0.5$.

In Fig. 4b the cost of CCS and OCS is compared. Among the three schemes, OCS with computation clone exhibits the lowest cost and always outperforms the other schemes, because OCS with clone yields the fastest speed to complete the sub-tasks. When $0.00002 \le \lambda \le 0.00014$, CCS outperforms OCS without computation clone in terms of energy cost, since OCS without computation clone needs to upload sub-task results to the cloud while CCS saves this cost. When λ increases, the contact duration becomes smaller, which may cause the failure of sub-task's execution in CCS. Thus, when λ is larger than 0.00014, OCS without computation clone performs better than CCS.

Figure 4c shows the comparison of computation duration with OCS and CCS modes. Given a fixed λ, the computation duration of OCS is shorter than that of CCS. With the increase of λ, OCS yields better delay performance, because the frequency of the computation node meeting the service nodes increases with a larger λ. For CCS, the computation duration gradually decreases from a small λ (e.g., 0.00002 to 0.0001). However, with the continuous increase of λ, the computation duration of CCS starts to raise again, because the contact duration (meeting time) becomes smaller, which causes insufficient contact time to enable a successful sub-task offloading, execution, and feedback. As discussed above, we can draw the conclusion that:

- RCS mode: If the computation task is high-

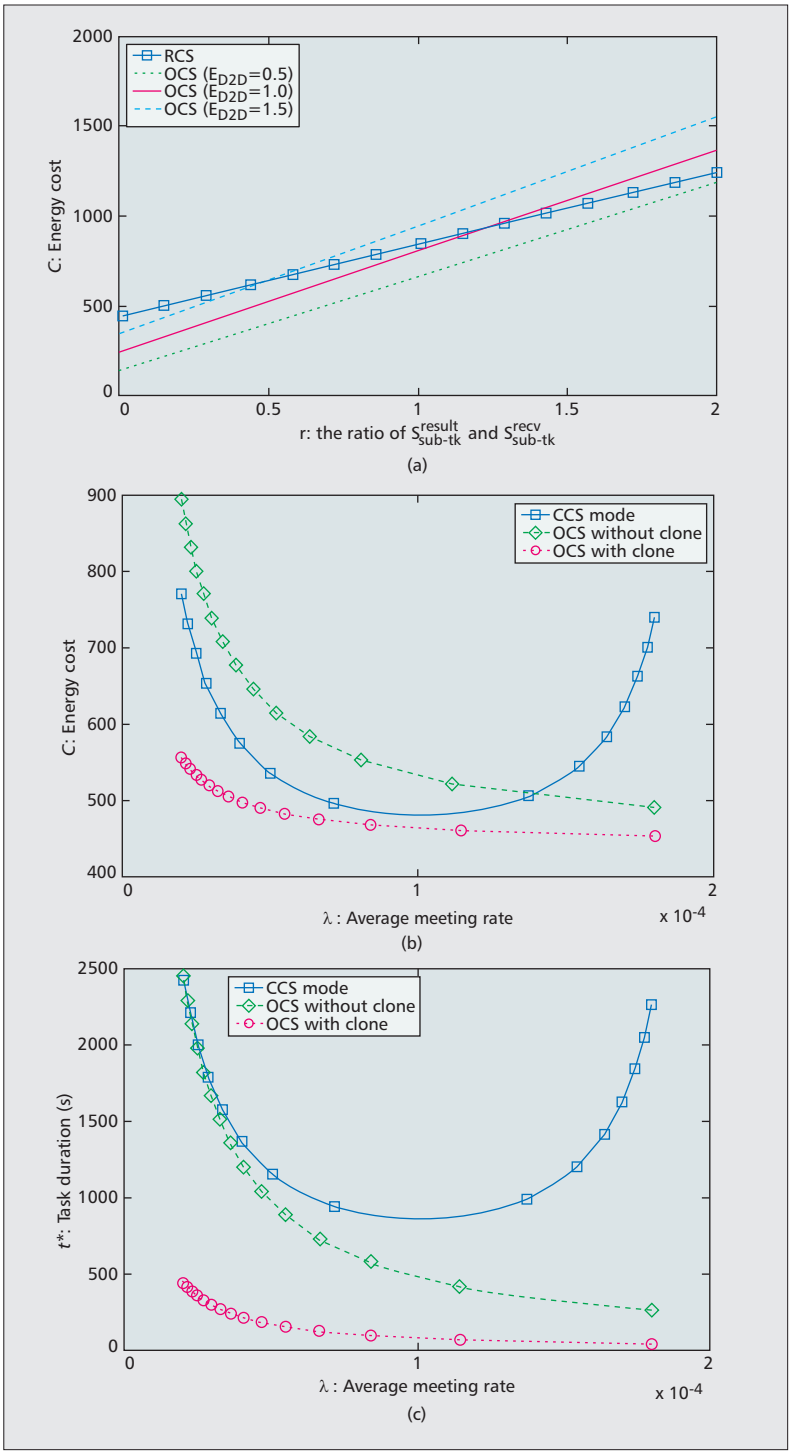


Figure 4. Performance evaluation among RCS, CCS and OCS: (a) comparison of energy cost between RCS and OCS with various r; (b) comparison of energy cost between CCS and OCS with various λ; (c) Comparison of task duration between CCS and OCS with various λ.

In the design spectrum, the OCS mode can be treated as an intermediate mode between CCS mode and RCS mode, thus yielding more flexibility and cost effectiveness to enable a more energy-efficient and intelligent strategy for computation offloading through the use of a cloudlet.

ly sensitive to delay, and the user can afford a higher cost to achieve good QoE by the use of 3G/4G RCS can be a good option.
- CCS mode: If the the computation node has a major concern in terms of communication cost while the movement of the service node is limited, CCS is a good choice.
- OCS mode: If the size of the computation result is much smaller than the size of the computation task, i.e., r is lower, OCS is more cost-effective while enabling maximum freedom for the computation node and the service nodes.

Conclusion

With an ever-increasing number of mobile devices and the resulting explosive growth in mobile traffic, the 5G networking system should be re-designed with more efficient resource utilization. One advanced technology to cope with the growing traffic and the associated computation demand is to offload computation intelligently. In this article we propose a novel service mode for cloudlet-assisted computing.

We call this new service mode "opportunistic ad hoc cloudlet service" (OCS). We categorize computation offloading into three modes: remote cloud service (RCS), connected ad hoc cloudlet service (CCS), and OCS. In the design spectrum, the OCS mode can be treated as an intermediate mode between CCS mode and RCS mode, thus yielding more flexibility and cost effectiveness to enable a more energy-efficient and intelligent strategy for computation offloading through the use of an ad hoc cloudlet. To the best of our knowledge, this article is the first to propose the OCS mode. In order to provide insights for facilitating the utilization of the newly proposed OCS mode, we build up a general and novel mathematical model, based on which optimal problems are formulated and solved.

Acknowledgement

This work is partially supported by the China National Natural Science Foundation under Grant 61300224 and Grant 61272397, the International Science and Technology Collaboration Program (2014DFT10070) funded by the China Ministry of Science and Technology (MOST), the Hubei Provincial Key Project under grant 2013CFA051, the Program for New Century Excellent Talents in University (NCET), and the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S20120011187.

References

[1] T. Taleb, "Towards Carrier Cloud: Potential, Challenges, and Solutions," *IEEE Wireless Commun.*, vol. 21, no. 3, June 2014, pp. 80–91.
[2] T. Taleb and A. Ksentini, "Follow Me Cloud: Interworking Federated Clouds and Distributed Mobile Networks," *IEEE Network*, vol. 27, no. 5, Oct. 2013, pp. 12–19.
[3] H. Flores and S. Srirama, "Mobile Code Offloading: Should It Be a Local Decision or Global Inference?" *Proc. ACM MobiSys*, June 2013, pp. 539–40.
[4] M. Valerio Barbera *et al.*, "To Offload or Not to Offload? The Bandwidth and Energy Costs for Mobile Cloud Computing," *Proc. IEEE INFOCOM*, Apr. 2013, pp. 1285–93.
[5] B. Han *et al.*, "Mobile Data Offloading Through Opportunistic Communications and Social Participation," *IEEE Trans. Mobile Computing*, vol. 11, no. 5, May 2012, pp. 821–34.
[6] X. Wang *et al.*, "TOSS: Traffic Offloading By Social Network Service-based Opportunistic Sharing in Mobile Social Networks," *Proc. IEEE INFOCOM*, Apr. 2014, pp. 2346–54.
[7] H. T. Dinh *et al.*, "A Survey of Mobile Cloud Computing: Architecture, Applications, and Approaches," *Wireless Communications and Mobile Computing*, vol. 13, no. 18, Oct. 2011, pp. 1587–611.
[8] L. Lei *et al.*, "Operator Controlled Device-to-Device Communications in LTE-Advanced Networks," *IEEE Wireless Commun.*, vol. 19, no. 3, June 2012, pp. 96–104.
[9] Y. Li and W. Wang, "Can Mobile Cloudlets Support Mobile Applications?" *Proc. IEEE INFOCOM*, Apr. 2014, pp. 1060–68.
[10] C. Wang, Y. Li, and D. Jin, "Mobility-Assisted Opportunistic Computation Offloading," *IEEE Commun. Lett.*, vol. 18, no. 10, Oct. 2014, pp. 1779–82.

Biographies

Min Chen [M'08, SM'09] (minchen@ieee.org) is a professor in School of Computer Science and Technology at Huazhong University of Science and Technology (HUST). He is the director of the Embedded and Pervasive Computing (EPIC) Lab. He was an assistant professor in the School of Computer Science and Engineering at Seoul National University (SNU) from September 2009 to February 2012. He worked as a post-doctoral fellow in the Department of Electrical and Computer Engineering at the University of British Columbia (UBC) for three years. Before joining UBC he was a post-doctoral fellow at SNU for one and half years. He has more than 180 paper publications. He received the Best Paper Award from IEEE ICC 2012, and the Best Paper Runner-up Award from QShine 2008.

Yixue Hao (yixue.epic@gmail.com) received the B.S. degree from Henan University , Kaifeng, China, in 2013 . He is currently a Ph.D. candidate in the Embedded and Pervasive Computing (EPIC) Lab led by Prof. Min Chen in the School of Computer Science and Technology at Huazhong University of Science and Technology (HUST). His research includes Internet of Things, body sensor networks, and mobile cloud computing.

Yong Li [M'09] (liyong07@tsinghua.edu.cn) received the B.S. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. From July to August 2012 and 2013 he was a visiting research associate with Telekom Innovation Laboratories and The Hong Kong University of Science and Technology, respectively. From December 2013 to March 2014 he was a visiting scientist with the University of Miami. He is currently a faculty member in the Department of Electronic Engineering, Tsinghua University. His research interests are in the areas of networking and communications.

Chin-Feng Lai [M'09, SM'14] (cinfon@ieee.org) has been an associate professor in the Department of Computer Science and Information Engineering, National Chung Cheng University since 2014. He received the Ph.D. degree from the Department of Engineering Science at National Cheng Kung University, Taiwan, in 2008. He received the Best Paper Award from IEEE EUC 2012. He has more than 100 paper publications. He is an associate editor-in-chief for the *Journal of Internet Technology*. His research focuses on Internet of Things, body sensor networks, E-healthcare, mobile cloud computing, cloud-assisted multimedia networks, and embedded systems, among other areas.

Di Wu [M'06] (wudi27@mail.sysu.edu.cn) is an associate professor and associate department head in the Department of Computer Science, Sun Yat-sen University, Guangzhou, China. He received the B.S. degree from the University of Science and Technology of China in 2000, the M.S. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2003, and the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong in 2007. From 2007 to 2009 he worked as a postdoctoral researcher in the Department of Computer Science and Engineering, Polytechnic Institute of NYU, advised by Prof. Keith W. Ross. He is the co-recipient of the IEEE INFOCOM 2009 Best Paper Award.

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

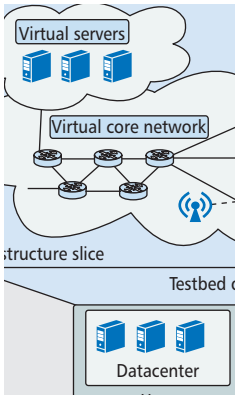
SERIES EDITORS

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

Software Defined and Virtualized Wireless Access in Future Wireless Networks: Scenarios and Standards

Future wireless networks are expected to provide augmented and data-intensive services in a multi-vendor multi-proprietor scenario. This scenario introduces relevant challenges to the networking infrastructure, especially in terms of flexibility and interoperability that could be addressed by extending the concept of Virtualization and Software Defined Networking to the wireless or wired-cum-wireless world.

Fabrizio Granelli, Anteneh A. Gebremariam, Muhammad Usman, Filippo Cugini, Veroniki Stamati, Marios Alitska, and Periklis Chatzimisios



Abstract

Future wireless networks are expected to provide augmented and data-intensive services in a multi-vendor multi-proprietor scenario. This scenario introduces relevant challenges to the networking infrastructure, especially in terms of flexibility and interoperability, that could be addressed by extending the concept of Virtualization and Software Defined Networking (SDN) to the wireless or wired-cum-wireless world. This paper provides a review of the perspectives to the extension of the SDN paradigm in the wireless domain by identifying current trends and proposed solutions, and providing the existing major standardization efforts and future trends in the field.

COMMUNICATIONS STANDARDS

Introduction

Wireless communications and networks nowadays are playing an integral role in how people interact and communicate, as well as how businesses operate and offer services to end customers. The increasing growth in the number of wireless devices and the introduction of a vast amount of applications, which are being used over the wireless access network, have led to an increasing demand for more bandwidth and have dictated the need for more powerful and faster networks.

With the introduction and enablement of unified communication services, wireless access networks have to be able to handle huge amounts of traffic nowadays, including data, voice, and video. There has been tremendous momentum in the advances in wireless technologies in recent years, due to the fact that they offer mobility and access to resources from almost anywhere. In support of these technologies, different standards have been introduced and coexist in order to serve different purposes and needs. The variety of these standards results in the creation of complex wireless heterogeneous networks, which has a negative effect on the way these networks operate and are being managed, increasing the overall complexity.

In order to deal with the aforementioned challenges, the focus of both the industry and the research community has shifted, in an effort to define and develop the next generation of wireless networks, namely 5G networks. 5G promises to support a massive number of simultaneously connected devices, high system spectral efficiency (data volume per area unit), low outage probability (better coverage), low latency, high versatility and scalability.

To achieve such goals, emerging technologies such as network virtualization and software defined networking (SDN) are being considered as technology enablers. These advancements are promising the introduction of programmability, flexibility, and elasticity for the managed networks, in order to better manage the high demand for enhanced IT resources and to satisfy customers' requests. Indeed, in wireless access networks, virtualization is expected to become an essential functionality, which will enable scaling and efficiency, resulting in easier network management in an effort to smooth the process of achieving interoperability and coexistence of different wireless technologies [1]. The need is to have a service oriented architecture (SOA) for convergence and a smooth transition between different wireless technologies, as it is currently being done in cloud computing and wired networks.

In the IT world, virtualization of resources (e.g. servers, compute, storage) has been prevalent and has changed the way IT services are being developed and offered. Following the same logic in recent years, the concept of network function virtualization (NFV) has been introduced, in order to enable the virtualization of network components. Thorough research has been conducted and various architecture definitions for virtualization of functions of the wireless access networks are being proposed. Various standards have been published on this issue, and the most relevant ones are discussed later.

In parallel with the research around NFV, in 2008 the OpenFlow protocol was introduced as part of university research and became the basis on which Open Networking Foundation (ONF) released the first version of the OpenFlow protocol in 2011, which enables the decoupling of the control plane from the data plane in networks. This led to an exponential increase in research and made possible the development of the concepts and technologies of SDN. Actually, SDN was originally oriented toward wired networks, but an increasing interest is driving the research of its application in wireless access networks as well, due to the inherit benefits offered by SDN approaches.

The current article provides an overview of the perspectives of using the software defined network paradigm at the service of the future wireless access networks, including both 5G mobile technologies and wireless local area networks (WLANs). The reported work is based on the activities developed under the framework of the recently established IEEE Standardization Research Group on Software Defined and Virtualized Wireless Access.[1]

Fabrizio Granelli, Anteneh A. Gebremariam, and Muhammad Usman are with the University of Trento.

Filippo Cugini is with CNIT.

Veroniki Stamati is with Sytel Reply.

Marios Alitska and Periklis Chatzimisios are with Alexander TEI of Thessaloniki.

[1] http://community.comsoc.org/groups/rg-software-defined-and-virtualized-wireless-access

0163-6804/15/$25.00 © 2015 IEEE

The article is organized as follows. We provide an overview of the future directions and opportunities in SDN/virtualized wireless access, while we address specific aspects related to the application of the SDN concepts to wireless access networks: enabling protocols and architectures for virtualization of wireless networks and end-to-end SDN in wired-wireless scenarios. We provide a list of current related standardization efforts as well as some real life examples and future trends. We then conclude the article.

Perspectives of Software Defined and Virtualized Wireless Access

The introduction of NFV for wireless access networks, complimented by the introduction of SDN technologies, offers tremendous opportunities and allows a number of benefits to be realized within the areas of deployment, operation, and management of wireless access networks [14, 15]. One of the main benefits is the decoupling of the network control and management function from data forwarding, which takes place in the hardware. Essential functions for the control and management of the network that previously had to be embedded in the hardware's ASIC, now can be deployed and developed in software, and by applying DevOps techniques these functions can be optimized further and faster. Furthermore, a number of network functions can now be implemented in the cloud and incorporated with other network access domains through the use of SDN. This flexible infrastructure reduces the dependency of emerging wireless technologies on hardware, enables better exploitation of the available infrastructure, and correspondingly shortens the research and development cycle of wireless technologies.

Based on the suggested architecture models and the various technologies of wireless access virtualization, three main perspectives of wireless access virtualization can be identified [1].

Flow Oriented Perspective: In this perspective, the wireless access domain can be defined as the data exchange and distribution network. This is the most common wireless access virtualization perspective that focuses on the management, scheduling, and service differentiation of different data flows from different slices. This perspective is commonly defined as mobile network virtualization. This can be implemented in two ways: either as an overlay over the wireless hardware, such as OpenRoads and virtual Base Transceiver System (vBTS), or it can be implemented as an internal scheduler inside the wireless hardware, such as Network Virtualization Substrate (NVS) and virtual Long Term Evolution (LTE).

Protocol Oriented Perspective: The protocol oriented perspective aims to isolate, customize, and manage the multiple wireless protocol instances on the same radio hardware. If the protocol processing is done purely in software for the all protocol layers, then software-based resources must be sliced. On the other hand, if the protocol processing is done purely on hardware, then the hardware resources must be sliced. In [2] a partial implementation of the protocol oriented perspective allows for the sharing of the same radio resources for different instances of the wireless protocol stack, while OpenRadio and Sora make the radio hardware fully customizable by introducing the full implementation of the protocol oriented perspective and allowing different protocols to operate on the same hardware.

Spectrum Oriented Perspective: In the spectrum oriented perspective, the resources to be sliced are radio frequency (RF) bands and raw spectrum. This perspective decouples the RF front end from the protocol, allowing multiple front ends to be used by a single node, or for a single RF front end to be used by multiple virtual wireless nodes. In this perspective the scheduling is done in a flow oriented approach, whereas the protocol oriented perspective can be overwritten by reshaping the signal.

Although the main idea and basis of virtualization is the same for both wired and wireless networks, the approaches used for the controllable medium (wired) have to be modified and adapted for the wireless medium and the time-varying characteristics of the mobile environment. In the related literature, different approaches have been proposed in order to have better control over wireless virtualized networks for a wide range of applications. We now summarize some of them.

Wireless Access Virtualization and Software Defined Networking: The main concepts, such as service awareness and function modularity, which have been introduced by SDN and virtualization in the wired network, can be extended to the wireless virtualized access. Furthermore, the fact that through SDN technologies one can achieve programmability, flexibility, and elasticity of the network makes SDN directly applicable to wireless networks in an effort to deal with the challenges of the increasing number of mobile devices. Overall, SDN and virtualization in wireless access networks can be considered as an extension of the wired network. One of the options is to consolidate the wireless functions in a centralized software controller, where the decoupling of a management and data plane is achieved by using a protocol such as CAPWAP (Control And Provisioning of Wireless Access Points). The Openflow extension of wireless access points is suggested in OpenRoads, where the data plane of the wireless access is virtualized through the use of FlowVisor. The configuration of the wireless access point is controlled using the Simple Network Management Protocol (SNMP).

Wireless Virtualization using SDRs: Software defined radio (SDR) offers the same functionalities for wireless networks as SDN does for wired networks. OpenRadio was proposed as an extension of SDN for the wireless domain, where baseband processing is separated into the processing plane and decision plane. The programmability of both planes increases the flexibility of hardware to be shared among different protocols.

WLAN Virtualization: IEEE 802.11 wireless LAN access points are virtualized by taking advantage of the existing functions of IEEE 802.11 WLAN (sleep state of power save mode

The introduction of NFV for wireless access networks, complimented by the introduction of SDN technologies, offers tremendous opportunities and allows a number of benefits to be realized within the areas of deployment, operation, and management of wireless access networks.

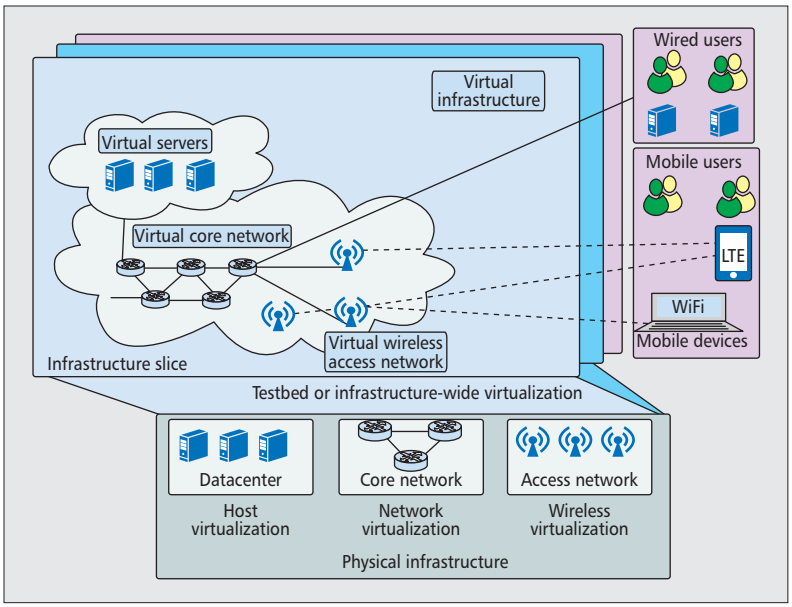


Figure 1. The extension of the concept of NaaS in the wireless domain.

(PSM)) to enable the network interface card (NIC) to communicate with different networks at the same time. Such heterogeneous networks are managed through a common mesh function management layer by using the interface management function and IEEE 802.21 extensions to the MAC abstraction layer. Another promising solution to WLAN virtualization is the decoupling of IEEE 802.11 MAC frames, which can be processed in the cloud, using the OpenFlow protocol. The MAC frames are processed in virtual access points (APs), which are present in the cloud. The technique is known as CloudMAC in the literature [3].

Cellular Base Station Virtualization: In the area of base station virtualization, each tenant can have its own scheduler over its slice. Different architecture models have been proposed in the literature to provide slice isolation, which can be based upon modifying the medium access technique. For WiMAX virtualization, a virtual base transceiver (vBTS) [4] and network virtualization substrate (NVS) [5] were proposed. Similarly, a modification in the MAC of LTE eNB is proposed in [6] for LTE virtualization to separate the traffic of different slices based on SLAs of each slice. In real life, this method is very prevalent in the telco on cloud applications, with some of the biggest global telco operators currently deploying these technologies in their network.

Wireless Spectrum Virtualization: Wireless resource virtualization can be performed below the physical layer using the spectrum virtualization layer (SVL) [7]. The SVL uses spectrum reshaping techniques to share the same RF front end on different portions of the spectrum.

In reality, different virtualized domains exist in the same geographical area and are interconnected to form the modern network infrastructure. These virtualized domains can be integrated with one another to form a cloud infrastructure. The logic of the domain results in the infrastructure sharing the same kind of resources and performing the same kind of functionalities. For example the wireless virtualization domain can be integrated with others. In this way, the wireless access network can be an extension of network as a service (NaaS), a concept borrowed from cloud computing. The concept is shown in Fig. 1, where an application or a service is no longer bounded to a domain or layer. Although both the industry and the research community have generally accepted the concept of the NaaS, a number of challenges and open problems still remain. Different domains are going to be managed by different controllers, and in a number of situations different protocols and standards would be applicable. This raises the question of how these different domains will be orchestrated and managed in a harmonized approach and view, offering to business users the end-to-end manageability that virtualization and SDN are promising.

Enabling SDN and Virtualization over Wireless Networks

In order to boost network capacity and efficiency in a self-automated manner while reducing CAPital EXpenditure (CAPEX) and OPerational EXpenditure (OPEX), the 3rd Generation Partnership Project (3GPP) introduced the concept of self organizing networks (SON). The main goal of SON is to make *planning*, *configuration*, and *optimization* of heterogeneous and mobile radio access networks simpler and faster in an automated manner with a minimum need of manual intervention. In [8], SON techniques are applied to multi-radio access technologies (RATs), targeting deployment optimization. Four different use-cases are described, depending on the deployments of the RATs, i.e. whether the communication services are transmitted from the same cell-site over *the same* or *different* antennas. This leads to different types of configurations and implementations of the SON objective functions and controllers (e.g. one centralized RAT objective functions and multiple controllers dedicated to each RAT, multiple RAT SON objective functions, and multiple dedicated controllers).

In a fairly similar goal to SON, SDN promises innovation in terms of network programmability by allowing network control and management whereby a high level of abstractions exist. Implementing SDN over wireless represents a challenging task, since it introduces many issues related to link isolation or channel estimation that are not necessary in wired networks. However, SDN is very promising in the wireless domain and carries great potential according to future perspectives, since it provides functions that could promote a better cooperation between access points/base stations in order to reduce interference or to enhance security.

The concept of network virtualization is defined as the process of combining hardware and software network resources and network functionality into a single, software-based entity called a *virtual network*. Network virtualization improves the resource utilization scheme by sharing the same hardware in a controlled and an isolated manner (i.e., each virtual network

believes it has its own hardware). To achieve these functionalities, a clear abstraction of the underlying hardware should be provided to the software-based entity. The virtualization of the underlying infrastructure (it could be of different RATs) allows *multiple service providers* to simultaneously control and configure the underlying infrastructure. Moreover, service evolutions as in the vision of future 5G networks could be achieved easily by gradually applying the changes in each network slice, which also mitigates the issues related to backward compatibility with the existing legacy infrastructure. Therefore, the concepts of SDN and virtualization will have a huge potential impact toward realizing future wireless network deployments.

In the following paragraphs we provide some frameworks and architectures that may be helpful for future SDN developments in the wireless domain.

One proposed approach to enable SDN over wireless is to integrate the SDN principles in wireless mesh networks (WMNs), formed by OpenFlow switches with one or multiple wireless interfaces, typically based on IEEE 802.11 protocols. A wireless mesh software defined network (wmSDN) can take advantage of OpenFlow and utilize the optimized link state routing (OLSR) protocol to route OpenFlow control and data traffic to avoid operating issues in the unwanted sudden scenario of controller unreachability. This wmSDN toolkit implementation is composed by a POX controller, Open vSwitch, OLSR daemon, Bash, and Python Scripts. Moreover, a wireless mesh router (WMR) is also needed to provide connectivity to the different access networks, as well as provide connectivity to the Internet and operate as a gateway. The WMR is connected to an OpenFlow controller through a wireless/wired connection interface. The SDN paradigm that is implemented by OpenFlow could foster WMNs as it provides simple management and flexibility. Wireless resource utilization can be enhanced by a central server that can perform processing actions on multiple levels of the protocol stack. Actually, wmSDN utilizes OLSR, Linux-based OpenFlow tools and scripts that can easily be developed in Linux-based wireless IP routers that operate in WMNs. A similar approach has been followed by the Clemson Openflow program, which has developed and deployed an outdoor mesh network on which GENI researchers can conduct experiments.

Figure 2 represents the concept of WMNs . As shown, the WMN includes several WMRs in order to provide Internet access to a group of APs, and a wireless or wired connection interface to end users as well. In this scenario, the SDN connections are represented by the dotted arrows from the OpenFlow controller in the top of the figure to each WMR. Such connections can be wired or wireless.

A very relevant scenario for 5G networks is the implementation of SDN for extremely dense wireless networks. In particular, SDN has been identified as a solution for this scenario since it tackles two key challenges of wireless dense networks: interference and mobility management.

The architecture proposed in the CROWD research project[2] offers the tools to orchestrate network elements in a way that intra-system interference is mitigated, channel opportunistic transmission/receipt techniques can be enabled, and energy efficiency can be boosted. An extremely dense and heterogeneous network composed of two domains of physical network elements, the reconfigurable backhaul and radio access network (RAN). The functionality of network optimization is entitled from the network elements to a set of controllers, which are virtual entities deployed dynamically over the physical devices (taking into account the actual network load and the capacity constraints). The control plane consists of two types of controllers: the CROWD regional controller (CRC), which is a logically centralized entity that executes long-term optimizations; and the CROWD local controller (CLC), which runs short-term optimizations. As energy efficiency is essential for operators as well as for environmental issues, CROWD is proposing a power cycling control application to dynamically reconfigure the network and the status of network nodes according to traffic demands. Ultimately, the CROWD project proposes control applications for networks consisting of both LTE and IEEE 802.11 devices (e.g., the offloading control application envisions the utilization of load balancing and relay techniques that span across multiple RATs and multiple technologies).

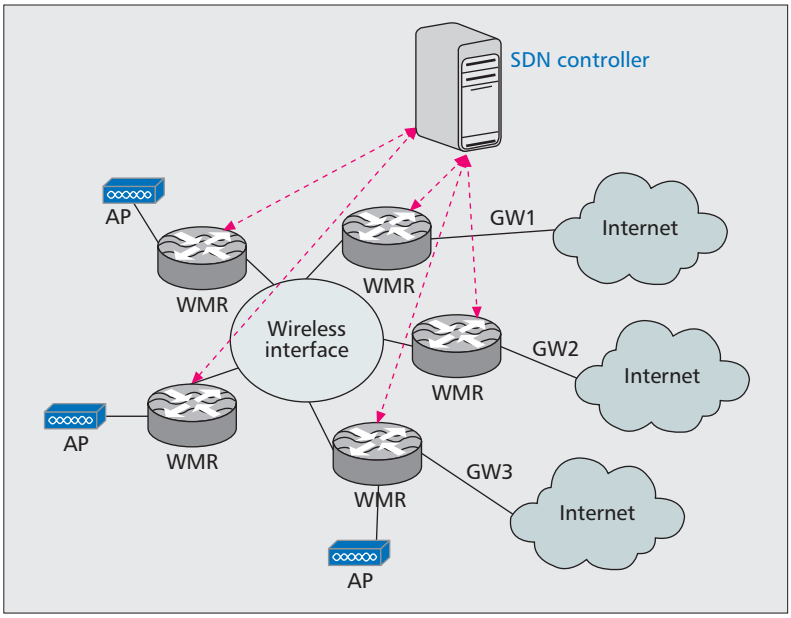


Figure 2. SDN over a wireless mesh network, where the SDN controller connects to WMRs.

Figure 3 represents the concept of the CROWD architecture under an extremely dense and heterogeneous network. As shown in the figure, there is a control plane, which consists of two types of controllers (CRC and CLC) with the help of OpenFlow and CAPWAP as well. The data plane consists of two heterogeneous network physical elements, the reconfigurable backhaul and the dense radio access network where the forwarding is being held.

Moreover, as mentioned earlier, OpenRadio

[2] http://www.ict-crowd.eu

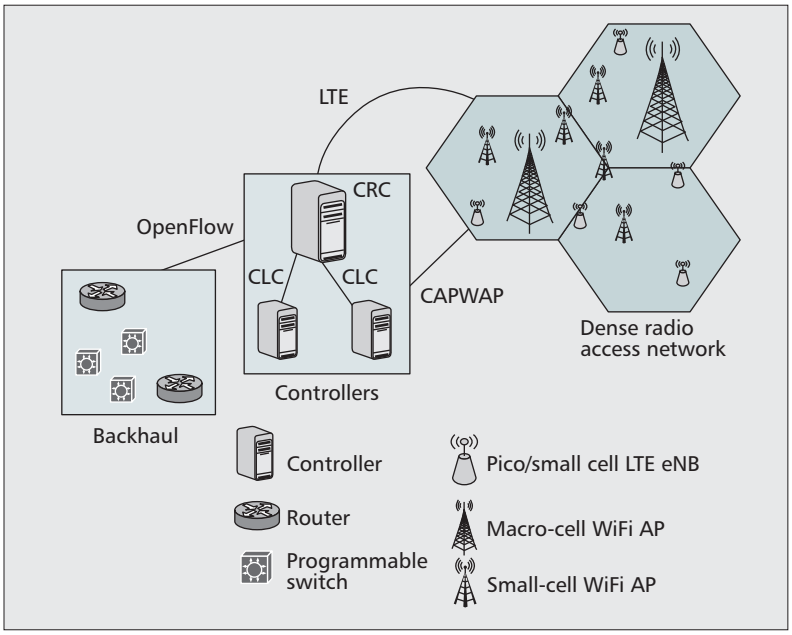


Figure 3. Overview of the CROWD network control architecture.

gives a modular and declarative programming interface by separating the wireless protocols into two planes, processing and decision planes, providing the right abstraction in-order to balance the trade-off between performance and flexibility. Following a similar concept, SoftRAN aims to build a centralized control plane for radio access networks (RANs) to address the issues related to radio resource allocation, interference reduction, handover, and load balancing. It achieves those goals by aggregating base stations as a virtual big-base station with a centralized control system.

The concept of a virtual cell (V-Cell) architecture is proposed in [9], aiming to overcome the technical limitations of *Layer 1* and *Layer 2* of the conventional wireless networks. In a similar analogy to SoftRAN, the V-Cell abstracts all the resources provided by a pool of base stations into a single large resource space to a centralized control-plane (i.e., the SDN RAN). In LTE, the resource space (also known as *Resource Pool*) is a *3-dimensional* (time, frequency, and space) matrix of the resource blocks (RBs). Furthermore, the authors introduce the concept of *no handover zone*, where the user equipment (UE) is assigned to different RBs from the centralized resource pool allowing the UE to jump from one base station to another without instantiating a handover procedure. In [10] the SDN approach is exploited by exposing the lower layers (i.e., PHY and MAC) of the LTE protocol stack to a centralized controller, such that it is possible to dynamically reconfigure the network by means of specifically designed algorithms.

Another approach, MultiFlow [11], aims at enhancing IP multicast over IEEE 802.11 networks. MultiFlow is based on the conversion of multicast transmission to a group of unicast transmissions. The MultiFlow implementation on SDN and the OpenFlow protocol can be carried out without the need for adopting any proprietary software or hardware. The programmable OpenFlow controller in the MultiFlow architecture is used to implement a multicast controller that will be responsible for detecting multicast packets. The use of OpenFlow further allows the efficient implementation of interfaces and mechanisms of a centralized WLAN controller protocol (e.g., CAPWAP) as part of the multicast controller. The estimation of the network wireless conditions is the main functionality of the controller. MultiFlow can provide an enhancement of the channel availability as well as a more efficient handling of the multicast load (this significantly improves the transmission of high definition multimedia over IEEE 802.11 WLANs). Moreover, when MultiFlow is utilized, the delay performance of the system is improved (especially when the AP is already overloaded) and enhanced power saving can be achieved.

Figure 4a represents an architectural network approach that implements the MultiFlow protocol. This scenario includes a set of multicast servers, a multicast router, a MultiFlow controller, and a typical network scenario with connected AP's to a switch. The MultiFlow controller is composed of an OpenFlow controller in order to create a multicast controller, which operates the multicast to multiple unicast conversion as shown in the Fig. 4b.

A very specific scenario is represented in the case of deploying SDN over wireless sensor networks (WSNs). The main weaknesses of WSNs are related to resource limitations, such as processing power, memory, energy, and communication capabilities. Those weak points may be addressed by smart management of network resources through SDN. Under WSN's scenarios, SDN can reach a higher potential since it provides functions that can allow a better collaboration between the base station and the forwarding nodes. The deployment of SDN can be useful for issues such as energy saving, sensor node mobility, network management, localization accuracy, and topology discovery. The proposed framework in [12] considers a WSN that includes a base station and a number of sensor nodes. In this framework, sensor nodes do not have to make routing decisions. Actually, they forward or drop packets according to a set of pre-installed rules stored in a special data-structure (known as the flow table) maintained at every sensor node, and the best routes are calculated by the controller according to specific matching criteria. Under this framework, the controller makes use of location information gathered by any localization technique for identifying the best routes. The controller then transforms these decisions to a set of rules that are to be imported into the flow tables of each node. The controller does not necessarily need be a stand alone node; it can also be implemented as a part of the base station. The controller architecture of the base station can address several issues in the management area such as mobility and localization, and simply provides reconfiguration abilities. This improves management features such as energy saving and topology discovery. The controller (base station) node maintains its functionality through the following five layers: physical, medium access control, network operating system, middleware, and application. For

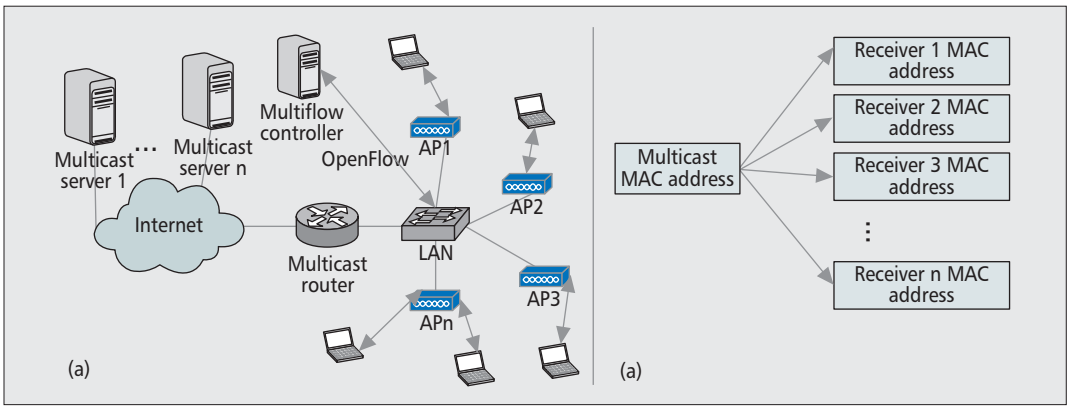


Figure 4. Architectural concept of a network that operates with a) MultiFlow; b) multicast to multiple unicast conversion.

certain applications a WSN needs to be deployed to cover a certain area and needs to maintain continuous monitoring. Using the proposed architecture, system behavior could be changed in an easier and more efficient way. This could be achieved by modifying the forwarding rules in the base station (controller), and thus new action policies would be sent to the network devices. Furthermore, maintaining the system's scalability becomes a simpler task. In order to extend the network to another area, new devices could be linked to the same WSN (same base station) that should allow for network policies to be sent to the new devices.

Figure 5 depicts a possible software defined wireless sensor network framework. There is no need for sensor nodes to make routing decisions. Decisions are made centrally at the base station and new rules are allocated to sensor nodes, e.g. add, remove, or update flow. From the software defined perspective inside the base station there is a logically-centralized controller because the control logic is implemented as part of the base station.

Despite the fact that the concepts of network virtualization are extensively explored in the wired domain, it cannot be directly applied to the wireless domain. For example, the bandwidth achieved by a particular virtual wireless network segment from a given amount of radio resources varies depending on the channel quality of the users. When a certain bandwidth is reserved for a particular network segment (slice), the dynamic nature of the air-interface should be properly considered. In addition, another limiting factor (i.e. interference) needs to be taken care off as well. Wireless network virtualization is in its very early stage and attracting a huge interest of research.

End-to-End SDN in a Wired-Wireless Scenario

As mentioned previously, one potential use case of SDN and virtualization is the enablement of end-to-end connectivity between wired and wireless networks. Two main networking scenarios have been considered for SDN-based wired-wireless integration and are described in this section.

The first scenario refers to future 5G mobile systems. Effective solutions for high-rate radio transmissions will have to be combined with advanced management functionalities, enabling a fully integrated solution between both the wireless and the wired part. The wireless network side will experience increased traffic volumes, higher data transmissions rates, and the emergence of new services based on cloud applications. This translates into the need to have an integrated, flexible, and programmable backhaul/fronthaul segment able to guarantee the necessary adaptability to service requirements and traffic conditions. In order to accomplish a programmable backhaul and to ensure smooth interoperation between the fronthaul and core layers, the industry is suggesting the use of SDN capabilities, to separate the bearer and control functions and to centrally manage and automatically configure the cell site gateway (CSG) and the small cell site gateway (SCSG) on the aggregation site gateway (ASG).

The second scenario where SDN can be used from an end-to-end perspective refers to wireless LAN systems. In LANs, wireless is becoming the primary access method. Also in this case, enhancements in network throughput have to be combined with offering better agility and flexibility, aiming at providing the same responsiveness and SLA of wired connections.

Overall, it is apparent that SDN represents the most suitable candidate technology to provide combined management of the wireless and wired segment, toward achieving an end-to-end network deployment and provisioning. Provided that an integrated SDN architecture will be adopted, a number of benefits are expected, as detailed below.

Unified Management of the Wired and Wireless Network. A single SDN orchestrator, equipped with different and technologically-specific southbound interfaces, is expected to provide a common view and control of the wired-wireless network. SDN solutions for mobile/backhaul/fronthaul access segments are expected to simplify network operations, lower total cost of ownership, and introduce manual-free operations. Similarly, SDN orchestration in Wi-Fi and wired LANs is expected to simplify operations

It is apparent that SDN represents the most suitable candidate technology to provide combined management of the wireless and wired segment, toward achieving an end-to-end network deployment and provisioning. Provided that an integrated SDN architecture will be adopted, a number of benefits are expected.

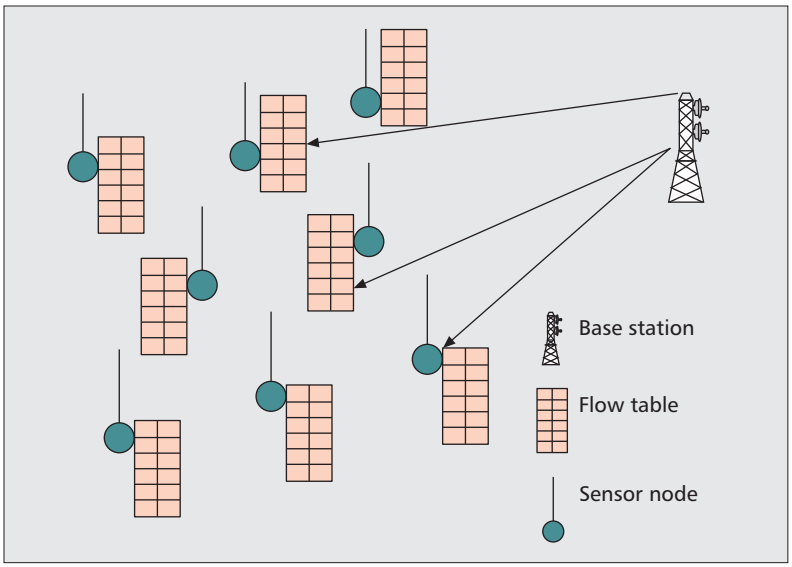


Figure 5. The software defined wireless sensor network framework.

and management functionalities that nowadays are still handled separately.

Unified Policy Enforcement. A common management of wireless and wired technologies will simplify IT policy enforcement. In particular, policies can be defined and enforced only once and applied across the whole network. Such functionality, already present in some advanced management tools, is becoming the standard approach integrated within SDN solutions that are using the group-based policy model.

Network Programmability and Network Function Virtualization. With a common SDN orchestration, the integration between the wired and wireless segments will enable more effective adaptation and virtualization strategies as well as the capability to dynamically react, in a coordinated way, to application and business needs, achieving the concept of offering the network as a service (NaaS).

Performance Improvement. The advanced programmability enabled by a common SDN architecture has the potential to improve overall network performance. For example, network throughput can be improved when users are located in overlapped service areas by enabling advanced programmability of migration and handoff strategies. Moreover, download rates can be increased by activating multiple parallel streams originated in the wired network and delivered, in a coordinated and synchronized way, by the wireless network. In addition, a common SDN orchestration could implement advanced power saving solutions (e.g., traffic migrations and sleep configurations) during off-peak traffic conditions.

Vendor Interoperability. The open standard and open source implementations of southbound SDN interfaces will significantly facilitate interoperability among different vendor devices. However, due to the different solutions and approaches that various vendors are offering, interoperability between them still remains an open issue.

Customized Applications. Standard northbound and open APIs will provide support to new SDN applications, enabling customized network behavior through the SDN controller.

To meet such high expectations, SDN has to guarantee open programmable access to the wireless infrastructure, adopting controller modules, abstraction layers, and enhanced northbound/southbound interfaces, able to be fully integrated within the open SDN-based solutions designed and operated in wired networks.

Related Standardization Efforts

This section provides an overview of ongoing standardization efforts in the framework of SDN/network softwarization that also include wireless networks and devices. It must be emphasized that currently most standards on SDN are technology- or data-transport-agnostic, therefore they specify interfaces and management approaches without specifically addressing the wireless/mobile scenario.

The International Telecommunications Union — Telecommunications Standardization Sector (ITU-T) is actively involved in the framework of standardization for future networks. Relevant standards by ITU-T are aimed at SDN in future networks, as in the case of ITU-T Recommendation Y.3300 (2014) — Framework of Software-Defined Networking. However, the document describes the framework of SDN by providing definitions, objectives, capabilities, and architecture at a high level. Wireless networks are included in the overall SDN deployment picture, but not explicitly addressed in the document.

The European Telecommunications Standards Institute (ETSI) is involved in standardization efforts with the publication of the framework of the Industry Specification Group for Network Function Virtualization (ETSI NFV SIG). The current version of document ETSI GS NFV-INF 001 addresses wireless (and specifically mobile base stations) as a possible domain for virtualization, and specifies standard interfaces and use cases, without addressing how virtualization should be performed.

The Internet Engineering Task Force (IETF) is also actively developing RFCs for standards on SDN and network virtualization. Most relevant documents in the framework of SDN standardization in mobile networks are related to the concept of Service Function Chaining (SFC) and especially to the SFC Architecture (draft-ietf-sfc-architecture-01) and SFC Use Cases in Mobile Networks (draft-ietf-sfc-use-case-mobility-01). The documents describe an architecture and related use cases for usage of SFC, i.e. a carrier-grade process for continuous delivery of services based on network function associations in mobile networks (3GPP as a reference).

The ONF explicitly addresses the scenario of SDN in wireless networks in the white paper "OpenFlow-Enabled Mobile and Wireless Networks." The proposed use cases include intercell interference management and mobile traffic management, outlining benefits in terms of additional flexibility in a 4G multi-vendor scenario as well as improved granularity in resource management.

IEEE is active in such a scenario, with the participation to the IEEE SDN Initiative and the definition of Standardization Working Groups and Research Groups on Virtualization in Wireless Networks. Relevant activity is being carried out in the framework of the Research Group on Software Defined and Virtualized Wireless Access and the Research Group on SDN/NFV — Structured Abstractions.

Future Research Trends and Real Life Scenarios

The evolution of wireless networks to 5G will change consumers' habits in using the Internet. The efficient management of frequency spectrum and bandwidth in 5G technology will face inevitable challenges in the future. Software defined wireless networking represents a valuable choice for integration between frequency spectrum and bandwidth, between suppliers and consumers, QoS and security. Both SDN and SDR have the capability to reconfigure, allowing network administrators to move forward to a self-adaptive environment by collecting signals and changing parameters at the packet level and quickly finding a suitable communication path and frequency band.

The key open issues to be addressed can be summarized as wireless network abstraction, programmability, security, quick reconfigurability, mobility, and orchestration. Some of the recent research topics and real life scenarios are discussed below which require further exploration to meet future needs.

SDN Performance in Dense Mobile Networks

Mobile networks tend to be more dense and large scale to meet the future needs of increased bandwidth and better QoS. Some of those issues are addressed by SoftCell [13] and CROWD [8], but the corresponding performance is not clear in all scenarios. An SDN enabled cross-layer MIMO solution could be necessary to meet the future bandwidth needs.

Internet of Things (IoT) and SDN

SDN, with its ability to intelligently route Internet traffic and efficiently use network resources, will make it easy to eliminate bottlenecks and efficiently process the data generated by IoT without placing a strain on the network. SDN capabilities of service changing, bandwidth calendaring, and dynamic load management will be particularly useful for IoT.

SDN Based Mobile Data Offloading

The rapid growth of wireless networks has created increased demand for mobile data services. The problem of energy consumption has also become more significant for mobile devices where battery time is a crucial factor. The need is to separate the computationally intensive and memory intensive applications and offload them to nearby servers using software defined networking, enabling programmable offloading policies that take into account real time network conditions and the status of devices and applications.

SDN Orchestrator

The employment of SDN in wireless networks raises the question of how these different domains will be orchestrated and managed in a harmonized approach and view, offering to business users the end-to-end manageability that virtualization and SDN are promising. This represents a huge challenge nowadays, because such a single orchestrator does not exist and efforts toward a common approach have not been successful due to the existence of different SDN solutions.

Conclusions

Software defined networking represents a promising paradigm in both the wireless and the wireless-cum-wired scenarios. As discussed in this article, targeted effort is being allocated to extend the benefits of virtualization and softwarization to the wireless domain. Such features make SDN over wireless a relevant technology to manage scenarios including multi-vendor and multi-owner setups, such as those envisaged in the framework of the current discussion on 5G and future wireless networks.

This article discussed standardization efforts on how to extend SDN to the wireless sections of the end-to-end path as well as how to control and manage wireless resources. Indeed, harmonization of current efforts will be useful to enable interoperability and seamless access to the wireless infrastructures of the future.

Targeted effort is being allocated to extend the benefits of virtualization and softwarization to the wireless domain. Such features make SDN over wireless a relevant technology to manage scenarios including multi-vendor and multi-owner setups, such as those envisaged in the framework of the current discussion on 5G and future wireless networks.

References

[1] H. Wen, P. K. Tiwary, and T. Le-Ngoc, "Current Trends and Perspectives in Wireless Virtualization," *Int'l. Conf. Selected Topics in Mobile and Wireless Networking (MoWNeT)*, Montreal, QC, Canada, Aug. 19–21, 2013, pp. 62–67.
[2] Y. Al-Hazmi and H. de Meer, "Virtualization of 802.11 interfaces for Wireless Mesh Networks," *Proc. 8th Int'l. Conf. Wireless On-Demand Networks and Services (WONS'11)*, Bardonecchia, Italy, Jan. 26–28, 2011, pp. 44–51.
[3] P. Dely *et al.*, "CloudMAC — An OpenFlow based Architecture for 802.11 MAC Layer Processing in the Cloud," *Proc. 2012 IEEE Globecom Workshops*, Anaheim, CA, USA, Dec. 3–7, 2012, pp. 186–91.
[4] G. Bhanage *et al.*, "Virtual Basestation: Architecture for an Open Shared WiMAX Framework," *Proc. 2nd ACM SIGCOMM Wksp. Virtualized Infrastructure Systems and Architectures (VISA'10)*, New Delhi, India, Aug. 30–Sept. 3, 2010, pp. 1–8.
[5] R. Kokku *et al.*, "NVS: A Virtualization Substrate for WiMAX Networks," *Proc. 16th Annual Int'l. Conf. Mobile Computing and Networking (ACM MobiCom'10)*, Chicago, Illinois, USA, 2010, pp. 233–44.
[6] Y. Zaki *et al.*, "A Novel LTE Wireless Virtualization Framework," *Mobile Networks and Management*, vol. 68, 2011, pp. 245–57.
[7] K. Tan *et al.*, "Enabling Flexible Spectrum Access with Spectrum Virtualization," *2012 IEEE Int'l. Symp. Dynamic Spectrum Access Networks (DYSPAN)*, Bellevue, WA, USA, Oct. 16–19, 2012, pp. 47–58.
[8] A. Oliva *et al.*, "Denser Networks for the Future Internet, the CROWD Approach," *Mobile Networks and Management*, vol. 58, 2013, pp. 28–41.
[9] R. Riggio *et al.*, "V-Cell: Going Beyond the Cell Abstraction in 5G Mobile Networks," *2014 IEEE SDNMO Wksp. Network Operations and Management Symp. (NOMS)*, Krakow, Poland, May 5–9, 2014, pp. 1–5.
[10] A. A. Gebremariam *et al.*, "A Framework for Interference Control in Software-Defined Mobile Radio Networks," *IEEE CCNC 1st Int'l. Wksp. Vehic. Networking and Intelligent Transportation Systems (VENITS)*, Las Vegas, NV, USA, Jan. 2015, pp. 853–58.
[11] S. Tajik and A. Rostami, "MultiFlow: Enhancing IP Multicast over IEEE 802.11 WLAN," *Proc. IFIP Wireless Days (WD)*, Nov. 13–15, 2013, pp. 1–8.
[12] A. De Gante, M. Aslan, and A. Matrawy, "Smart Wireless Sensor Network Management Based on Software-Defined Networking," *Proc. 27th Biennial Symp. Commun. (QBSC)*, June 2014, pp. 71–75.
[13] X Jin *et al.*, "SoftCell: Taking Control of Cellular Core Networks," http://arxiv.org/abs/1305.3568.
[14] T. Taleb, "Towards Carrier Cloud: Potential, Challenges, & Solutions," *IEEE Wireless Commun.*, vol. 21, no. 3, June 2014. pp. 80–91.
[15] T. Taleb and A. Ksentini, "Follow Me Cloud: Interworking Federated Clouds & Distributed Mobile Networks," *IEEE Network*, vol. 27, no. 5, Sept./Oct. 2013. pp. 12–19.

Biographies

Anteneh A. Gebremariam (anteneh.gebremariam@unitn.it) received the B.Sc. degree in electrical and computer engineering from Addis Ababa University, Addis Ababa, Ethiopia, in 2007, and the MSc. degree in telecommunication engineering from the University of Trento, Trento, Italy in 2013. Currently he is a Ph.D. candidate at the University of Trento. His research focuses on abstraction, virtualization, and efficient resource utilization for wireless communication based on the SDN paradigm. He is collaborating with the CREATE-NET (headquartered in Trento, Italy) research group on activities for software-defined mobile networks. In 2013 (1 March to 31 October) he did his master's thesis at Nokia Siemens Networks (Munich, Germany) titled "SON: Tilt based Optimization on LTE-Advanced Networks." From 2007 to 2011 he worked as a GSM-BSS engineer at ZTE Corporation (Chinese multinational telecommunications equipment and systems company). He is a member of the Ethiopian Society of Electrical Engineers (ESEE) and the IEEE Communication Society.

Fabrizio Granelli (fabrizio.granelli@unitn.it) is an IEEE ComSoc Distinguished Lecturer for the period 2012–15, and an associate professor in the Dept. of Information Engineering and Computer Science (DISI) at the University of Trento (Italy). He received the Laurea (M.Sc.) and Ph.D. degrees from the University of Genoa, Italy, in 1997 and 2001, respectively. He spent six months as a visiting professor at the State University of Campinas (Brasil). He has authored or co-authored more than 150 papers on topics related to networking, with a focus on wireless communications and networks, cognitive radios and networks, green networking, and smart grid communications. He is the founder and general vice-chair of the First International Conference on Wireless Internet (WICON'05) and general chair of the 11th, 15th, and 18th IEEE Workshop on Computer-Aided Modeling, Analysis, and Design of Communication Links and Networks (CAMAD). He is TPC co-chair of the IEEE GLOBECOM Symposium on "Communications QoS, Reliability and Performance Modeling" in the years 2007, 2008, 2009, and 2012.

Muhammad Usman (muhammad.usman@unitn.it) received his B.E. degree in electronics engineering from the School of Electrical Engineering and Computer Science, National University of Science and Technology, Pakistan, the M.S. degree in telecommunication engineering from the University of Trento, Italy, and an M.S. in computer networks from the Sant'Anna School of Advanced Studies, Pisa, Italy. He is currently a student in the Ph.D. program at the University of Trento, working in the green networking research group. His research interests include software defined networks, cloud computing, and device to device communication in 5G networks.

Filippo Cugini (filippo.cugini@cnit.it) received the M.S. degree in telecommunication engineering from the University of Parma, Italy. Since 2001 he has been with the National Laboratory of Photonic Networks, CNIT, Pisa, Italy. His main research interests include theoretical and experimental studies in the field of optical communications and networking. He has co-authored 12 international patents and more than 150 international publications.

Veroniki Stamati (v.stamati@replyltd.co.uk) has been working in the UK since 2011. At the beginning of her career she had been working alongside a variety of industries consulting in data protection and privacy. For the past three years she has been specializing as a senior consultant in the telco industry, helping clients to realize the benefits of cloud, SDN and NFV, identifying specific use cases, conducting proof of concepts, and defining their strategy. She is currently the SDN/NFV competency lead within Sytel Reply UK. She studied informatics at Aristotle University of Thessaloniki in Greece and obtained her masters in computing, IT law, and management from King's College London.

Marios Alitska (alitska@it.teithe.gr) is currently a fourth year B.Sc. student in the Department of Informatics at Alexander Technological Educational Institute of Thessaloniki. His research focused on wireless communication networks at the CSSN Research Lab, working with associate professor Periklis Chatzimisios. He is interested in LTE and WiMAX wireless networks, social media strategies, and information retrieval.

Periklis Chatzimisios [SM IEEE] (peris@it.teithe.gr) serves as an associate professor at the Alexander TEI of Thessaloniki (Greece). Recently he has been a visiting academic/researcher at the University of Toronto (Canada) and Massachusetts Institute of Technology (USA). He is involved in several standardization activities, serving as a member of the Standards Development Board for the IEEE Communication Society (ComSoc) (2010–today), as the secretary of the IEEE 1907.1 Standardization Working Group, and lately as an active member of the IEEE Research Groups on IoT Communications & Networking Infrastructure and on Software Defined & Virtualized Wireless Access. He holds editorial board positions for several IEEE/non-IEEE journals and he is the director (co-director during 2012–2014) for the E-letter of the IEEE Technical Committee on Multimedia Communications (MMTC). He is the author/editor of eight books and more than 85 peer-reviewed papers on the topics of performance evaluation and standardization activities of mobile/wireless communications, with more than 1300 citations received by other researchers. He received his Ph.D. from Bournemouth University (UK) (2005) and his B.Sc. from Alexander TEI of Thessaloniki, Greece (2000).

Call for Papers
IEEE Communications Magazine
Wireless Communications, Networking, and Positioning with Unmanned Aerial Vehicles

Background

Enabled by the advances in computing, communication, and sensing as well as the miniaturization of devices, unmanned aerial vehicles (UAVs) such as balloons, quadcopters, and gliders, have been receiving significant attention in the research community. Indeed, UAVs have become an integral component in several critical applications such as border surveillance, disaster monitoring, traffic monitoring, remote sensing, and the transportation of goods, medicine, and first-aid. More recently, new possibilities for commercial applications and public service for UAVs have begun to emerge, with the potential to dramatically change the way in which we lead our daily lives. For instance, in 2013, Amazon announced a research and development initiative focused on its next-generation Prime Air delivery service. The goal of this service is to deliver packages into customers' hands in 30 minutes or less using small UAVs, each with a payload of several pounds. 2014 has been a pivotal year that has witnessed an unprecedented proliferation of personal drones, such as the Phantom and Inspire from DJI, AR Drone and Bebop Drone from Parrot, and IRIS Drone from 3D Robotics.

Among the many technical challenges accompanying the aforementioned applications, leveraging the use of UAVs for delivering broadband connectivity plays a central role in next generation communication systems. Facebook and Google announced in 2014 that they will use a network of drones which circle in the stratosphere over specific population centers to deliver broadband connectivity. Such solar-powered drones are capable of flying several years without refueling. UAVs have also been proposed as an effective solution for delivering broadband data rates in emergency situations through low-altitude platforms. For example, the ABSOLUTE, ANCHORS, and AVIGLE projects in Europe have been investigating the use of aerial base stations to establish opportunistic links and ad-hoc radio coverage during unexpected and temporary events. They can serve as a temporary, dynamic, and agile infrastructure for enabling broadband communications, and quickly localizing victims in case of disaster scenarios.

This proposed Feature Topic (FT) issue will gather articles from a wide range of perspectives in different industrial and research communities. The primary FT goals are to advance the understanding of the challenges faced in UAV communications, networking, and positioning over the next decade, and provide further awareness in the communications and networking communities on these challenges, thus fostering future research. Original research papers are to be solicited in topics including, but not limited to, the following themes on communications, networking, and positioning with UAVs.

- Existing and future communication architectures and technologies for small UAVs
- Delay-tolerant networking for cooperative UAV operations
- Design and evaluation of wireless UAV test beds, prototypes, and platforms
- Multi-hop and device-to-device communications with UAVs
- Interfaces and cross-platform communication for UAVs
- QoS mechanisms and performance evaluation for UAV networks
- Game-theoretic and control-theoretic mechanisms for UAV communications
- Use of civilian networks for small UAV communications
- Integrating 4G and 5G wireless technologies into UAV communications, such as millimeter wave communications, beamforming, moving networks, and machine type communications
- Use of UAVs for public safety and emergency communications, networking, and positioning
- Integration of software defined radio and cognitive radio techniques with UAVs
- Channel propagation measurements and modeling for UAV communication channels

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 (pre-ferred) or MSWORD formatted paper via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to Author Center. Follow the instructions there. Select "May 2016 / Wireless Communications, Networking and Positioning with UAVs" as the Feature Topic category for your submission.

Schedule for Submissions

- Submission Deadline: November 1, 2015
- Notification Due Date: January 15, 2016
- Final Version Due Date: March 1, 2016
- Feature Topic Publication Date: May 2016

Guest Editors

Ismail Guvenc
Florida International Univ., USA
iguvenc@fiu.edu

Walid Saad
Virgina Tech, USA
walids@vt.edu

Mehdi Bennis
Univ. of Oulu, Finland
bennis@ee.oulu.fi

Christian Wietfeld
TU Dortmund Univ., Germany
christian.wietfeld@tu-dortmund.de

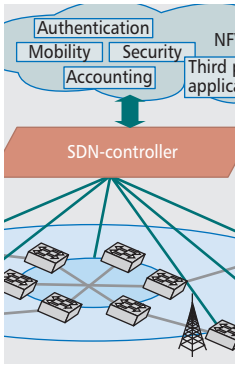
Ming Ding
NICTA, Australia
ming.ding@nicta.com.au

Lee Pike
Galois, Inc., USA
leepike@galois.com

Cellular Software Defined Networking: A Framework

The authors propose a cellular network architecture called CSDN (Cellular SDN), which is based on Software Defined Networking and Network Functions Virtualization. This architecture enables network operators to simplify network management and control. It also enables the creation of new services, in a flexible, open, and programmable manner.

Abbas Bradai, Kamal Singh, Toufik Ahmed, and Tinku Rasheed



Abstract

Today's mobile customers desire to remain connected anywhere, at any time, and using any device. This phenomenon has encouraged mobile network operators to build complex network architectures by incorporating new features and extensions, which are harder to manage and operate. In this article we propose a novel and simplified architecture for mobile networks. The proposed architecture, which we call CSDN (Cellular SDN), leverages software defined networking (SDN) and network functions virtualization (NFV). SDN abstracts the network and separates the control plane from the data plane; NFV decouples logical network functions from the underlying hardware, for dynamic resource orchestration. Furthermore, we argue that dynamic resource orchestration and optimal control need real-time context data analyses to make intelligent decisions. Thus, in the proposed architecture we exploit the capability of the mobile edge networks to gather information related to the network as well as the users. This information can be used to optimize network utilization and application performance, and to enhance the user experience. In addition, the gathered data can be shared with third party service providers, enabling the realization of innovative services.

COMMUNICATIONS STANDARDS

Introduction

Data traffic in mobile networks has recently witnessed an explosive growth with the increasing penetration of smartphones. In addition, users' interests have evolved from voice and short message texts (SMS) to high quality real-time audiovisual content consumption and production. This evolution has pushed 3G networks to their limits, and have motivated network operators to adopt the Long Term Evolution (LTE) network, which is also known as the fourth generation (4G) mobile network. Now the industry is creating the roadmap toward 5G, the fifth generation of mobile network standards.

The LTE architecture proposed by the 3rd Generation Partnership Project (3GPP) [1] is considered an all-IP network that uses an orthogonal frequency division multiplexing (OFDM) air interface. Figure 1 illustrates the main components of the 3GPP LTE architecture. It is composed of Evolved Universal Terrestrial Radio Access Network (E-UTRAN), containing a set of base stations (e-NodeBs), connected to each other via X2 interfaces. The e-NodeBs connect the user equipment (UE) to the LTE core network, called the evolved packet core (EPC). The EPC is an overlay network that uses IP and Ethernet based packet switched communication.

In an LTE network, mobility management schemes are used to ensure connectivity and to keep the IP addresses of users unchanged, even when users move and change their network. The serving gateway (S-GW) serves as a mobility management anchor and, thus, has to maintain a high number of states related to the mobile users. S-GW is then connected to the packet data network gateway (P-GW), which links the network to the Internet and other data networks. Additionally, the P-GW performs several functions such as monitoring, billing, access control, and enforcement of varied policies. Note that each device in this architecture uses specialized hardware and software to implement varied functionalities, which in turn increase the complexity as well as the cost of the network.

The LTE architecture has been widely adopted by dozens of mobile service providers (MSP) around the world. However, with the emergence of new technologies and services such as cloud computing and content delivery networks (CDNs), the actual structure of the LTE network has become an obstacle to future evolution. The centralized data plane functionalities, such as QoS, access control, and monitoring features, introduce scalability issues at the P-GW. Indeed, in an LTE network, all the traffic should pass through the P-GW, even if the communication is between the UEs of the same cell. This makes the P-GW the hot element of the network [2]. With such a configuration, strategies like caching of the popular content in the mobile network find limited use because all flows have to pass through the P-GW. Additionally, the EPC elements are specialized components with standardized interfaces, where each component achieves a specific task and each interface has a unique definition.

In this context, the introduction of new network functionalities takes a long time from the standardizing process to implementation and market entry. This long process deters network operators from innovating and investing in new network services. Furthermore, the specialized network components, defined exclusively for EPC, suffer from inflexibility and lack of openness. Each EPC element is controlled through standardized interfaces, and cannot be controlled by open interfaces or through application programming interfaces (APIs). In case the network operator plans to introduce new network services or needs to adopt new functionalities, the existing network components become useless, or of little use, to implement the envisaged functionalities.

Abbas Bradai is with CNRS Grenoble Informatics Laboratory.

Kamal Singh is with Laboratoire Hubert Curien/Université de Saint-Etienne.

Toufik Ahmed is with University of Bordeaux.

Tinku Rasheed is with Create-Net Research Center.

 0163-6804/15/$25.00 © 2015 IEEE

Cellular Software Defined Network: Overview

In order to overcome the challenges introduced in the previous section, we propose a cellular network architecture called CSDN (Cellular SDN), which is based on Software Defined Networking (SDN) [3] and Network Functions Virtualization (NFV) [4]. This architecture enables network operators to simplify network management and control. It also enables the creation of new services, in a flexible, open, and programmable manner.

Figure 2 provides an overview of the proposed architecture. CSDN leverages the benefits of SDN and NFV for dynamic resource management and intelligent service orchestration. The NFV platform is a cloud-based radio access network (C-RAN) [5] which allows the orchestration of resources using virtualization techniques. It uses different virtual machines running multitudes of applications over the cloud infrastructure of an MSP. An application within this framework, for instance MME (mobility management element) can not only run on virtual machines, but can also adapt its capabilities in an elastic way depending on the network load. This makes it very easy and flexible for network dimensioning and resource allocation.

Additionally, the proposed architecture leverages the SDN concepts for the orchestration of intelligent services. Here, by intelligent services we refer to the services that allow an MSP to implement subscriber policy, profile-aware service provisioning at the level of individual flows, etc. However, to realize these intelligent services, the MSP needs to take decisions based on data analytics. This requires a data component similar to a user data repository (UDR) [5]. We extend this data component to create the context data repository (CDR), which includes additional information such as network data, user profiles, and usage data.

In our CSDN architecture, mobile operators gather data related to the network as well as related to users. Network data includes traffic load, bandwidth availability, wireless channel information, and network health information such as network points suffering from congestion or other problems. This data is combined with subscriber data related to usage, quality of experience (QoE), as well as subscriber profiles including user behavior and preferences. Combining network data with user data allows the MSP to connect network-centric data with the user-centric data, allowing for intelligent resource allocation and provisioning, while considering the network conditions in real time. With this process, the MSP can optimize the user experience, reducing user churn, while minimizing network resource utilization at the same time. It also enables easier network evolution for better performance to support new technology. In the following, we provide a brief overview of SDN and NFV concepts, and then we illustrate the features and benefits of our CSDN architecture.

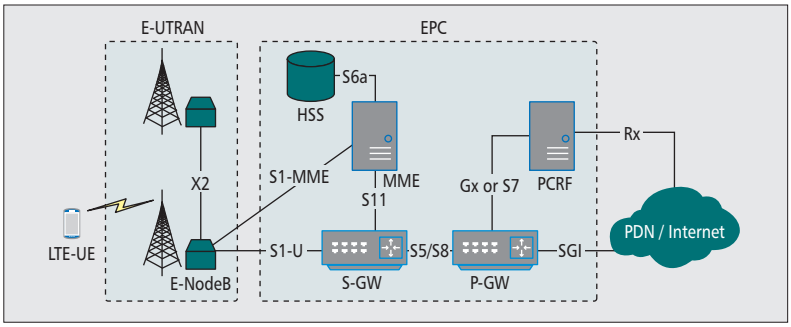


Figure 1. LTE network architecture.

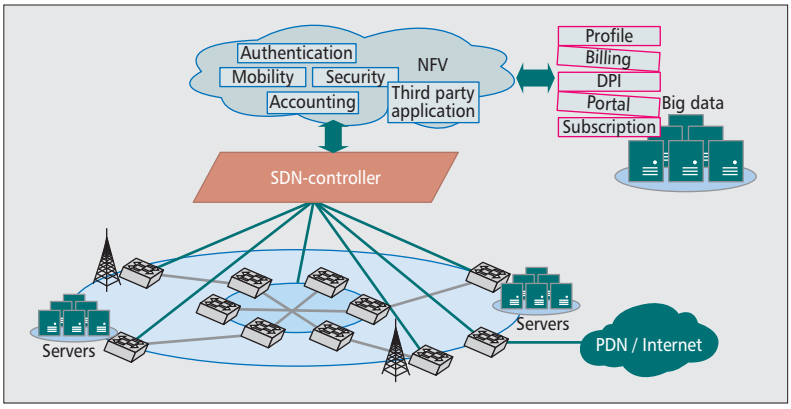


Figure 2. CSDN overview.

SDN and NFV for Intelligent Services Orchestration and Dynamic Resources Management

SDN (software defined networking) is clearly climbing the technological hype cycle every day as a novel approach that separates the control plane from the data plane (Fig. 3a). The control plane is logically centralized and controls the data forwarding elements of the network using an open interface. SDN originated from the need to solve the problems in managing fast-evolving Telco networks that require cumbersome configuration measures of several network equipments, which is a complex task. The key to solving such problems is through the automation of the control through application programming interfaces, and enabling virtualization of network functions by hiding the detailed configuration process from network control. SDN provides a flexible and centralized way to configure network equipment where network applications can be built on top of the SDN control plane for intelligent management of the network. These management applications can take optimal decisions, which in turn can be automatically compiled into network configuration rules. The latter are simple rules based on packet headers that are handled by the SDN controller to communicate to the SDN-enabled switches that implement them. This centralized approach significantly reduces the effort required to configure a network, and at the same time reduces the occurrence of problems such as downtime due to misconfigurations or failures.

Together, SDN and NFV can enable MSPs to make their network services dynamic, which will allow them to optimize their network resources, increase the agility of network, implement novel services, and hasten the process from service design to service production.

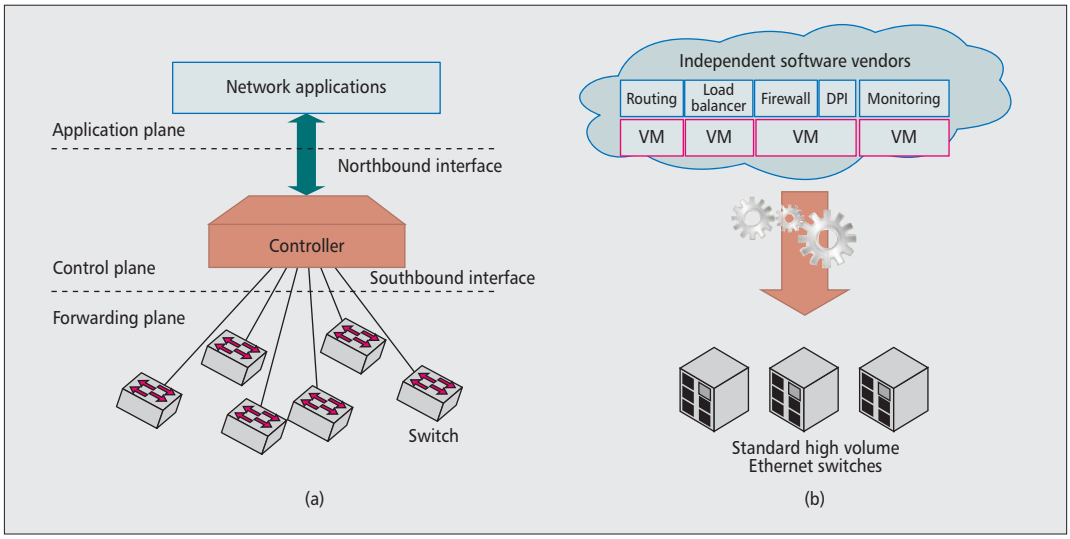


Figure 3. Architecture of SDN and NFV: a) SDN architecture; b) network function virtualization.

While SDN allows the network layers to be programmable by separating the data plane from the control plane, NFV technology provides the capability of flexible networking service placement. The concept of NFV, as defined by the European Telecommunications Standard Institute (ETSI) committee [4], is to decouple the network function from the hardware, as shown in Fig. 3b. The decoupled network functions are implemented using software. NFV combines virtualization and cloud computing techniques and applies them to telecommunication networks. It virtualizes the networking functions referred to as virtualized network functions (VNF). These network functions are then interconnected or chained to create networking services. The VNFs are deployed over one or more virtual machines and use generic hardware, such as off-the-shelf servers, instead of vendor-specific hardware. The use of generic hardware leads to significant cost reductions. Additionally, NFV can benefit from the centralized control and network layer programmability offered by SDN. Together, SDN and NFV can enable MSPs to make their network services dynamic, which will allow them to optimize their network resources, increase the agility of network, implement novel services, and hasten the process from service design to service production. Further, this approach enables easy upgrading and network capacity expansion to support resource sharing between multiple-tenants and support for multi-cell collaborative signal processing.

Framework for Software Defined Cellular Network Architecture

While several SDN based architectures and solutions have been proposed [3], most of these works concern wired networks. Adapting SDN technology to the context of mobile networks is challenging due to mobile network specific issues such as handling user mobility, management of the radio resources, and addressing resource scarcity in the wireless environment. Indeed, scalability problems related to the mobile SDN need to be addressed such as the update of many fine-grained rules related to the traffic management of a high number of mobile users. The network should be able to keep multitudes of states required for mobility management, monitor flows, detect if the user traffic exceeds its preassigned quota, perform billing, assure QoS, congestion control, and optimize resource utilization. In the literature there are few works addressing SDN applied to wireless and mobile networks [7, 8]. While the authors in [7] propose a high level SDN-based architecture for future mobile networks, it lacks details and requirements for the real deployment of such an architecture. While the authors in [8] focus mainly on radio virtualization to provide effective resource virtualization, the approach can compromise overall system performance.

Figure 4 provides a detailed overview of the proposed cellular SDN architecture (CSDN). CSDN follows the SDN layering architecture and comprises the following layers: forwarding layer, control layer, and the network application layer. We expand this model by considering an additional functional layer called the knowledge layer that allows the MSP to gain insights into the intelligent vision of its network and users environment. In the following, we detail the most important aspect of the proposed architecture with a focus on mobile network-specific components.

LTE Virtualized Functions

Figure 5 illustrates the mobile network application layer of CSDN. The majority of the LTE EPC's functional elements are implemented in a centralized cloud-based infrastructure at the application level of CSDN. These virtualized network functionalities interact at the management and control level with the CSDN switches via the controller. In addition to these basic network functionalities, our architecture allows new applications or virtual functions to be implemented and instantiated, such as video adaptation and optimization, which can be easily

introduced using the northbound interface of the controller.

This flexibility makes it possible to build network applications that are not only 3GPP compliant, but also implement innovative schemes from third party providers, such as location based services, Internet of Things (IoT) applications, and optimized content distribution and content caching. In Fig. 5 we show the network applications complying with the 3GPP standard. The V_e-NodeB with the CSDN switch corresponds to the e-NodeB functionalities, the same for V_S-GW, V_P-GW, V_MME, and V_PCRF that correspond with their switches to the S-GW, P-GW, MME, and PCRF respectively. Due to the lack of space, we provide details for the functionalities of V_e-NodeB and V_MME below.

Radio Resource Management (V_e-NodeB): Radio resource management (RRM) is a big part of a mobile network and several RRM functionalities are present in LTE e-NodeBs. The base stations use distributed protocols for the allocation of shared radio resources and participate in the management of sessions, handovers, interference, admission control, etc. However, distributed protocols are sub-optimal as compared to centralized optimization schemes having a global vision. Thus, following SDN philosophy, we decouple the control plane from the radio equipment. The radio equipment is controlled from a centralized control plane, and this centralization makes it easier to perform radio resource allocation as well as backhauling. For example the controller can perform load balancing between several base stations by moving users from a congested base station to another base station, cooperative MIMO techniques can be employed to enhance the signal quality, and lightly loaded base stations can be put in sleep mode to reduce energy consumption.

Mobility Management (V_MME): Mobility management is another important issue in cellular networks. Mobility management can be implemented in CSDN conforming to the 3GPP standard as an application on top of the CSDN controller. This becomes possible thanks to the northbound interface. In Fig. 5a, we show the V_MME as a part of the virtual EPC (V_EPC) which respects the 3GPP standard. While the V_e-NodeB manages the radio resources and ensures physical mobility, i.e., the handover, the V_MME ensures the MME functionalities. The difference between CSDN and normal mobile networks is that the signaling and tunneling functionality of mobile networks, such as GTP tunneling, is compiled into CSDN packet flow rules. The controller sends these rules to the CSDN switches. Thus, the networking policies and functionalities related to QoS, metering, tunneling, and routing are compiled into packet flow rules, which in turn are used to configure the CSDN switches.

Design Consideration at the Forwarding Plane

Openflow is an open SDN standard that allows communication between the control plane and the network switches [9]. The switches defined by Openflow already provide functionalities such as

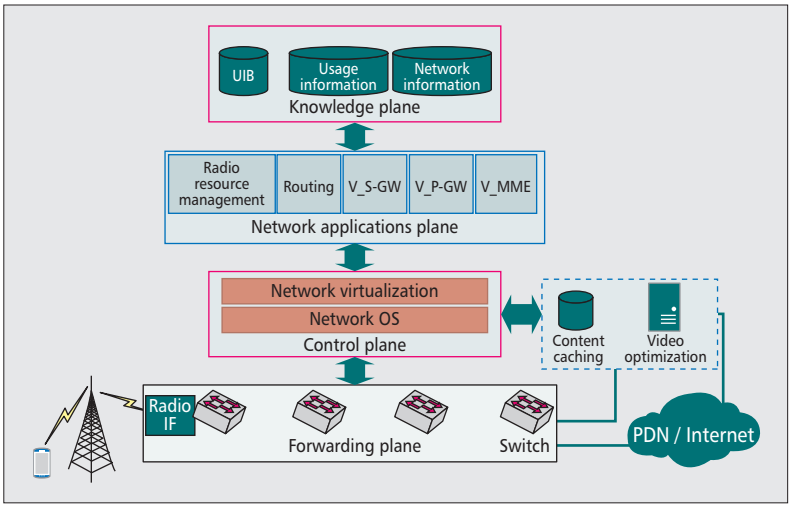


Figure 4. Cellular software defined network.

network traffic measurement and management to network operators, allowing measurement of network traffic, subscriber's usage statistics, perform billing, assess the QoS and then flexibly change the path of a user's flow to optimize as well as enforce QoS network policies. However, applying current Openflow based SDN architecture remains insufficient for mobile networks. This is due to the peculiarities of mobile networks that suffer from scalability issues, essentially due to an increasing number of users, their mobility, the fine grain access control, QoS policies, their nature of being a wide area network, and the scarcity of the radio resources. Centralized management of mobile networks addressing these issues makes the SDN controller a bottleneck. In addition, the frequent controller solicitation introduces additional delay, which can impact QoS. For these reasons, the local features in a CSDN network switch should be increased to discharge some functionalities from the controller. In particular, the switch should be able to execute a local program for performing simple tasks under the supervision of the controller, such as notifying the controller if the traffic exceeds a certain threshold, tagging some packets to be redirected to a transcoder, etc. There needs to be more in-depth investigation of this balance of delegating controller functionalities to CSDN switches and at the same time keeping these switches as simple as possible.

In addition, new switch capabilities will be beneficial to software-defined cellular networks. Today, TCP/UDP port numbers are no longer a reliable way to identify applications. Instead, support for deep packet inspection (DPI) would enable finer-grain classification of the applications, such as Web, peer-to-peer, video, and VoIP traffic. This is important to divide traffic into separate traffic classes for differential packet scheduling and routing policies, as commonly done in today's cellular networks. DPI will also help intrusion detection and prevention systems that analyze packet contents to identify malicious traffic. The DPI functionality could be enabled only on some switches, and applied only for some packets for better scalability and performance of the switches. In addition to DPI,

One of the main advantages of the SDN paradigm is the fact that it provides an API' s interface to easily develop new applications and services. The controller receives policies from the applications and provides them a virtual view of the network.

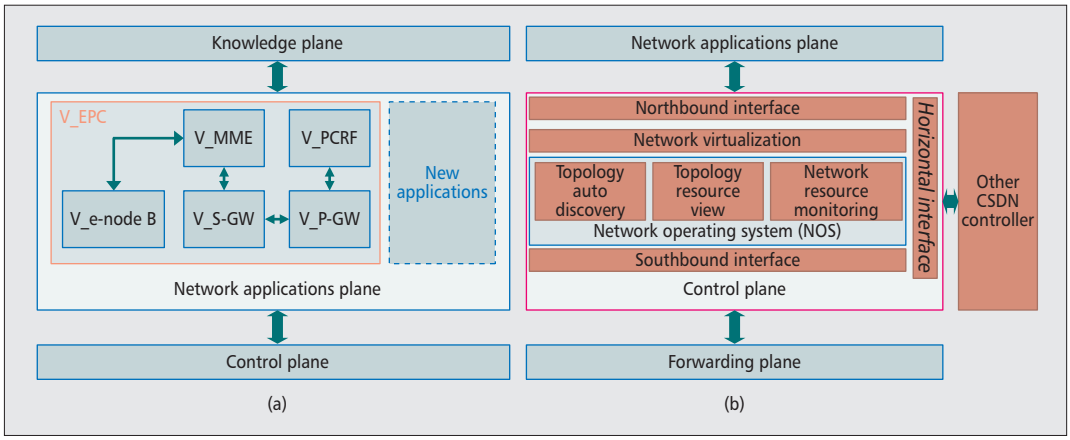


Figure 5. CSDN architecture details: a) network applications plane; and b) control plane.

some CSDN switches should perform other tasks such as header compression, mainly for small payload packets, such as VoIP packets. These switches could be deployed in low bandwidth regions of the network.

Design Consideration at the Control Plane

The CSDN controller (Fig. 5b) consists of a network operating system (NOS), a network virtualization block, and three communication interfaces. The core component of the controller is the NOS, whose main feature is to abstract the distributed state of the network and provide a global view of the network. The NOS is composed of three main building blocks: network resource monitoring, topology auto-discovery, and topology resource view. The second level of abstraction is performed by the *network virtualization* building block, which provides an abstract view of the network. This view in turn is exploited by the applications.

The three communication interfaces are the northbound, the southbound, and the horizontal interfaces (east-west). The *southbound interface* controls the CSDN switches via simple forwarding rules. It receives the network measurements and in some cases it also receives packets from the switches. In the case where a packet does not match with any rule in the forwarding table, the packet is sent to the controller. The controller analyses the packet with respect to the policies of the application layer and a new forwarding rule is pushed to the switch. The *horizontal interface* allows a controller to communicate with other CSDN controllers and to support third party interactions. Indeed, as shown in [3], a single controller for a large network introduces scalability issues. In CSDN, we propose a controller for each public land mobile network (PLMN). For this purpose coordination is needed between the different controllers in order to get a global view of the network and to coordinate actions in case of inter-PLMN communications. Finally, the *northbound interface* allows communication with the network applications. One of the main advantages of the SDN paradigm is the fact that it provides an API's interface to easily develop new applications and services. The controller receives policies from the applications and provides them a virtual view of the network.

Context Data for Network Optimization and Innovative Applications

Mobile operators gather information related to the mobile access network such as ongoing traffic, bandwidth availability, points of congestion, etc. The big data analysis techniques using network information data combined with the subscriber information data related to their profile and behavior can provide valuable insights to the network operator for network optimization and innovative personalized applications [10, 11]. Based on the network conditions, channel information, user profile, and device properties among many other data that could be considered, the MSP can leverage connection prioritization, audio/video transcoding, and quality selection to enable multi-screen customization and adaptability as an innovative service.

The main advantage of using the gathered data in CSDN is that the central decision making entity can gather a global view of the network. Boosting this central entity with big data tools allows it to have a global and insightful view of subscriber behavior, situational awareness, and the service evolution. Furthermore, MSPs are many steps ahead, as compared to digital service providers, in terms of the breadth and depth of data collection as well as data quality and reliability. The MSPs' role as the connection provider grants them access to individual-level information continuously and accurately, given that they have 'around-the-clock' connection to each and every subscriber, regardless of their location. The MSPs can share their data with third-party partners, allowing them to rapidly deploy innovative applications and services for mobile subscribers, enterprises, and other vertical segments. For example, this information enables enterprises to better understand local markets, such as local commercial areas, popular products and companies, and user behavior in order to serve their customers via just-in-time location-based services based on the gathered information.

Proximity, context, agility, and speed can be translated into value and can be exploited by the MSPs, service and content providers, over the top (OTT) players, and independent software

vendors (ISVs), enabling them to play complementary and profitable roles within their respective business models, and allowing them to monetize the mobile broadband experience and create an energized ecosystem. Based on innovation and business value, this value chain will allow all players to benefit from greater cooperation. To this end, a standardized, open architecture should be designed toward efficient and agile integration of such applications across multi-vendor mobile network platforms. In the following sections we present the different components and interfaces of the context data plane in our CSDN architecture, and then we explore the main obstacles and challenges to overcome.

Before we go through the details, it is essential to identify the different data gathered by mobile operators. The data gathered by mobile operators can be categorized into the following three classes of information, grouped into what we call the context data repository (CDR):

User Information Base: Mobile operators have accurate subscriber information. They have the true identity of the subscriber as well as some basic information about them such as the address for user identification and addressing, which corresponds to the international mobile subscriber identity (IMSI) and the mobile subscriber ISDN number (MSISDN) or telephone number. All this information is stored in the User Information Base (UIB). The latter also contains the user profile information, which concerns the service subscription status and user-subscribed QoS information such as maximum allowed bit rate or allowed traffic class. In addition, UIB can give an idea about the subscriber's economic status via device type, subscription type, and social activity via call minutes, frequency dial number, and exchanged messages.

Network Information Base: Mobile operators have the ability to gather fine grain information about any user, including real time information about the user's location as well as their activity and navigation information, in addition to the device information and connection characteristics. This information can be gathered thanks to the signaling through the base station, femtocells, and switches. The gathered information, once analyzed, can be converted into high level patterns of users' activity, consumption, and lifestyle, whereas wireless channel information and scheduling information (QoS, channel quality indicator, load) can be used to enhance network performance.

Usage Information: Using DPI and other related technologies, the network operators are also able to extract information about the subscriber's consumed content and applications used to learn about their browsing activities, preferences, and centers of interest. Network operators can also have access to the subscriber's economic information through their online shopping and e-banking activities. This information allows the operators to decipher subscribers' social and economic ecosystems, including their brand choices.

In addition to having real-time information, MSPs also have historical data that is captured continuously over long periods of time. This is possible thanks to their subscription management systems, billing and charging management systems, portals, and customer service systems, which are configured to record this information over the lifetime of each subscription. MSPs can use this historical data to conduct trends analysis and understand long term usage patterns.

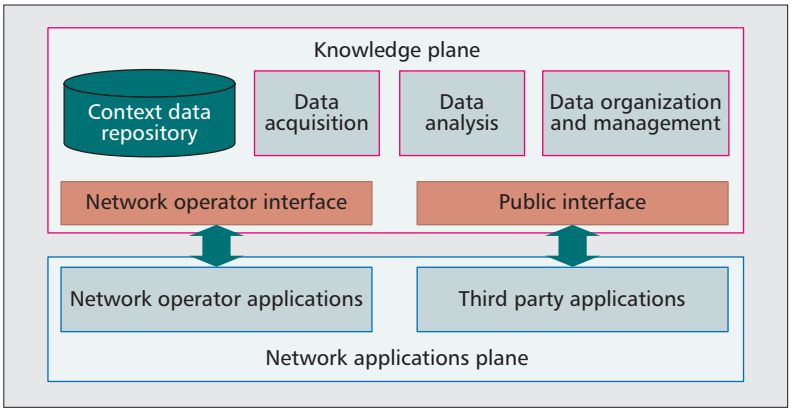


Figure 6. Knowledge plane.

Knowledge Plane in CSDN Framework

Current SDN forwarding devices match basic packet headers with the forwarding table rules to perform actions such as routing, packet header modification, packet dropping, etc. In order to support more advanced network applications, the CSDN maintains and updates the user, network, and usage data. The NOS can then "compile" these policies expressed in terms of user information into forwarding rules executed by the switches. In Fig. 6 we provide an overview of the proposed knowledge plane in our CSDN framework. The knowledge plane is composed of three functional blocks and two interfaces. The data acquisition block allows gathering of information either from the network, through a network application running in the application layer of CSDN, or through other information sources such as the network operator portal. After the acquisition of data, the next step is to analyze the data to provide valuable insights to the network operator and enable a comprehensive analysis of this data in terms of subscribers' network conditions, navigation patterns, lifestyles, events, and the movement of people. This is realized by the data analysis block. Finally, the data organization and management block enables organizing the raw data to facilitate its exploitation. The knowledge plane uses a bidirectional communication interface to communicate with the MSP applications that have privileged access to all the data. However, third party applications have limited access to the data, through the public interface, in order to preserve subscribers' privacy and deny public access to sensitive data.

Practical Considerations for Context Data in CSDN

Once we have defined the information that can be gathered by the MSPs, we need to identify the challenges and practical limits of contextual data for mobile SDN. In the following paragraphs we explore the main issues, the obstacles and the challenges to overcome in order to move toward big-data centric cellular SDN networks.

The ONF is a key organization that is standardizing SDN technologies. It produced the OpenFlow standard that is used in the current article. The OpenFlow standard specifies the OpenFlow switch and the OpenFlow protocol that is used by the SDN controller to communicate with the OpenFlow switches.

There are a number of practical considerations to be addressed before moving toward a dynamic, flexible, and centralized mobile network infrastructure that exploits multiple data sources [12].

Centralized vs. Distributed Collection of Data and Data Sampling Granularity: In order to collect data about the subscriber and their environment, the mobile operator exploits traditional methods such as the mobile operator offices and call centers. However, the subscribers are more and more reluctant to provide information about themselves. Thus, network operators try to gather user information within their network (using DPI for example), from the subscriber's devices or from service portals. However, using techniques such as DPI faces many regulatory and technical challenges. Governments are more and more restrictive about their citizens' privacy. Using DPI to collect information about subscribers could be regulated. For example, the type of information, their anonymization, and their storage are strictly regulated. In addition, gathering huge and real time information distributed over different network devices is technically challenging and can reduce network performance. The challenge is then to design scalable and failure-resilient data gathering approaches. Once gathered, the placement of the data storage is an important challenge, mainly to identify if a centralized or distributed solution is the most appropriate choice.

In addition, most data analysis tools use sampling in their process. In the case of real-time network adaptation, fine grain sampling is required. This will impact network performance and will create additional complexity. Furthermore, in the case of non-real-time network sampling, there needs to be the right balance in choosing the frequency of sampling in order to avoid the data becoming obsolete.

Data Processing: Once the network operator collects the data, the next challenge is to process this huge amount of data. Such data is multisource, heterogeneous, real-time, voluminous, and continuous, as well as static, ever-expanding, and having spatio-temporal dimensions. Streaming data analysis and knowledge extraction techniques are required to process this data coming from heterogeneous sources in order to convert it into actionable knowledge. The data processing needs to detect correlations in the data and discover patterns or abnormalities in dynamically evolving situations. Regarding the real-time aspects, we argue that the processing method depends on the real-time needs of the target application. Thus, the collected data should be classified into real-time data and non-real-time data. For instance, data related to user position should be processed in near real time, while data related to user preferences could be processed in a non-real-time manner.

Advanced Decision Making and Data Antiquity: Historical data can be combined with advanced models for the purpose of forecasting and advanced decision making. The objective is to forecast problems in the network and take preventive measures. The open issues are to design these forecasting models and advanced decision making tools. Note that network operators store historical information about the user in order to build a global image about their environment, behavior, and preferences in order to forecast requirements and enhance their experience. The amount of data and the time for which the respective data needs to be stored depends on different parameters, including the respect of the privacy jurisdiction, the utility degree of the data, and performance impact of the data quantity on the processing process.

User Privacy and Data Security Issues: The security of the gathered data is another issue. Mechanisms should be implemented to guarantee a satisfactory level of data safety. Therefore, data anonymization techniques should be considered to respect user privacy.

Relevance to Standardization Bodies and Fora

The Open Network Foundation (ONF)[1] is a key organization that is standardizing SDN technologies. It produced the OpenFlow standard that is used in the current article. The OpenFlow standard specifies the OpenFlow switch and the OpenFlow protocol that is used by the SDN controller to communicate with the OpenFlow switches. ONF has several working groups, and one of the groups related to mobile networks is the Wireless and Mobile Working Group (WMWG). This working group collects use cases and determines the architectural and protocol requirements for extending ONF technologies, such as the OpenFlow standard, for mobile and wireless networks. It has already drafted many use cases related to SDN utilization for network management and control, mobility management, flexible and scalable packet core, and virtualization for mobile networks.

The European Telecommunications Standards Institute (ETSI)[2] has an industry standardization group dedicated to Network Function Virtualization (NFV)[3] technology. At the completion of its Phase I, this working group published several NFV specifications in October 2013. Many of these specifications are related to defining the requirements and architecture for the virtualization of network functions, management and orchestration of virtual network appliances, and requirements and gap analysis for future technical specifications. In addition, the ETSI NFV working group provides a framework called proof of concept (PoC) for open demonstrations of NFV concepts. The PoC framework helps in building awareness and confidence in NFV technologies. It also provides feedback about challenges related to the implementations of NFV concepts and their interoperability. The ETSI NFV working group is now entering Phase 2 and its objectives are to ensure end-to-end inter-working of equipment and services. One of the objectives of ETSI NFV Phase 2 is also to clarify how NFV intersects with SDN technology and standards, which is one of the topics of the current article.

The Internet Engineering Task Force (IETF)[4] has several years of working experience on concepts related to SDN. One such example is the existing working group called Forwarding and Control Element Separation (FORCES). Some recent working groups have also emerged, such as

[1] https://www.opennetworking.org/

[2] http://www.etsi.org

[3] http://www.etsi.org/technologies-clusters/technologies/nfv

[4] https://www.ietf.org

Interface to the Routing System (I2RS), whose objective is to define the interfaces for the control of the routers and routing protocols. Moreover, more new working groups are being initiated such as Virtual Network Function Pools (VNF pools) to provide support mechanisms for the reliability of a group of virtual network functions, and the working group called Abstraction and Control of Transport Networks (ACTN) that targets network partitioning and resource slicing for clients, and also network automation and orchestration. Additionally, the Internet Research Task Force (IRTF) has a working group called the Software-Defined Networking Research Group (SDNRG) which aims at helping the other standardization organizations by investigating the interesting open research issues related to SDN.

This article targets the orchestration of services and network resources using SDN and NFV technologies, which are the topics of ongoing standardization activities as described above. This article also treats the intersection of SDN and NFV technologies, which is the objective of Phase 2 of the ETSI NFV group. Moreover, this article proposes a new architecture for mobile networks with the addition of a knowledge plane, which will help mobile operators gain more granular and user-aware control as well as orchestration of services and network resources.

Conclusions

We introduced CSDN, an architecture for future mobile networks that adopts the separation of the data plane from the control plane for intelligent service orchestration, and virtualization of network functions for centralized dynamic resource management. We discussed the design considerations of the proposed framework at different architectural levels, and we outlined the amendments that need to be introduced to current SDN approaches in order to adopt it in future cellular networks. Furthermore, CSDN takes advantage of the data generated at the last mile by MSPs to compile comprehensive information about users and their network conditions in order to optimize network utilization and to enhance the user experience by providing user-adapted, context-related, and personalized services. Such information can transform the network operator into a big data operator with the capability to share their data with third party service providers for innovative applications. In this context, we investigated the different challenges and obstacles to overcome before having a standardized and open architecture for extraction, management, and exploitation of contextual data in mobile networks. We expect that CSDN with big data analyses can catalyze the innovation in future cellular networks while reducing the cost and time-to-market for new adapted and personalized services.

References

[1] 3GPP TS 36.300 (v10.8.0), "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2 (Release 10)," July 2012.

[2] J. H. Lee *et al.*, "Distributed IP Mobility Management from the Perspective of the IETF: Motivations, Requirements, Approaches, Comparison, and Challenges," *IEEE Wireless Commun. Mag.*, vol. 20, no. 5, Oct. 2013, pp. 159–68.

[3] A. Reaz and R. Boutaba, "Design Considerations for Managing Wide Area Software Defined Networks," *IEEE Commun. Mag.*, vol. 52, no. 7, 2014, pp. 116–23.

[4] ETSI NFV White Paper, available: http://portal.etsi.org/nfv/nfv_white_paper.pdf.

[5] T. Taleb, "Towards Carrier Cloud: Potential, Challenges, & Solutions," *IEEE Wireless Commun. Mag.*, vol. 21, no. 3, June 2014. pp. 80–91.

[6] User Data Convergence (UDC), 3GPPP Technical Specification (TS 29.335), available: http://www.3gpp.org/DynaReport/29335.htm.

[7] C. J. Bernardos *et al.*, "An Architecture for Software Defined Wireless Networking," *IEEE Wireless Commun.*, 2014, vol. 21, no 3, pp. 52–61.

[8] L. E. Li, Z. M. Mao, and J. Rexford, "CellSDN: Software-Defined Cellular Networks," Techinical Report, Princeton University, 2012.

[9] N. McKeown *et al.*, "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Comp. Commun. Rev.*, 38.2, 2008, pp. 69–74.

[10] T. Taleb and A. Ksentini, "Follow Me Cloud: Interworking Federated Clouds & Distributed Mobile Networks," *IEEE Network*, vol. 27, no. 5, Sept./Oct. 2013, pp. 12–19.

[11] P. Makris, D. Skoutas, and C. Skianis, "A Survey on Context-Aware Mobile and Wireless Networking: On Networking and Computing Environments' Integration," *IEEE Commun. Surveys & Tutorials*, no.99, pp. 1–25.

[12] M. Bushong, 2013, Oct., SDN, Big Data and the Self-Optimizing Network, available: http://www.infoworld.com/article/2612727/sdn/sdn--big-data--and-the-self-optimizing-network.html.

Biography

Abbas Bradai (bradai@imag.fr) received his M.S in computer science from the National Institute of Computer Science (ESI ex-INI), Algiers, Algeria, and from The University of Rennes1, France in 2009, and his Ph.D. at the LaBRI/University of Bordeaux-1, France, in 2012. He is an assistant professor at the University of Grenoble and a research fellow at LIG lab, Grenoble. His main research interests are multimedia communications over wired and wireless networks, cognitive radio, software defined networking, and virtualization. He is/was involved in many French and European projects (FP7, H2020) such as ENVISION, VITAL, Data-Tweet.

Kamal D. Singh (kamal.singh@univ-st-etienne.fr) obtained his Ph.D. degree in computer science from the University Rennes 1, France in 2007. He then worked as a post doc in the Dionysos group at INRIA, where he co-developed many components of quality of experience estimation tools, and worked mainly on the analysis of video-based applications. He is currently an associate professor at the University of Saint Etienne/Telecom Saint Etienne, France. His research interests include quality of experience, video streaming, software defined networking, big data and semantic web.

Toufik Ahmed (tad@labri.fr) is currently a professor at ENSEIRB-MATMECA School of Engineers in the Institut Polytechnique de Bordeaux (IPB), and is performing research activities at the CNRS-LaBRI Lab-UMR 5800 at the University Bordeaux 1. His main research activities concern quality of service (QoS) management and provisioning for multimedia wired and wireless networks, media streaming over P2P network, cross-layer optimization, and end to-end QoS signaling protocols. He has also worked on a number of national and international projects. He is serving as TPC member for international conferences including IEEE ICC, IEEE GlobeCom.

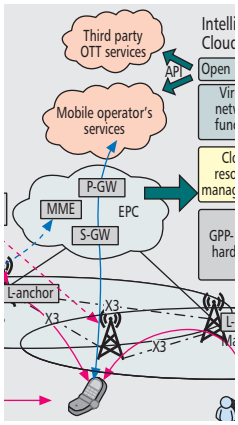
Tinku Rasheed (tinku.rasheed@create -net.org) is a senior research staff member at Create-Net. Since May 2013, he has been heading the Future Networks R&D Area [FuN] within Create-Net. Before joining Create-Net in December 2006, he was a research engineer with Orange Labs R&D from May 2003 until November, 2006. He received his Ph.D. degree from the Computer Science Department at the University of Paris-Sud XI., in 2007. He completed his M.S. degree in 2003 from Aston University, U.K. specializing in telecommunication engineering, and his bachelor degree in 2002 in electronics engineering from the University of Kerala, India. He has extensive industrial and academic research experience in the areas of mobile wireless communication and data technologies, and end-to-end network architectures and services. He has several granted patents on distributed protocols for wireless networking, and has published his research in major journals and conferences.

CSDN takes advantage of the data generated at the last mile by MSPs to compile comprehensive information about users and their network conditions in order to optimize network utilization and to enhance the user experience by providing user-adapted, context-related, and personalized services.

i-Net: New Network Architecture for 5G Networks

In order to deal with the challenges and opportunities brought about by mobile Internet, the authors propose a new Internet-oriented mobile network architecture, dubbed i-Net, for 5G mobile communication systems. Based on direct communications established between the base stations, i-Net can implement local data routing for mobile data traffic, which adapts to increased service localization for mobile internet.

Jianquan Wang, Zhaobiao Lv, Zhangchao Ma, Lei Sun, and Yu Sheng



Abstract

In order to deal with the challenges and opportunities presented by the mobile Internet, this paper proposes a new Internet-oriented mobile network architecture, dubbed i-Net, for fifth generation (5G) mobile communication systems. Based on direct communications established between the base stations (BSs), i-Net can implement local data routing for mobile data traffic, which adapts to increased service localization for the mobile Internet. We first present the network architecture and main technical features, i.e., inter-connection, integration, and intelligence. Then the key techniques are discussed to show their potential in improving the network efficiency of i-Net. Field trials for inter-BS direct communications to implement cooperative multi-point oPeration (CoMP) in i-Net are carried out to demonstrate the feasibility of i-Net. It is expected that i-Net can provide a good quality of experience for end users through the efficient integration of the multi-BS, multi-band, and multi-radio-access-technology (RAT) radio resources in 5G.

COMMUNICATIONS STANDARDS

Introduction

With surging mobile Internet services, mobile operators are facing unprecedented opportunities and challenges. To meet the demands due to the rapid development of multimedia services, it is far from enough to only expand and upgrade current networks, which results in a substantial increase in CAPital EXpenditure (CAPEX) and OPerating EXpenditure (OPEX) without reasonable investment returns [1]. Therefore, in order to provide a good quality of experience (QoE) for users with affordable costs, a new network architecture is desired by telecommunications operators to enable sustainable network evolution to the fifth generation (5G).

Moreover, the mobile network does not have to be limited to a pipe between mobile users and over-the-top (OTT) services. Instead, operators have the ability to provide differentiated QoE and develop new services for mobile customers. Thus, close collaboration between the mobile network and OTT service providers is another important factor to consider when designing 5G mobile networks, with the aim of providing an effective means to maintain sustainable development of mobile telecommunications eco-systems.

Until now, much attention has been paid to emerging 5G technologies ranging from information processing to networking techniques. Most of the focus has been on further improving radio access efficiency through advanced signal processing, e.g., massive MIMO, millimeter-wave communications, cooperative relaying, and so on [2–4]. On the other hand, the existing shaft-like network structure with centralized data processing is not suitable for flexible and diverse mobile Internet services, which undoubtedly restricts the service provisioning ability of the mobile network. Hence, designing a new network architecture becomes urgent.

For example, cloud radio access network (RAN) has been proposed, in which centralized processing baseband units (BBUs) are connected to remote radio units (RRUs) via optical fibers [5]. In [6, 7], the proposed framework for cloud computing aims at a smooth migration of all or only part of an ongoing IP service based on mobile networks. All these emerging new network architectures can efficiently exploit advances in new techniques such as cloud computing and software defined networking (SDN) [9]. However, there is still much work that needs to be done to improve the mobile network architecture with good backward compatibility. This paper emphasizes how to maximize the utilization of millions of currently deployed base stations (BSs), and to smoothly evolve toward converged 5G networks.

The rest of this article is organized as follows. We introduce the motivation for i-Net. The architecture and main features of i-Net are described. The key techniques enabling i-Net are investigated. Finally, we present our conclusions and a description of future work that needs to be done.

Motivation

The rapid development of the mobile Internet is promoting the prosperity of mobile broadband networks (MBNs), as well as presenting new challenges, including:

New mobile services. Nowadays, users of all ages want to share information at anytime and anywhere via MBNs, the so called SoLoMo (Social, Local and Mobile) trend. Also, the growth of the Internet-of-Things (IoTs) has had serious effects on mobile network performance [10]. However, current mobile networks were originally designed to enable mobile subscribers (MSs) to gain access to resources and services provided by the fixed Internet.

Boosting demands for capacity and user experience. It is expected that by the year 2020 average user data rates will increase from around 20 Mb/s to 1 Gb/s, and connection density will also increase from 140 thousand connections per square kilometers to 6 million connections per square kilometers [8]. Moreover, new applications such as the Internet-of-Vehicles are envisioned for 5G, which will demand much shorter latency from the underlying network.

The authors are with China Unicom Group Company.

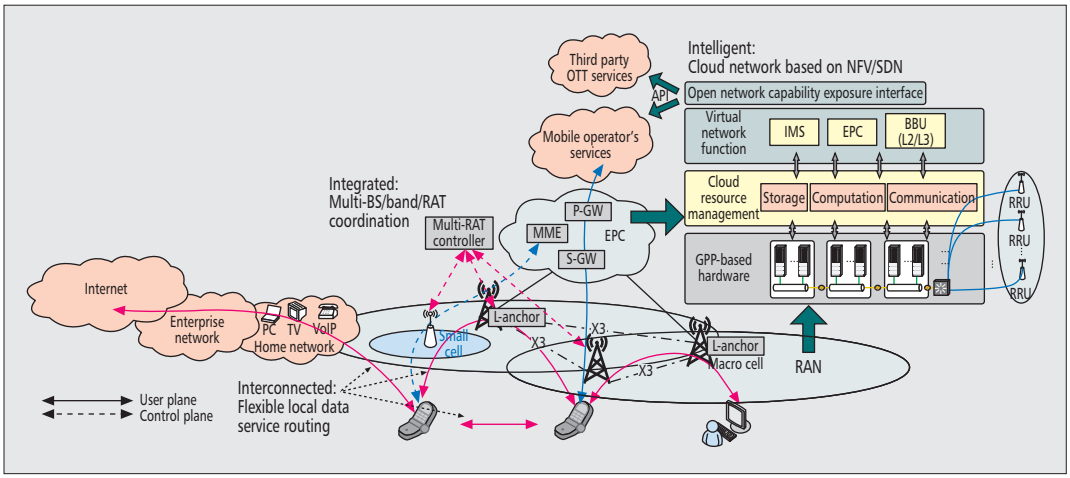


Figure 1. i-Net architecture vision.

The i-Net architecture is an evolution of existing networks, which includes base stations for macro cells and small cells, and CN elements including the mobility management element, the serving gateway, the packet data network gateway, and so on.

Severe investment/income imbalance. Considering the existing network infrastructure, telecommunications operators need to invest heavily on both network maintenance and upgrade to support ever increasing capacity requirements. However, revenues from increased data traffic cannot keep up with the investment. How to reduce CAPEX/OPEX and create new revenue flows are crucial issues facing telecommunications operators.

The existing mobile network architecture has its inherent limitations, which are described as follows.

Vertical Shaft-Like Architecture

Traditional networks are built on a shaft-like architecture. All service data flows, including communications between users served by a single BS, go through both the RAN and the core network (CN), and then ultimately to the destination. With increased data traffic and emerging interactive mobile traffic, this architecture suffers from inevitable congestion and the endless expansion of core and transport networks. Moreover, data routing for traffic between users is redundant, which may cause unnecessary latency.

Isolated Island-Like Structure

The isolated island effect is another main drawback of current mobile networks. First, millions of BSs have been deployed and are "loosely connected." That is, the interaction between BSs is only limited to the handover and static inter-cell interference coordination (ICIC) functions via the X2 interface. Efficient radio signal cooperation and resource sharing between BSs are not possible. From the spectrum perspective, each operator owns multiple frequency bands, but cannot aggregate them into a single large band, since frequencies are allocated to different systems independently. From the system aspect, quite a few different wireless systems co-exist, e.g., the second and third Generation (2/3G) networks, Long Term Evolution (LTE), and wireless local access networks (WLANs). However, efficient coordination among these systems is lacking, leading to serious inter-operability and complex spectrum re-farming issues.

Application-Specific Hardware Equipment

An entire mobile network, including the RAN, the transport network, the CN, and the service chains, is constructed by application-specific hardware equipment with little flexibility and scalability. As a result, network upgrades and expansion is usually complex and expensive. Moreover, it is time-consuming to deliver new services, which stifles service innovation. For instance, if a service provider wants to accelerate the mobile access speeds of its VIP customers, the operator needs to modify the quality of service (QoS) policy executed at a gateway, e.g., a gateway GPRS support node (GGSN) for 3G, with a command line interface, which usually takes several months for approval and operation. Besides, gateway memory is very limited and not scalable to accommodate the ocean-level rules required by service providers.

From the above discussions, the evolution toward 5G networks needs to provide reliable support for mobile Internet services; provide a much stronger pipe with substantially lower cost; and enable service innovation to generate new revenue flows. In order to achieve these goals, a new network architecture, i.e., i-Net, is expected to offer the features of flexible data plane steering consistent with service routes; integration of various radio access techniques and coordination of radio resources; and virtualization of infrastructure to facilitate fast network deployment and service innovation.

i-Net Network Architecture and Features

Architecture

The i-Net architecture is an evolution of existing networks, which includes base stations for macro cells and small cells, and CN elements including the mobility management element (MME), the serving gateway (S-GW), the packet data network gateway (P-GW), and so on.

As illustrated in Fig. 1, the RAN is enhanced with access nodes capable of direct inter-BS communications. This kind of access node is termed an i-Node, which integrates the radio access capability of the traditional BS and the new direct inter-connection interface, defined as

	LTE-Advanced R10	i-NET Stage 1	i-NET Stage 2	i-NET Stage 3
RAN	Loose inter-connection between BSs	• Cluster i-Nodes with local anchor • Enable inter-BS coordinated scheduling • Add multi-RAT controller	• Centralize BBUs with clusters • Add inter-BS data switch • Enable inter-BS CoMP and CA	• New BBU-RRU interface • vRAN achieved • vEPC-U integrated to achieve flexible data routing globally
CN	P-GW plays the only anchor	MME is enhanced for local routing decision	vEPC achieved	vEPC splits into vEPC-C and distributed vEPC-U merged with vRAN

Table 1. Comparison between i-Net under different stages and LTE-Advanced R10 networks.

the X3 interface. The service data flows between mobile users under various scenarios, including enterprise and home networks, can be directly routed among i-Nodes. In addition, the CN should be capable of establishing and storing the relationships between i-Nodes in order to determine whether a service flow occurs within an inter-connected region.

Moreover, the coordination capability among multi-BS, multi-bands, and multi-RATs is also enhanced. The i-Nodes can exchange information via an inter-connected channel to achieve signal and resource coordination. A general control node across the 2G/3G/LTE BSs and WLAN access points (APs) is introduced, and then all information across different RATs can be collected to facilitate multi-RAT coordination. In addition, the entire network infrastructure may evolve toward a virtual cloud-based structure for both RAN and CN.

Evolving from existing networks, the long-term targets of i-Net can be achieved in three steps. In the first stage, the current LTE BSs can be grouped into clusters, with a cluster-head enhanced with the local data anchor (L-anchor) function, which can build a direct inter-BS data bearer within the cluster for local service routing. The inter-BS direct data tunnel can be utilized to exchange scheduling information to realize coordinated scheduling. In addition, 2G/3G/LTE BSs and WLAN APs can be coordinated using the introduced multi-RAT controller.

The second stage is the transitional period toward full network virtualization. It is feasible that the CN is first to be virtualized, where the evolved packet core (EPC) of LTE can be implemented with network functions virtualization (NFV), namely vEPC. The BBUs of the BS cluster can be deployed within the same site, and connected with a high speed Ethernet switch to strengthen the information exchange capability. In this way, highly efficient coordination between BSs can be achieved, e.g., inter-BS cooperative multi-point operation (CoMP) or inter-BS carrier aggregation (CA).

The third stage is characterized by the realization of NFV and SDN for the entire network. At the RAN side, the interface and functional split between the BBU and RRU need to be redefined, i.e., a part of the physical (PHY) layer of the BBU may be moved to the RRU part, in order to facilitate the introduction of 5G RAT and the inter-connection of BBUs. Virtual RAN (vRAN) can be implemented with new BBUs based on general-purpose-processor (GPP) servers, and thus resource sharing among i-Nodes can be achieved. On the other hand, based on the principle of splitting the control plane and the user plane, vEPC can be separated into two components, i.e., EPC with C-plane (vEPC-C) and EPC with U-plane (vEPC-U), while EPC-U is distributed to each vRAN side. Thus, through the unified control of EPC-C, flexible data routing within the entire network can be achieved. Table 1 compares i-Net under different stages with the LTE-Advanced Release 10 (R10) network.

Main Features

The main technical features of i-Net are illustrated in detail as follows.

Inter-Connection: The inter-connection feature of i-Net aims to tackle potential problems caused by intensive interactivity between users of mobile Internet applications. It intends to modify the traditional vertical shaft-like structure, and establish a direct data channel between users only via i-Nodes. With controllable direct data transmission via i-Nodes, i-Net is capable of offloading traffic within the RAN without the involvement of the CN. In this way, the traditional two-way transmission of "User↔RAN↔CN↔Server↔CN↔RAN↔User is simplified into "User↔i-Nodes↔User," or even "User↔User" [11]. Thus the traffic in the CN and transport networks can be offloaded and the latency for local services can be effectively reduced.

A great number of application scenarios can benefit from service localization. For hotspot areas such as enterprises, shopping malls, sport stadiums, and campuses, i-Net can provide efficient local file transfer, local multi-media streaming, local video-conference, local advertisements, a local content delivery network (CDN), etc. Moreover, proximity-based interactive services, e.g., mobile social networks and mobile P2P sharing, can also benefit from service localization.

Integration: i-Net aims to integrate and exploit all available radio resources via tight coordination among multiple BSs, multiple frequency bands, and multiple RATs, in order to achieve the aims of "single network," "aggregated carrier" and "consistent experience."

Cooperative multi-BS signal processing can be adopted to mitigate multi-cell interference and improve the cell-edge user experience. Although the CoMP technique has been investigated in Third Generation Partnership Project (3GPP) Release 11 [12], it only applies to limited scenarios, e.g., BBUs are co-located at the same site. On the other hand, more flexible use

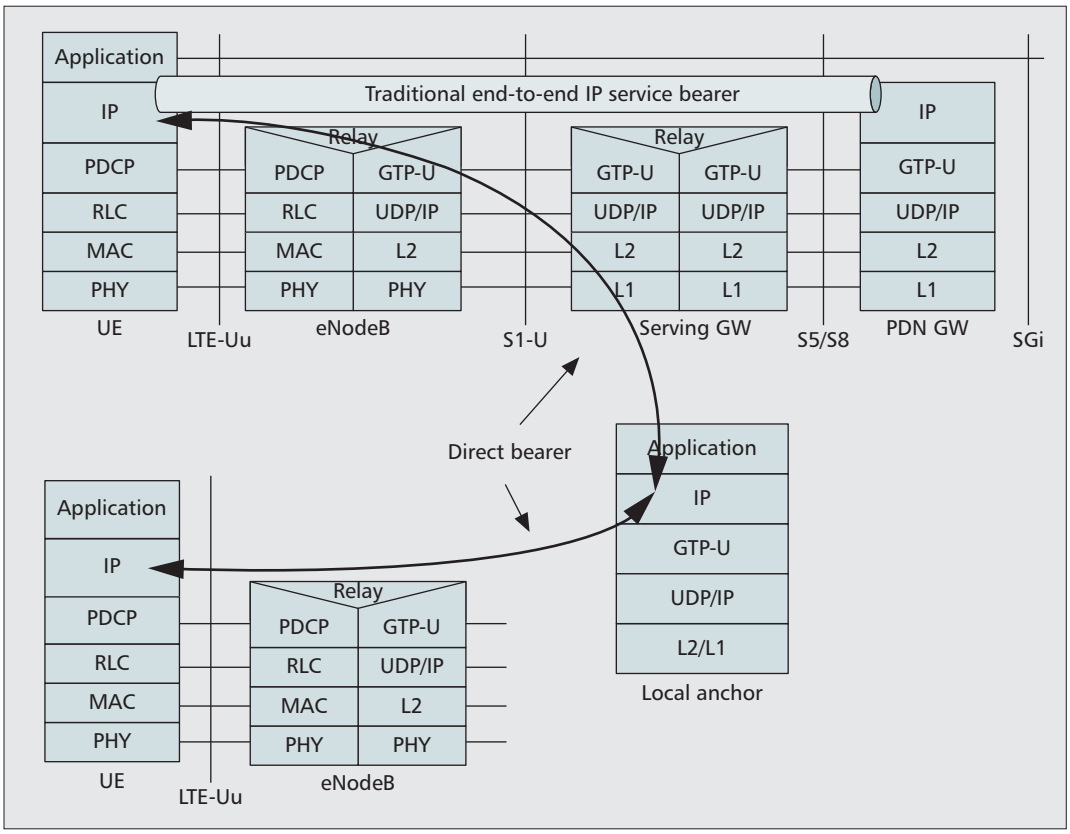


Figure 2. Inter-connected network based on local anchor.

of CoMP is expected in i-Net, i.e., no limitation of co-located BSs, and sharing processing resources among BSs to balance BBU utilization.

Moreover, i-Net intends to make full use of the multiple frequency bands via the evolved CA function of LTE-Advanced [13] in an attempt to provide high data rates for mobile customers. First, an operator may own a larger number of licensed and unlicensed frequency bands including TDD and FDD. Because the coverages of these bands are very different, the carries perceived by a user may come either from the same site or from different sites. This situation can be well accommodated by i-Net, where all the bands, including TDD and FDD, licensed and unlicensed, intra-site and inter-site, can be aggregated for the end user.

In addition, since the co-existence of the existing and evolving RATs such as 3G, LTE, and future 5G networks, is expected to last for a long time, the i-Net integration also focuses on enhanced coordination among different RATs. Via the multi-RAT controller, i-Net can improve the interworking experience for users roaming among different RATs. i-Net supports smoother spectrum re-farming for legacy systems evolving toward advanced RAT with dynamic frequency sharing, which has been investigated for 2G/3G and LTE in 3GPP [14]. That is, i-Net is able to dynamically allocate spectra among a variety of RATs when their traffic load varies on a cell-by-cell basis.

All coordination operations require an interconnected channel to exchange information between i-Nodes, which can be conveniently provided by i-Net with the inter-connection feature.

Intelligence: It is envisioned that i-Net will greatly reduce the costs of network construction and maintenance through the intelligent feature, while enabling more flexible service deployment and value-added innovation.

In lieu of traditional network elements based on application-specific hardware, NFV based on the GPP can greatly enhance mobile network scalability and flexibility, resulting in reduced cost. Moreover, network virtualization can provide enabling infrastructure for advanced interconnection and integration features, i.e., flexible data routing and baseband sharing. It can also facilitate new service development and deployment, which could possibly reduce the time-to-market (TTM) from months to seconds. It can provide a flexible network application programming interface (API) to third party service providers. Thus, network-as-a-service (NaaS) is envisioned to provide new competitive edges in the mobile Internet era.

Key Technologies in i-Net

In this section, only selected techniques enabling the i-Net features are investigated. Much work is needed to make i-Net feasible in the near future.

Service Localization Enabling Techniques

As shown in Fig. 2, in the current RAN, the end-to-end IP-based service bearers originate from the user equipment (UE) and terminate at the CN gateway, e.g., the P-GW in LTE. Moreover, the BS only forwards data packets to its pre-configured destination, e.g., either a radio

In lieu of traditional network elements based on application-specific hardware, NFV based on the GPP can greatly enhance mobile network scalability and flexibility, resulting in reduced cost. Moreover, network virtualization can provide enabling infrastructure for advanced inter-connection and integration features.

A compromise solution is to deploy a local anchor in i-Net, which plays the role of the local gateway at places with abundant local services. It can be co-located with i-Node or deployed stand-alone with a cluster of directly connected i-Nodes.

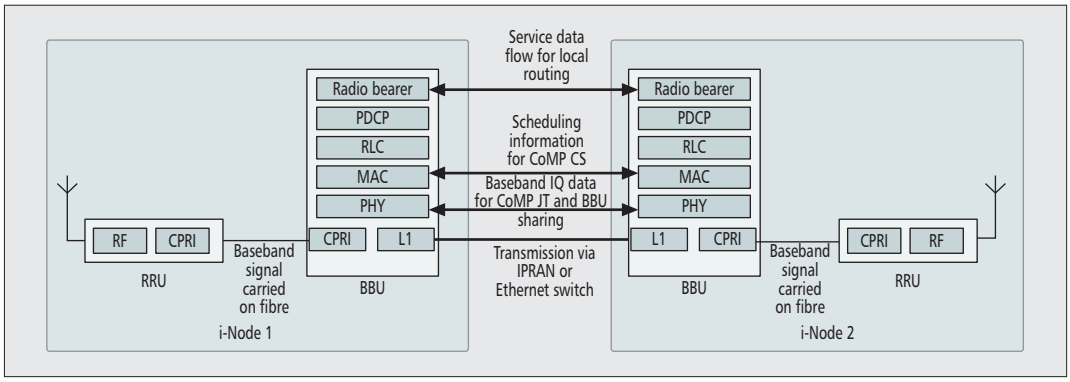


Figure 3. Inter-BS information exchange via the inter-connected channel.

network controller (RNC) in the 3G network or a S-GW in the LTE network.

One advantage of i-Net lies in its ability to achieve RAN-level service localization. The establishment of direct communications between i-Nodes will provide a fundamental architecture to support local data transmission. However, how to enable the efficient transmission of "User↔i-Node↔User" still requires much more effort in designing resource control schemes.

Apparently, simple implementation of the local routing function based on BS enhancements is extremely difficult and costly, which entails adding IP layer routing functions and much storage memory to millions of BSs.

A compromise solution is to deploy a local anchor in i-Net, which plays the role of the local gateway at places with abundant local services. It can be co-located with i-Node or deployed stand-alone with a cluster of directly connected i-Nodes. When a UE initiates a service request, the MME of the CN will detect if the service originator and destination are served by i-Nodes with a direct tunnel within the same local anchor, and then decide whether the service flow can be locally routed or still works as usual, i.e., through the RAN and CN.

After the CN decides on service localization, the local anchor needs to set up a direct bearer between the source UE and destination UE, as illustrated in Fig. 2. Both the source and destination UEs establish a direct bearer terminated at the IP layer of the local anchor, where one-to-one mapping between direct bearers is made in order to construct the end-to-end tunnel for local data routing. Part of the CN functions, e.g., mobility, charging, and security, needs to be moved down to the local anchor, so as to guarantee effective control and management for local services. However, if the UE moves out of the coverage area of i-Nodes connected to the local anchor, the local data flow inevitably encounters service interruption. Hence, in order to avoid interruption due to user mobility, the UE needs to be capable of establishing two kinds of bearers at the same time, i.e., one for local data services, and the other for CN services with strict continuity, e.g., voice calling.

This scheme has little impact on existing mobile networks, while the drawback is that only areas with a local anchor deployed can achieve such local traffic offloading.

Multi-BS Cooperative Resource Management

Multi-BS cooperative resource management is expected to further enhance radio network efficiency through tight coordination among BSs, which requires the efficient direct communications between i-Nodes provided by i-Net. Unlike IP-based service data flows in local service routing, the information delivered between i-Nodes is signal processing related and dependent on the coordination purpose, as illustrated in Fig. 3. Therefore, it is feasible to implement intelligent signal cooperation and baseband sharing techniques.

Intelligent Signal Cooperation: In order to mitigate co-channel interference, coordinated BSs exchange information to implement joint signal processing to reduce or even avoid multi-cell interference. There have already been several CoMP techniques developed in LTE-Advanced, e.g., coordinated scheduling (CS) and joint transmission (JT). However, current CoMP techniques assume perfect backhaul, and thus only intra-BS CoMP, i.e., coordination between the different sectors of the same BS, can be performed since the latency of the X2 interface defined in 3GPP cannot fulfill CoMP requirements. But in a practical network, both intra-BS and inter-BS interferences are strong, especially in densely populated urban areas. According to the field tests in the urban area at Shanghai, the strongest interference in more than 60 percent of cases comes from cells belonging to other sites.

Thanks to the inter-connection capability between i-Nodes, the coordinated BSs can effectively exchange information to perform intelligent signal cooperation and to suppress inter-BS interference in i-Net. The requirements for inter-BS information exchange depend on different CoMP algorithms. For CS, it is the scheduling information that needs to be exchanged between BSs. It is estimated that the exchange of this information requires a bandwidth for 5–10 Mb/s transmission per BS, and results in latency of less than 4 ms for a 20 MHz LTE cell with two transmit and two receive (2T2R) antennas. For JT, U-plane physical signals need to be shared among coordinated cells for joint transmit/receive operation. The resultant bandwidth requirement is higher than that of CS, estimated to be around 400 Mb/s per BS for a 20 MHz LTE cell with 2T2R antennas. The latency

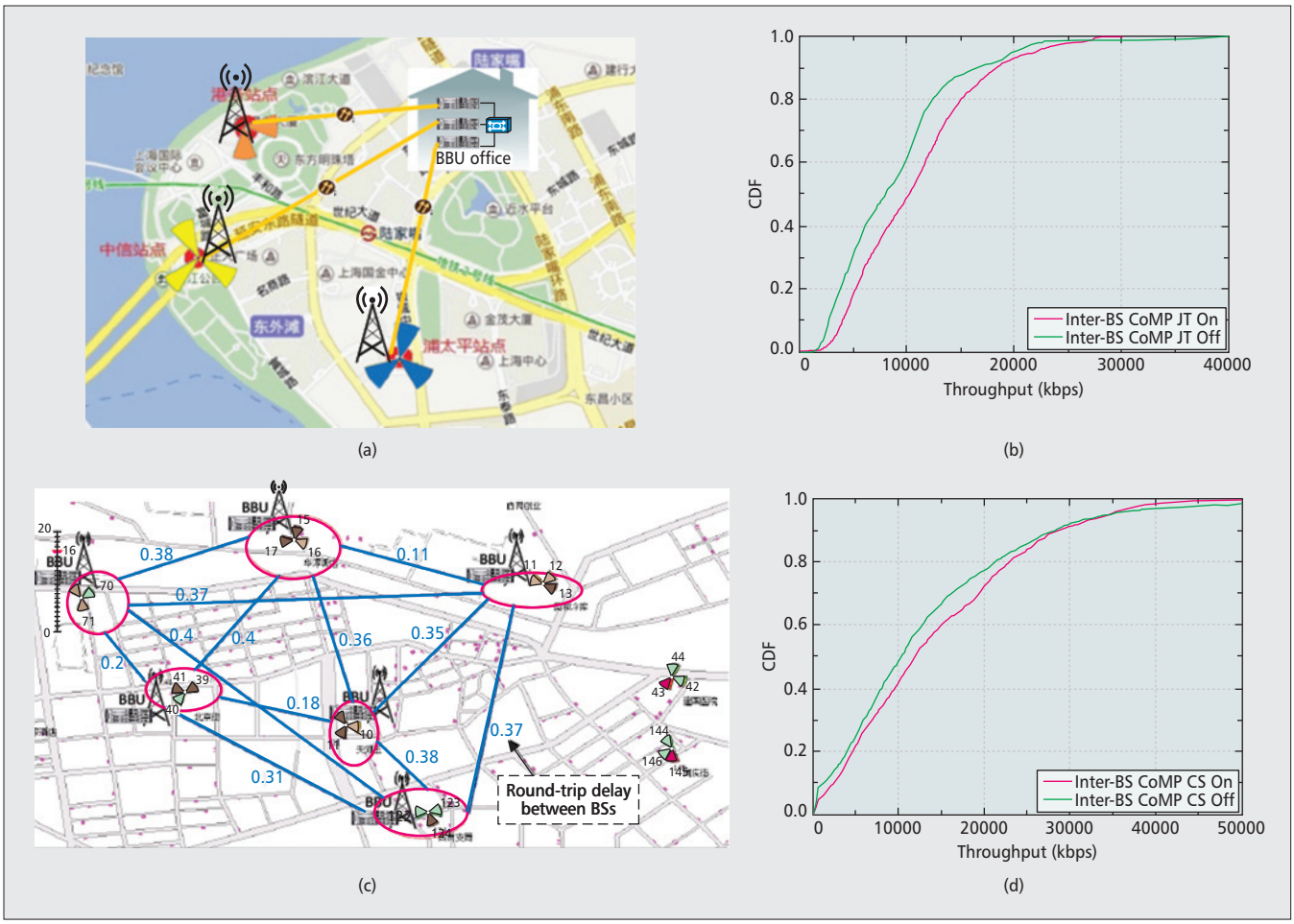


Figure 4. Performance evaluation on Inter-BS CoMP in i-Net: a) centralized BBUs test scenario; b) inter-BS CoMP JT performance; c) distributed BBUs test scenario; and d) inter-BS CoMP CS performance.

requirement is also tighter, i.e., lower than 0.3 ms, in order to satisfy the 1 ms transmit time interval (TTI) of LTE.

Based on existing networks, there are two feasible deployment scenarios. For the centralized BBU case, the BBUs can be connected with a high-speed switch located in the same-site office, and thus CoMP JT can be applied. For the distributed BBU case, the BBUs are connected via the transport network based on the IP radio access network (IPRAN) scheme, where only CoMP CS can be implemented with a low bandwidth requirement.

Field trials were carried out to verify the inter-BS CoMP implementation and performance. In the centralized scenario as shown in Fig. 4a, there are three RRU sites associated with eight LTE 20 MHz 2T2R cells, and the associated BBUs are centralized in the given site to perform UpLink(UL) CoMP JT. In the distributed scenario as shown in Fig. 4c, there are six RRU sites with 11 LTE 20 MHz 2T2R cells, and the BBUs are co-located with the RRUs, inter-connected by IPRAN to perform CoMP CS. The inter-site distance is around 300 meter. At least one test UE per cell is distributed in the test area to create a real interference environment with full buffer service applied, and there are two additional UEs moving around the coverage area to test how the ergodic performance is affected by inter-BS CoMP. As shown in Fig. 4b, the gain of cell edge user throughput (i.e., 5 percent cumulative probability distribution function (CDF)) of inter-BS CoMP JT compared to intra-BS JT is around 50 percent, and the average system gain is 10 percent. Figure 4d shows that a cell edge gain of 80 percent can be achieved while the average system gain is negligible with inter-BS CoMP CS compared to the case without CoMP (frequency selective scheduling is enabled).

Intelligent Baseband Sharing: The unbalanced traffic distribution results in varying traffic loads among different cells. For example, network statistics show that 70 percent of data traffic is carried by 30 percent of BSs. Dramatic traffic variances among adjacent cells are also observed.

However, in the traditional BS structure, a RRU can only be connected to its corresponding BBU, where baseband processing resources are reserved for cells associated with its respective RRU. In this case, each BBU has to be configured to meet the demands of the maximum traffic loads during peak hours, although the BBU is not fully utilized most of the time.

With the i-Node inter-connection capability, the RRU baseband signal can be transferred directly across BBUs. This can break the limitation that one BBU can only process its connected RRU's data in traditional networks, and thus

The latency for the CPRI interface is below approximately 800 ms depending on the BBU processing delay, which sets strict requirements on the inter-BS interface. These issues make baseband sharing difficult for application in existing networks, which is expected to be solved by i-Net in Stage 3.

enable the sharing of a BBUs' processing powers.

In this way, a large number of BBUs can form a "base station pool" with aggregated processing capabilities to serve the traffic requirements of the coverage area. The baseband signal from the RRUs can be processed by any BBU in the BS pool, instead of only one single BBU. For instance, when there are certain regions with heavy traffic loads and certain BBUs are overloaded, the corresponding RRU baseband data can be transferred to and then processed by other idle BBUs using i-Node direct communications.

However, the inter-BS sharing of baseband data places strict requirements on inter-BS data exchange bandwidth and latency. The required baseband signal bandwidth over the BBU-RRU interface following the CPRI standard for a 20 MHz LTE cell with 2T2R antennas is about 2.5 Gb/s. When there are ten three-sector BS sites sharing the baseband, the required bandwidth for the cross-BBU switch is up to approximately 75 Gb/s. The latency for the CPRI interface is below approximately 800 μs depending on the BBU processing delay, which sets strict requirements on the inter-BS interface. These issues make baseband sharing difficult for application in existing networks, which is expected to be solved by i-Net in Stage 3.

Network Virtualization Techniques in i-Net

Network virtualization provides the foundation for the i-Net intelligent feature and boosts other potential features in i-Net. The NFV architecture has been discussed by the European Telecommunications Standards Institute (ETSI) [15]. In the virtualized infrastructure layer, GPP-based hardware can be widely adopted instead of traditional application-specific hardware, and all kinds of resources for communications, computation, and storage can be shared through cloud resource management. Above the virtual infrastructure layer, the original network elements can be implemented by software functions in the cloud, such as EPC, IP multimedia subsystem (IMS), and even RAN. Moreover, this virtualized network is capable of sharing information and control abilities with a third party via an open interface, which makes it easy to add value to mobile networks.

There are huge challenges in virtualizing i-Net, especially for the RAN functions. First, the physical layer radio signal transmission and reception incurs huge computation complexity and real-time processing, for which GPP-based hardware is difficult to implement with low energy consumption. Second, cloud resource management for BBUs encounters a similar problem of inter-BS transmission bandwidth and latency requirement. as mentioned in the above baseband sharing.

Nevertheless, it is envisioned that more antennas, e.g., massive MIMO, and larger frequency bandwidth, e.g., mm-wave bands, will be applied in 5G. As a result, the baseband in-phase and quadrature (IQ) data bandwidth transferred on the current BBU-RRU CPRI interface will increase linearly with the number of antennas and frequency bandwidth. Moreover, the bandwidth requirement for coordination between i-Nodes for either intelligent signal cooperation or baseband sharing may also scale up linearly, which makes coordination more difficult. Even if the BBUs of i-Nodes are centralized to facilitate information exchange, it still brings about the challenge of long distance transmission of the large bandwidth CPRI signal between the BBU and RRU. Therefore, it is necessary to propose new designs of the BS architecture and a new definition of the BBU/RRU interface for i-Net.

One possible solution is that part of the PHY layer functions located at the BBU can be relocated to the RRU side, where radio signal processing can be done locally at the RRU. The original MAC and higher layers are left at the BBU side. Besides, there are still parts of the PHY functions remaining at the new BBU for CoMP related physical layer signal processing. In this way, the bandwidth required for the new BBU-RRU interface can be greatly reduced.

The new BS architecture not only facilitates centralized BBU deployment, but also facilitates the application of NFV techniques to the BBU, since the MAC and higher layer computation complexity and delay requirements are not as strict as those for the physical layer. Based on centralized BBUs, the virtual RAN can be realized by NFV, and thus intelligent baseband sharing and signal coordination can be achieved. Then, i-Net services can be realized in a convenient manner. Here, only two examples are given for the purpose of illustration.

Localized Services: It becomes much easier to implement the full inter-connection feature in the virtualized cloud network. The BBU functions can be implemented as logical virtual machines running in the cloud. Thus, all i-Nodes can be viewed as logically inter-connected. Therefore, the direct logical bearer between the service originator and terminator can always be set up. Then, a SDN controller of RAN transportation, which can intelligently detect the transport layer status, determines the optimized route for the direct bearer between i-Nodes. In this scheme, the mobile network can precisely adapt to mobile Internet traffic via flexible local traffic steering.

CDN Services: It is difficult to implement CDN in a traditional mobile network since the CDN servers can only be deployed outside an operator's gateway. Moreover, CDN service requires good scalability and an open interface to the service providers. With the i-Net interconnection feature in conjunction with the NFV/SDN technique, it becomes easy to deploy CDN services in the mobile network. First, the local anchor used for local routing can be implemented using NFV, and the CDN functions can be deployed within the virtualized local anchor. Then a direct bearer between the end user and the nearest local CDN server can be established. In this way, it can provide scalable CDN services within the mobile network, and also give a network API to service providers, who can flexibly call the CDN API to accelerate their services.

Conclusion and Future Work

This paper has pointed out the challenges and importance to break through traditional network constraints, while the requirements of the mobile Internet are increasing rapidly. Toward this end, we have proposed the i-Net architecture, which combines the advantages of inter-connected,

integrated, and intelligent networks. The evolution path toward i-Net was also discussed as well as the key techniques for i-Net implementation. Based on direct communications among i-Nodes, service localization can be achieved so that the transmission route within RAN can be optimized to align with the service route of mobile Internet traffic. With close coordination between i-Nodes, multi-BS cooperative resource management can be performed to achieve a more consistent experience. Field trials have been carried out to demonstrate the gains for inter-site CoMP in i-Net, which shows 50 percent cell edge gains for UL CoMP JT compared with intra-site CoMP only, and 80 percent cell edge gain for CoMP CS compared with the scenario without CS.

However, there is still much research needed to fully exploit the potential of i-Net. For instance, the current cell-edge user throughput, which is improved with i-Net inter-site CoMP, still has a large gap between the visions of "no edge" consistent experience. In particular, the issue of severe interference with ultra-dense deployment of i-Nodes needs to be addressed. Therefore, i-Net needs to be further refined to take advantage of the opportunities in the mobile Internet era.

References

[1] G. Piro *et al.*, "Information-Centric Networking and Multimedia Services: Present and Future Challenges," *Trans. Emerging Telecommunication Technologies (ETT)*, vol. 25, no. 4, 2014, pp. 392–406.

[2] F. Boccardi *et al.*, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol.52, no. 2, 2014, pp. 74–80.

[3] K. Zheng *et al.*, "10 Gb/s Hetnets with Millimeter-Wave Communications: Access and Networking-Challenges and Protocols," *IEEE Commun. Mag.*, vol. 53, no. 1, 2015, pp. 222–31.

[4] G. Wang, W. Xiang, and J. Yuan, "Generalized Wireless Network Coding Schemes for Multi-Hop Two-Way Relay Channels," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, Sept. 2014, pp. 5132–47.

[5] C-RAN: The Road Towards Green RAN white paper, CMCC, Nov. 2011

[6] T. Taleb and A. Ksentini, "Follow Me Cloud: Interworking Federated Clouds & Distributed Mobile Networks," *IEEE Network*, vol. 27, No. 5, Sep./Oct., 2013, pp. 12–19.

[7] A. Ksentini, T. Taleb, and F. Messaoudi, "A LISP-based Implementation of Follow Me Cloud," *IEEE Access Mag.*, vol 2, Oct. 2014, pp. 1340–47.

[8] White paper of IMT vision towards 2020 and Beyond, IMT-2020 (5G) Promotion Group, China, 2014.

[9] S. Sezer *et al.*, "Are We Ready for SDN? Implementation Challenges for Software-Defined Networks," *IEEE Commun. Mag.*, vol. 51, no. 7, July 2013, pp. 36–43.

[10] K. Zheng *et al.*, "Radio Resource Allocation in LTE-Advanced Cellular Networks with M2M Communications," *IEEE Commun. Mag.*, vol. 50, no. 7, 2012, pp. 184–92.

[11] L. Lei *et al.*, "Operator Controlled Device-to-Device Communications in LTE-Advanced networks,"*IEEE Wireless Commun.*, vol. 19, no. 3, June 2012, pp. 96–104.

[12] 3GPP TR 36.819, "Coordinated Multi-Point Operation for LTE Physical Layer Aspects," V11.0.0, 2011-12.

[13] 3GPP TR 36.815, "LTE-Advanced Feasibility Studies," V9.1.0, 2010-06.

[14] 3GPP TR 37.870, "Study on Multi-RAT joint coordination(Release 13)," V1.0.0, 2015-02.

[15] ETSI, "Network Functions çirtualization (NFV), Achitectual Framework," V1.1.1, 2013-10.

Biographies

Jianquan Wang (jianquanwang2015@163.com) received the Ph.D degree from Beijing University of Posts and Telecommunications (BUPT) in 2003. His research interests include network technology and next generation mobile network. He is currently working in the network construction department of China Unicom, focusing on the deployment of HSPA+ and LTE network, responsible for mobile network technology strategies.

Zhaobiao Lv received the Ph.D degree from Beijing University of Posts and Telecommunications (BUPT) in 2006. His research interests include next generation wireless techniques network and 4G/5G service innovations. He is currently working for Guangdong company of China Unicom, responsible for new service research and development.

Zhangchao Ma received the B.S. and Ph.D degree from BUPT in 2006 and 2011, respectively. His focus is on LTE and UMTS technologies, standards and architecture evolution. He is currently working in the network construction department at China Unicom.

Lei Sun received the B.S. and Ph.D. degree from BUPT in 2006 and 2011, respectively. He is currently working in the network construction department at China Unicom. His research interests lie in the field of mobile communication technologies.

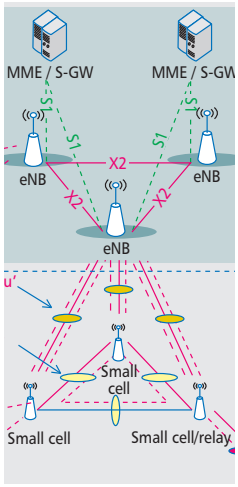
Yu Sheng received the Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT) in 2009. His research interests include LTE and HSPA+ wireless networking technology and next generation mobile network evolution. He is currently working in the network research institute of China Unicom.

There is still much research needed to fully exploit the potential of i-Net. For instance, the current cell-edge user throughput, which is improved with i-Net inter-site CoMP, still has a large gap between the visions of " no edge" consistent experience.

Virtual RATs and a Flexible and Tailored Radio Access Network Evolving to 5G

The authors describe a new method for radio access technologies called virtual RATs. The proposed scheme can provide a flexible and tailored access network that can meet the requirements and overcome the challenges in mobile broadband networks. They also present a high level design of the overall architecture and protocol stack for virtual RATs.

Shanzhi Chen, Jian Zhao, Ming Ai, Dake Liu, and Ying Peng



Abstract

This article first analyzes the requirements and challenges in the future 5G mobile wireless network. Then the authors describe a new method for radio access technologies called virtual RATs. The proposed scheme can provide a flexible and tailored access network that can meet the requirements and overcome the challenges in mobile broadband networks. The authors also present a high level design of the overall architecture and protocol stack for virtual RATs. Two concepts, virtual RAT Types and interface sets, are introduced. Two essential features, flexible control/user (C/U) plane separation and coordination between virtual RATs, are also discussed. Examples are provided to show how to achieve a flexible and tailored access network for services by using virtual RATs. Finally, the capability of hardware to support implementation of virtual RATs is analyzed.

Introduction

Nowadays, the dramatic growth of mobile data services, driven by the wireless Internet, the Internet of Things (IoT), mobile cloud, and smart devices, is presenting more challenges to radio access network design for the fifth generation (5G) mobile communication systems. Different people may have different views about the future radio access network. However, one common idea has been widely accepted, i.e., that multiple RATs will simultaneously exist in 5G [1], meaning there will be more new RATs to be developed in order to satisfy the increasing requirements of different user cases. For example, one new RAT might be developed to support the high data rate of 10 Gbps or above, while another new RAT might be designed to support machine to machine (M2M) services with a large number of users and low data rate. Subsequently, the issue arises of how to efficiently design those different RATs, and how to define and deal with various RATs will be an inportant factor when designing the architecture of the future access network. In this article we categorize three evolving phases for various RATs according to the level of inter-working efficiency in a heterogonous radio network: stand-alone RATs, integrated RATs, and virtual RATs. Then, the high level design of virtual RATs is investigated to provide a flexible and tailored network in the future.

The remainder of this article is organized as follows. We analyze future requirements and challenges and indicate that virtual RATs will be needed to meet those requirements and overcome those challenges. We present a virtual RATs design at a high level of the overall architecture and protocol stack by introducing two concepts: virtual RATs types and interface sets. Two other essential features, flexible C/U plane separation and coordination between virtual RATs, are also discussed. Examples are provided to show how to achieve flexible and tailored access networks for services by using virtual RATs. We then analyze the capability of hardware to support implementation of virtual RATs.

COMMUNICATIONS STANDARDS

Future Challenges and the Concept of Virtual RATs

There will be many more new services and applications in future networks. New applications such as the Internet of Things, social networks, mobile cloud, and public safety will be efficiently provided by radio access networks. Those new applications have different requirements than those of traditional mobile services. For example, the number of subscribers for IoT will increase 100 to 500 times while power consumption will decrease several ten times than that of normal mobile devices. Social network applications may request direct communications between terminals instead of communicating through a core network. Mobile cloud services will depend on huge amounts of information being exchanged between core networks and terminals with very low transport latency. Public safety will require the ability to communicate in emergency situations, especially in natural disasters such as earthquakes, when the public mobile communication systems are down. A method of multiple radio access technologies (multi-RATs) is a possible way to support multiple services. Actually, multi-RATs such as 3G/4G and WiFi have already been used to provide voice and data services in the current network under different scenarios. However, there are still new challenges presented by multi-RATs design. First, many rich and diverse services need to be supported by as few RATS as possible, to minimize the cost and complexity of devices. Second, seamless handover and dynamic offloading between different RATs needs to be supported to provide a high quality of experience (QoE) for users. Third, sharing and optimization of radio access resources between different RATs is needed to achieve higher spectrum efficiency as spectrum resources are becoming scarce and the data rate is increasing. Besides, the installation and cost for multiple networks of multi-RATs need to be significantly reduced. All of those challenges are

Shanzhi Chen, Jian Zhao, Ming Ai, and Ying Peng are with China Academy of Telecommunications Technology (CATT) and Datang Telecom Technology & Industry Group.

Dake Liu is with Beijing Institute of Technology (BIT).

0163-6804/15/$25.00 © 2015 IEEE

difficult to handle by using traditional inter-working methods of multi-RATs.

The first traditional inter-working method of multi-RATs is through stand-alone RATs. By using this method, different RATs work alone and provide different services. Each RAT keeps its protocol stack, and runs in each software and hardware platform. There is no information exchange between different RATs. The switch between different RATs is performed by manual operation. It is obvious that stand-alone RATs cannot meet the requirements of multi-RATs.

The second traditional inter-working method is integrated RATs, in which different RATs keep their respective architecture and protocol stack running in each software and hardware platform. However, integrated RATs can have some coordination in the radio access and the core network. The switch between different RATs can be executed automatically. One typical example is the case of 3GPP-WLAN interworking. Several solutions for WLAN offloading based on dual connectivity are defined to offload the traffic between 3GPP and WLANs seamlessly in 3GPP Release 12 [2]. Recently, another work item (WI) for LTE-WLAN Radio Level Integration and Interworking Enhancement was approved in 3GPP Release 13. It allows for real-time channel and load aware radio resource management across WLAN and LTE. 3GPP Release 13 can thus provide significant capacity and QoE improvements [3]. Although the integrated RATs method offers significant improvements over the stand-alone RATs method, it is still not able to deal with the challenges that we discussed above. Integrated RATs does not combine different RATs at the software and hardware levels, and thus cannot achieve the low cost and low complexity of devices, especially when more RATs are needed to provide richer services and applications. The interworking in the Packet Data Convergence Protocol (PDCP) has limitations to achieve higher spectrum efficiency.

Therefore, we propose a new multi-RATs method called virtual RATs. In virtual RATs, a common architecture and protocol stack are based on the 3GPP network to achieve a smooth evolution from legacy networks. The architecture and protocol stack for different RATs can be tailored or revised from the common architecture and protocol stack. Virtual RATs can be programmed after that. The information can be exchanged at each level to optimize the utilization of radio resources and to improve spectrum efficiency. Different RATs may work in a common software and hardware platform. However, there are differences between virtual RATs and the open source software method. By introducing the concepts of interface sets and virtual RAT types, virtual RATs can gain more advantages in security and interworking than the open source software method. Comparisons are shown in Table 1.

	Stand-alone RATs	Integrated RATs	Virtual RATs
Architecture and protocol stack	Different	Different	Common, flexible, and tailored
Hardware platform	Independent	Different	Common
Coordination and interworking between different RATs	No	Automatically, in the core network	Automatically, in the core network, L2 layer, and physical layer
Information exchange	No	Limited	Optimize
Cost	High	Medium	Low
Complexity	Low	High	Medium
QoE	Low	Medium	High

Table 1. Comparisons among stand-alone RATs, integrated RATs, and virtual RATs.

Virtual RATs Design for Future Radio Access Networks

In this section we present the high level design of Virtual RATs, including the overall architecture design by introducing a concept of interface sets, and the protocol stack design by introducing virtual RAT types. Meanwhile, the differences between the virtual RATs design and the open source software RAT design are also pointed out. In addition, we discuss two other important topics: flexible C/U plane separations and the coordination between RAT types, which are closely related to the virtual RATs design.

A Flexible and Tailored Overall Architecture Design by using Interface Sets

In general, an overall architecture design at the high level for a mobile system mainly consists of the definition of network components and the definition of interfaces between the network components.

Different kinds of low power nodes such as small cell, relay, and device-to-device (D2D) will appear beside traditional macro cell and micro cell [4]. Although those components of low power nodes have been heavily discussed in 3GPP, they are not explicitly included in the current overall architecture that is shown in Part 1 in Fig. 1 [5]. The reason is that the legacy mobile system such as the second generation (2G), 3G, and even the early version of 4G are mainly focussed on macro-coverage based design [6]. In other words, the air interface technologies have always been kept the same in order to keep the compatibility for both the macro-coverage scenario and the indoor/hotspot coverage scenario. So we called Part 1 the architecture for the macro-based International Mobile Telecommunications (IMT) path. It is obvious that the architecture in Part 1 is not ribust enough to be the architecture for future radio access networks. As analyzed in [7], low power nodes will play a very important role in the future network to meet the new requirements in traffic volume, frequency efficiency, energy, cost, and so on. Since there are significant differences between the indoor/hotspot scenario and the macro scenario, it is reasonable that an additional part of the architecture, for low power nodes used in

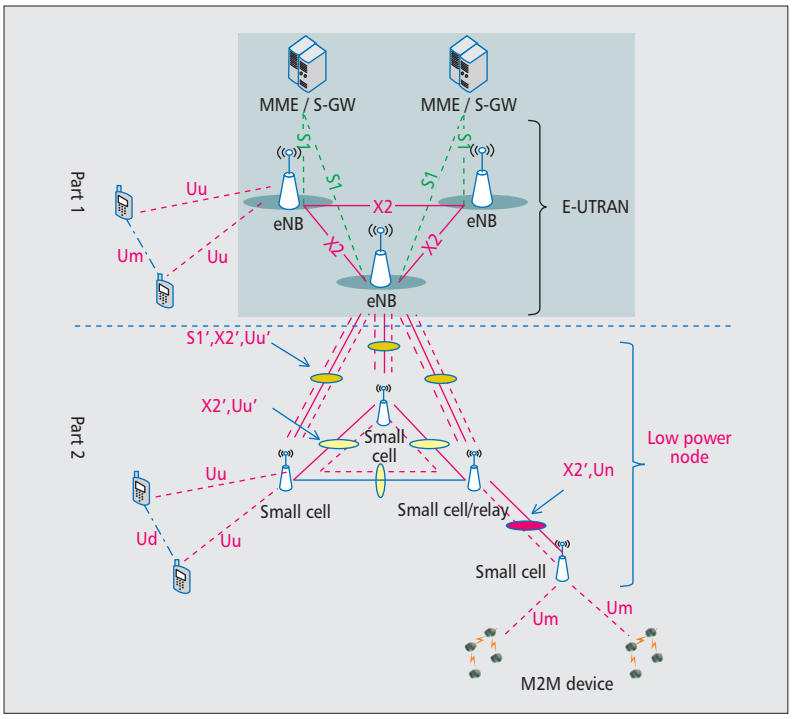


Figure 1. Overall architecture for of future radio access network.

indoor/ hotspot scenario, needs to be introduced in the overall architecture for future radio access networks. This additional part of the architecture we proposed for low power nodes is shown as Part 2 in Fig. 1.

The introduction of the architecture for low power nodes may raise another issue, i.e., how to define the interfaces between low power nodes as well as between low power nodes and macro cell base stations. The interface definitions always have difficulties in practice since several key requirements should be taken into account together including flexibility, programming, inter-working, security, system performance, and the cost to upgrade from legacy architectures.

To deal with those difficulties, we put forward the concept of "interface sets" in future radio access networks (especially for 5G). "Interface sets" is a set of interfaces that might be used in two peered entities. Each interface in the set can be programmed for one scenario and the interfaces can be selected and downloaded to run correspondingly to the actual scenario. Each interface in the set needs to be validated by a third party to provide inter-working between node and node, node and terminal. The interface is designed based on existing interfaces. We define three interface sets in Fig. 1: an interface set of "S1', X2', Uu'" is between an E-UTRAN Node B (eNB) and a small cell; an interface set of "X2', Uu'" is between a small cell and a small cell; and an interface set of "X2', Un'" is between a small cell and a relay. By introducing "interface sets," an optimized balance can be achieved to improve the flexibility, programmable, security, system performance, and updating cost. Since an interface can be selected according to an actual scenario and a new interface can be added by programming according to a new scenario, a fairly high degree of flexibility is achieved. For example, an interface set "S1', X2', Uu'" between an eNB and a small cell can support different kinds of backhaul methods for a small cell. S1' can be selected when wired backhaul is used. Uu' can be selected when wireless backhaul is used. X2' is used while requiring a direct backhaul between a small cell and a core network. More important, security can be guaranteed because the programming of the interface set can only be done by operators. Validation of the programming of interface sets may be needed to maintain inter-working between the nodes and devices from different manufactures. In addition, S1', X2', Uu' are defined respectively by revising them from the existing interfaces of S1, X2 and Uu, so the impact to the current architecture can be minimized.

A Flexible Protocol Stack Design for Radio Access Network by Using Virtual RAT Types

As we have discussed before, having one RAT provide one kind of service is not feasible when many more kinds of services need to be supported in future radio access networks. The best scheme is to support multiple kinds of services by using one common protocol stack, with information exchanged easily between each peered layer to achieve high QoE and efficiency.

Thus, we introduce the concept of virtual RAT types to achieve this goal. As shown in Fig. 2, one common protocol stack based on the 3GPP network is used which consists of the physical layer (PHY), the media control layer (MAC), radio link control (RLC), and radio resource control (RRC). The protocol stack for one RAT is revised from the common protocol stack and is programmed to one software package. The created software packages, for different RATs to support different kinds of services, run in the same hardware platform. Thus, the cost and integration complexity of devices can be reduced by using one hardware platform; and owing to software-based programming, the information exchange between peered layers can be easily achieved to support seamless handover and dynamic offloading between different RATs. In addition, the radio resource for all RATs can be managed by single self organizing networks (SON). So the optimization of radio access resources between different RATs is also available to improve spectrum efficiency.

Although virtual RAT types are software-defined, they are different from common open programming. First, virtual RAT types are revised from one common protocol stack based on the 3GPP network. Second, the air interface corresponding to one virtual RAT type is always selected from the interface sets we discussed above. So the virtual RAT types also receive the benefits of inter-working and security just as in the interface sets.

In Fig. 2 we illustrate an example of implementing two virtual RAT types in the same hardware with a single SON controller. The differences between virtual RATs and standalone RATs and integrated RATs are also shown.

Different from Open Source Software Design

Programmable technology seems to be the promising trend for the development of future radio access networks. The designs of interface

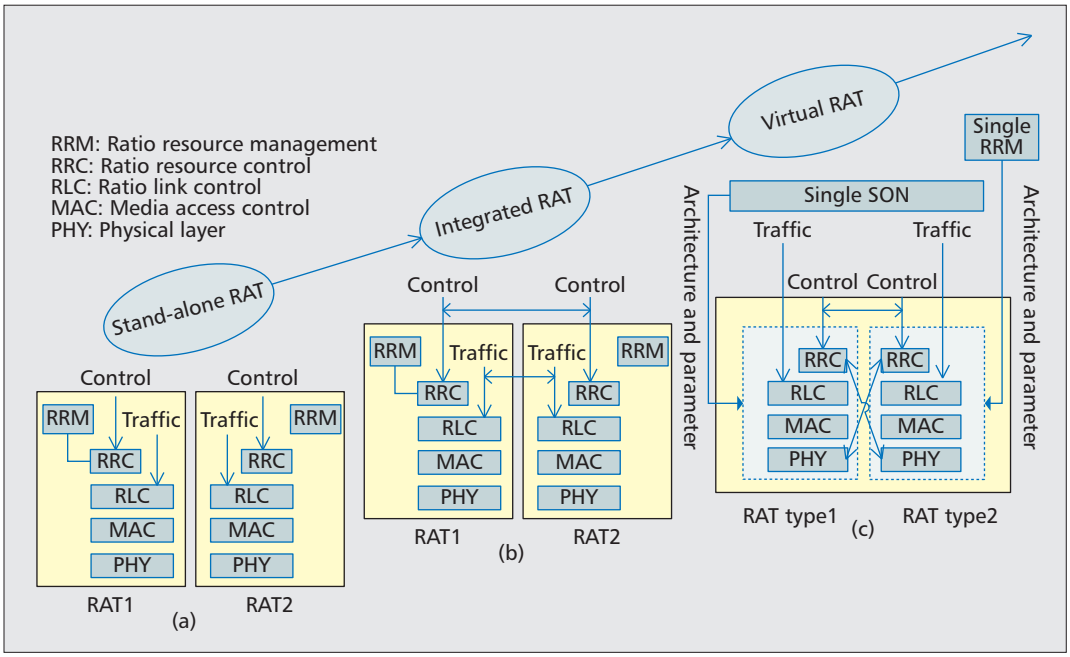


Figure 2. Illustration of three kinds of multi-RATs.

The validation of the design for Interface Set and Virtual RAT Types will provide good inter-working between the eNB and devices from different manufactures. In addition, the design for Interface Set and Virtual RAT Types are based on the re-use of the interface and RAT architecture of the legacy network.

sets and virtual RAT types are both software-based programmable designs, which offers the benefits of flexibility. However, our programmable RAT is different from open source software. Open source software offers freedom to freely use, change, and share the software. Third party developers may develop different radio systems according to different applications through programming on our platform. Open source software may have critical difficulties in security, inter-working, system performance, and updating cost. For example, open source software in radio architectures may facilitate the development of malicious masquerade eNBs and cause trouble for mobile users. Open source software might decrease system performance while optimizing for an individual user. On the other hand, constraints shall be set for the design of programmable interface sets and virtual RAT types, so benefits such as flexibility from programmable methods are achievable while the drawbacks from programmable methods are avoided. For example, the design for interface sets and virtual RAT types shall be manageable by operators or other administration supervisors to guarantee security. The validation of the design for interface sets and virtual RAT types will provide good inter-working between the eNB and devices from different manufactures. In addition, the design for interface set and virtual RAT types are based on the re-use of the interface and RAT architecture of the legacy network. Thus, the updating cost will be lower. The design for interface sets and virtual RAT types achieve good balance among programmable method, security, inter-working, system performance, and updating cost.

Flexible C/U Plane Separations and Combinations

Flexible C/U plane separations and combinations is an important feature in future networks, and has a very close relationship with virtual RATs. Flexible C/U plane separations and combinations are necessary in order for virtual RATs design to accomplish the real flexible C/U plane separations and combinations. In current radio networks, the C plane and the U plane can be separated only between one terminal and one eNB, as shown in the cases of user equipment (UE) 1 and UE 4 in Fig. 3. This method is suitable for basic mobile services such as voice, moderate data services, and machine to machine (M2M) services. In order to support more complicated cases such as high mobility and high data rate mobile services in heterogeneous network, more flexible C/U separations and combinations need to be introduced. For example, as shown the case of UE 2 in Fig. 3, the C plane is connected with a macro eNB while the U plane is connected with a small cell. The advantage is that the high data rate can be offloaded to the small cell while the good mobility is kept by the C plane connection to the macro cell. The related research work on this kind of separation has been ongoing in 3GPP. Furthermore, a wide separation can be supported for cases between one terminal and multiple stations crossing different RATs, as shown in the case of UE 3 in Fig. 3. Thus, capacity is provided by one RAT and mobility is provided by another RAT for one terminal. With this approach, flexible separations and combinations of the C/U plane will be supported, and different kinds of C/U planes can be selected according to the service case and the terminal's ability.

Coordination between Virtual RATs

Coordination between the different RATs is very useful to provide a high QoE and good user experience as well as to improve spectrum efficiency by optimizing utilization of radio access resources.

There are rich coordination paths for virtual RATs, including in the core network, the data link layer (L2 layer), and the physical layer. Furthermore, coordination in radio resource management (RRM) is also feasible by using a

Virtual RATs can achieve carry aggregation, interference coordination, and mitigation between different RATs. System performance can be significantly improved, and a new scheme for RF coordination named Centralized, Cooperative, Cloud and C-RAN is under investigation in the Next Generation Mobile Network.

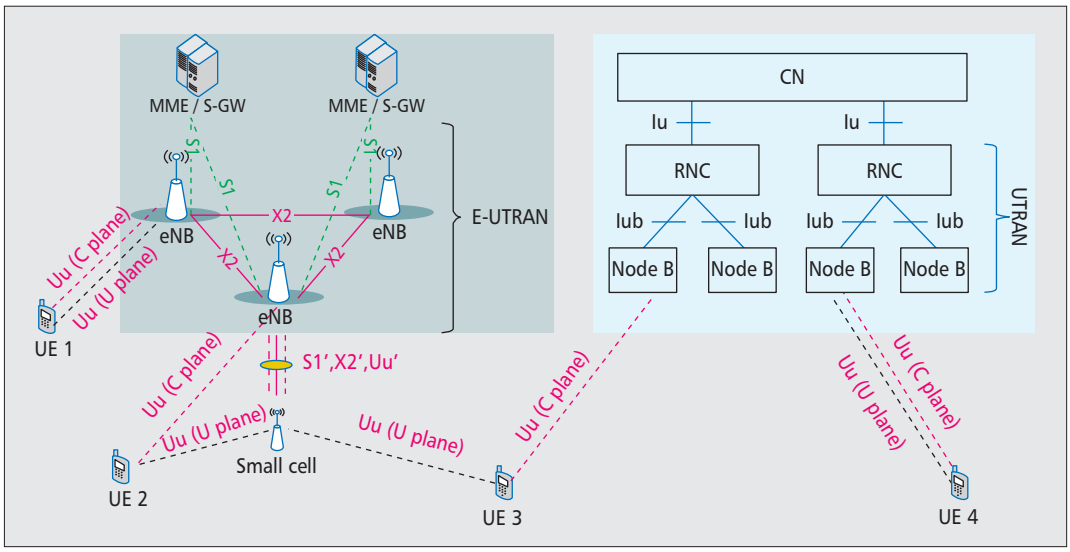


Figure 3. Examples of C/U flexible separations and combinations.

converged SON for different virtual RATs. Before Release 12, 3GPP mainly focused on the standardization in the core network limited by the integrated method of multi-RATs [8–10]. Owing to coordination in the L2 layer, virtual RATs can maximize radio resource efficiency and minimize interference and delay. Owing to coordination in the physical layer, virtual RATs can achieve carry aggregation, interference coordination, and mitigation between different RATs. System performance can be significantly improved, and a new scheme for RF coordination named centralized, cooperative, cloud and clean RAN (C-RAN) is under investigation in the Next Generation Mobile Network (NGMN) [11].

Examples of Using Virtual RATs

We can achieve the high level design of the overall architecture and protocol stack for virtual RATs to support typical new services in the future in a flexible and tailored way. Some examples are given as followed.

Figure 4a presents an overall architecture and protocol stack for D2D services. In this case, two different RAT types are tailored. One is the D2D data virtual RAT type transmitting the D2D data between the two UEs. Another is the normal cellular virtual RAT type providing the cellular traffic. The D2D virtual RAT type can be tailored to be the same with the U plane of the Uu interface of cellular virtual RAT type below the RLC layer so that the software and hardware resources in both the terminal and the eNB can be shared.

Figure 4b gives an illustration of the overall architecture and the protocol architecture for a group call to be used when the public cellular network has been damaged in a disaster. An emergency virtual RAT type can be tailored to terminals to accomplish direct communications between terminals nearby by removing PDCP, and simplifying RRC, as well as adding group call control (GCC) with gateway functionalities. For temporary eNBs, both the emergency virtual RAT type and the normal Virtual RAT type will be tailored to provide access between terminals and terminals. A terminal can act as a temporary eNB such as UE 3.

Figure 4c gives an overall architecture and protocol architecture for M2M services. In general, M2M services always have small data packets in a long idling period with low mobility and extremely low power consumption [12]. To meet those requirements of M2M, the architecture and protocol architecture need to be tailored to a large extent to efficiently meet such types of services. M2M access points will bring huge numbers of M2M subscribers to clusters, not directly to eNBs. PDCP and RRC modules can be removed to dramatically save the terminal's power and cost.

Implementation Feasibility in Hardware to Support Virtual RATs

Hardware flexibility for virtual RATs is essential. Most challenges will be in the physical layer, including the RF and baseband. Above the physical layer, the implementation of protocol stacks and interfaces are already based on software and can be adapted to future virtual RATs.

The first challenge in the eNB physical layer is the cost of the RF radio frequency transceiver (TRx). To support multiple RATs in different RF bands, we should design a RF TRx with different bandwidth, dynamic range, sensitivities, and output power levels. For each band, we need many passive RF components for filtering, switching, matching, and connections. By integrating more RATs (more bands), the cost of passive components increases rapidly and dominates the total cost of an eNB. To reduce the cost, we need integration technologies for passive RF components.

The second challenge is the design complexity of the digital base band (DBB) integrated circuit (IC) in an eNB for multiple RATs. Total computing cost is more than $2x10^{11}$ operations per second for a full band LTE with single antenna in a sector of an eNB. An eNB usually consists of three sectors with four antennas in each sector. The DBB computing cost is thus more than three trillion

[1] https://www.opennetworking.org/

[2] http://www.etsi.org

[3] http://www.etsi.org/technologies-clusters/technologies/nfv

[4] https://www.ietf.org

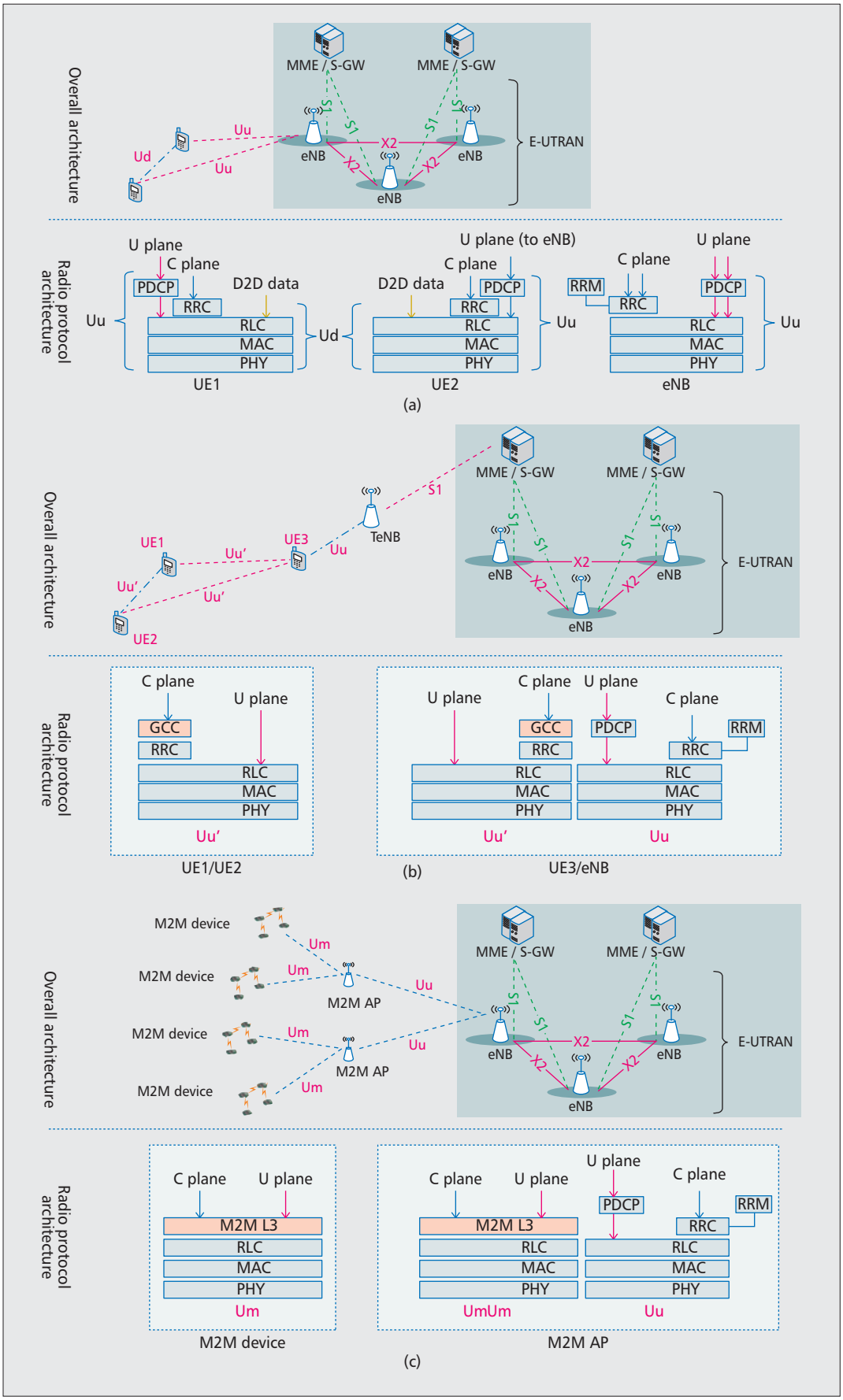


Figure 4. Examples of overall architecture and protocol: a) tailored overall and protocol architecture for D2D; b) tailored overall architecture and protocol for group call; (c) tailored overall architecture and protocol for M2M

Hardware flexibility for Virtual RATs is essential. Most challenges will be in the physical layer, including the RF and baseband. Above the physical layer, the implementation of protocol stacks and interfaces are already based on software and can be adapted to the future Virtual RATs.

Our solution provides a good tradeoff between flexibility, programmability, security, interworking, system performance optimization, and controlloing the cost of network upgrades. Meanwhile, improved hardware technologies will make it possible to implement our solution with reasonable cost and complexity in the future.

operations per second (TOPS). To meet the computing latency constraints, two GPU (GTX-590) modules are needed, each consuming 600 Watts with the cost of $1400. An application specific instruction-set processor (ASIP) for SDR will thus be a necessary technology offering both performance and flexibility with acceptable power [13]. To support baseband for OFDM (LTE and WLAN) and single carrier (GSM and CDMA), we need an efficient, high performance, and low power DBB ASIP. A DBB ASIP consists of three processor clusters: the symbol processor (for transformation, filtering, matrix, and function computing of complex/ real data); the FEC processor (forward error correction of decoding Turbo, LDPC, RS, and Viterbi); and the bit parallel processor (handling low latency bit level parallel processing). Symbol and FEC processors were on the market, there are early bit parallel ASIP prototypes currently available. The cost and power of ASIP DBB for volume products will thus be in the tens dollar range and on the tens Watts level.

The cost of silicon technologies and IC continues to decrease. The cost of passive components also can be decreased by the integration of passive components. It is predictable that the extra cost induced by virtual RATs will be acceptable in the next few years thanks to the use of new technologies and successful hardware sharing and integration.

Conclusions

In summary, new methods of multi-RATs need to be investigated in order to meet the new requirements and challenges of services in the future 5G networks. In this article we have presented a virtual RATs solution to support a flexible and tailored radio access network in the future that would provide high performance at low cost. Two concepts, virtual RATs types and interface sets, were put forward. Instead of building open source software, our solution emphasizes the requirements for security and inter-working. Two important features, flexible C/U plane separation and coordination between RATs, were discussed. Toward 5G, the evolving flexible overall architecture and its radio protocol are based on LTE and LTE-Advanced networks, in order to make a smooth evolution from legacy networks. Our solution provides a good tradeoff between flexibility, programmability, security, inter-working, system performance optimization, and controlloing the cost of network upgrades. Meanwhile, improved hardware technologies will make it possible to implement our solution with reasonable cost and complexity in the future.

Acknowledgment

This work was partly supported by the National Science and Technology Major Project (No. 2013ZX03001025-001) in China, and the National Natural Science Foundation of China for Distinguished Young Scholar under Grant 61425012.

References

[1] Afif Osseiran *et al.*, "Scenarios for 5G Mobile and Wireless Communications: The Vision of the METIS Project," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp 26–35.
[2] 3GPP Technical Specification 23.402: "Architecture Enhancements for Non-3GPP Accesses; (Release 12)."
[3] 3GPP New WI Proposal RP-150510: "LTE-WLAN Radio Level Integration and Interworking Enhancement," March 2015.
[4] L. Lei *et al.*, "Operator Controlled Device-to-Device Communications in LTE-Advanced Networks," *IEEE Wireless Commun.*, vol. 19, no. 3, June 2012, pp. 96–104.
[5] 3GPP Technical Specification 36.211, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2 (Release 12)."
[6] S. Chen *et al.*, "Technical Innovations Promoting Standard Evolution: From TD-SCDMA to TD-LTE and Beyond," *IEEE Wireless Commun.*, vol. 19, no. 2, Feb 2012, pp. 60–66.
[7] S. Chen and J. Zhao, "The Requirements, Challenges and Technologies for 5G of Terrestrial Mobile Telecommunication," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 36–43.
[8] 3GPP Technical Specification 23.327: "Mobility between 3GPP-WLAN Interworking and 3GPP Systems (Release 12)."
[9] 3GPP Technical Report 23.861: "Network based IP Flow Mobility (Release-13)."
[10] 3GPP Technical Specification 37.834, "Study on WLAN/3GPP Radio Interworking (Release 12)."
[11] http://www.ngmn.org/uploads/media/NGMN_CRAN_Suggestions_on_Potential_Solutions_to_CRAN.pdf.
[12] T. Taleb *et al.*, "Lightweight Mobile Core Networks for Machine Type Communications," *IEEE Access Mag.*, vol. 2, Oct. 2014, pp. 1128—37.
[13] D. Liu *et al.*, "Bridging Dream and Reality: Programmable Baseband Processor for Software-Defined Radio," *IEEE Commun. Mag.*, vol. 47, no. 9, Sept. 2009, pp. 134–40.

Biographies

Shanzhi Chen [SM'04] (chensz@datanggroup.cn) received his Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT), China, in 1997. He joined Datang Telecom Technology & Industry Group in 1994, and has served as CTO since 2008. He was a member of the steering expert group on information technology of the 863 Program of China from1999 to 2011. He received the Outstanding Young Researcher Award from the Nature Science Foundation of China in 2014. He is the director of State Key Laboratory of Wireless Mobile Communications, and a board member of the Semiconductor Manufacturing International Corporation (SMIC). Since 2004 he has devoted his work to the development of TD-SCDMA 3G and TD-LTE-advanced 4G. He received the State Science and Technology Progress Award in 2001 and 2012. His current research interests include network architectures, wireless mobile communications, Internet of Things (IoT), and emergency communications.

Jian Zhao (zhaojian@catt.cn) received his M.S. degree from Beijing University of Posts and Telecommunications (BUPT), China, in 1998. He joined Datang Telecom Technology Industry Group in 2002. From 2003 to 2008 he participated in 3GPP meetings as a standard manager of Datang. Since 2007 he has been acting in ITU meetings as a technical manager. Before joining Datang he worked at Huawei as a software R&D manager. His research interests include mobile communication technology, spectrum, and standards activities.

Ming Ai (aiming@catt.cn) joined Datang Telecom Technology Industry Group in 1998. Since 2008 he has been participating in 3GPP CT1 and SA2 meetings as a standard delegate and the co-coordinator of Datang. Before 2008 he worked as a software engineer and R&D manager for telecommunication equipment. His research interests include mobile communication technology, Internet technologies, and standards activities.

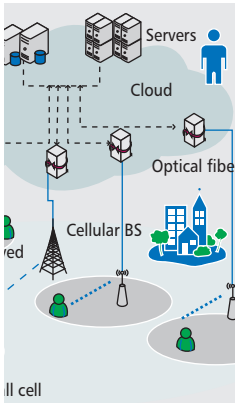
Dake Liu [SM'08](dake@bit.edu.cn) received the technology doctor degree from Linköping University, Sweden in 1995, He has been the director and a professor at the ASIP Lab (Application Specific Instruction-set Processor)of BIT, (Beijing Institute of Technology), China since 2010. He has also been a professor in the Dept. of EE (ISY), Linköping University, Sweden since 2001. He is a co-founder, and was the board director and chief scientist officer at Coresonic AB, Sweden from 2005 to 2012. He is a co-founder, and was the VP and chief engineering officer of FreeHandDSP AB, Sweden from 1999 to 2002. He was the board member and chief scientist officer of VIA Swedend AB from 2002 to 2005. He was a senior specialist of low power and communication IC at Ericsson Sweden from 1995 to 1998. He was an associate professor in the Dept. of Physics at Linköping University from 1998 to 2001. He was a lecturer of automatic control in the Dept. of EE, Beijing Jiaotong University from 1882 to 1990. Dake's research interests are ASIP for communications, high performance embedded computing, including architecture, parallel programming, and programming environments. Dake's research interests also include RF CMOS circuits and PA for radio communications.

Ying Peng (pengying@catt.cn) received her M.Sc. and Ph.D. in electrical and electronic engineering from the University of Bristol, U.K. She is presently a senior standardization researcher at the Wireless Mobile Innovation Center of the China Academy of Telecommunication Technology (CATT). She is a regular delegate in 3GPP RAN1 and ITU-R WP-5D. Her research interests include 5G techniques, 3GPP/ITU standardization, LTE/LTE-A physical layer design, D2D–CoMP, heterogeneous networks, and cognitive radio. She received First Prize on "4G TD-LTE-Advanced" of the Science and Technology Prize from CCSA in 2011.

Cloud Assisted HetNets Toward 5G Wireless Networks

To support the massive growth of mobile data, a large number of small cells are expected to be deployed, giving rise to HetNets, considered to be the key path toward 5G. With such large-scale HetNets, network operators face many serious challenges in terms of operation and management, cost-effective small cell deployment, and intercell interference mitigation. To deal with those issues, the authors introduce a cloud based platform designed to simplify the deployment, operation, and management, and facilitate round-the-clock optimization of the network, to pave the way for the development of 5G.

Ning Zhang, Nan Cheng, Amila Tharaperiya Gamage, Kuan Zhang, Jon W. Mark, and Xuemin Shen



The authors are with the University of Waterloo.

Abstract

With the proliferation of connected devices and emerging data-hungry applications, the volume of mobile data traffic is predicted to have a 1000-fold growth by the year 2020. To address the challenge of this data explosion, industry and academia have initiated research and development of 5G wireless networks, which are envisaged to cater to the massive data traffic volume, while providing ubiquitous connectivity and supporting diverse applications with different quality of service (QoS) requirements. To support the expected massive growth of mobile data, a large number of small cells are expected to be deployed indoors and outdoors, giving rise to heterogeneous networks (HetNets), which are considered to be the key path toward 5G. With such large-scale HetNets, network operators face many serious challenges in terms of operation and management, cost-effective small cell deployment, and intercell interference mitigation. To deal with those issues, a cloud based platform is introduced, aiming to simplify the deployment, operation and management, and facilitate round-the-clock optimization of the network, to pave the way for the development of 5G. Two case studies are provided to illustrate the benefits of the cloud based architecture. Finally, the related standardization activities are provided and some research topics essential for a successful development of 5G are discussed.

COMMUNICATIONS STANDARDS

Introduction

There has been an explosive growth in mobile data traffic volume, which almost doubles every year. According to the Cisco Visual Networking Index (VNI) [1], mobile data traffic volume is expected to grow at a compound annual growth rate (CAGR) of 61 percent from 2013 through 2018. Furthermore, it is predicted that data traffic volume will continuously grow up to 1000 times its value in 2010 by 2020, referred to as the 1000x data challenge.

The main factors accounting for the significant growth in mobile data traffic are two-fold: the proliferation of devices and the emergence of data-hungry applications. It is predicted by Qualcomm that 25 billion devices will be connected by 2020, due to the ever increasing use of smart phones and the promise of machine-to-machine (M2M) communication, Internet of things (IoT), and Internet of Vehicles (IoV) [2]. Data-hungry applications such as high resolution image transfer and video streaming will furher add to the volume. A standard definition and a high definition movie have 2.49 GB and 5.93 GB of data, respectively. Moreover, diverse types of applications have various quality of service (QoS) requirements. For instance, a high-definition movie transfer requires a very high data rate, while a safety message exchanged by a connected car requires a very low latency. Consequently, wireless networks are facing great challenges to accommodate such a large volume of data and meet the users' service demands [3].

The 5G wireless network promises to address the 1000x data challenge and handle a multitude of applications with diverse QoS requirements. It is expected to be a fully connected network and provide different QoS provisioned services to anyone and anything, at anytime and anywhere. It requires 1000x of 4G capacity, latency as low as 1 millisecond, and high energy and cost efficiency. Due to the urgency and importance of 5G, industry and government agencies initiated the research worldwide to facilitate the development of 5G. For instance, the European Commission pledged 50 million Euros in 2013 to fund 5G research, and a project called METIS has started to establish the foundation of 5G.

To develop 5G, efforts have been made in three areas: spectrum expansion, spectrum efficiency enhancement, and network densification. Network densification is considered to be the paramount and dominant approach to address the data challenge. It can be achieved by deploying a large number of small cells, such as microcell, picocell, femtocell, relay nodes, and Wi-Fi access points, which are low-powered radio access nodes and have smaller coverage areas compared with marcocells. Due to the dense deployment of small cells on a very large scale, the increased level of complexity, coupled with reduced control over applications and data, network operators are facing great challenges:

- Difficult to manage and control.
- Costly to deploy small cells densely.
- Intercell interference.

Those factors affect and limit the development of 5G.

In this article, to overcome the above issues, a cloud based platform is introduced, where small cells are connected to the cloud for management and control. With the cloud based platform, the operation and maintenance can be easily performed. Furthermore, centralized management facilitates round-the-clock optimization of the network, allowing the network operator to efficiently use network resources to support different applications while adapting to the changing service demands. Three main applications are discussed: operation and maintenance, a cloud assisted business model for small cell

 0163-6804/14/$25.00

Compared with conventional cellular communication, only half of the resources are required in direct communication. As a result, it can significantly improve the spectral efficiency. Moreover, transmission power can be saved since the communication is carried out between two adjacent nodes.

deployment, and cloud assisted interference coordination. Two case studies are provided, where the cloud is utilized for deployment and interference mitigation. Finally, the related standardization activities and some research topics are provided.

The remainder of this article is organized as follows. The evolution of 5G is presented. The main challenges to achieve 5G are discussed. The cloud platform for 5G is introduced, followed by the case studies. Standardization progress is provided, followed by the conclusion and future works.

Evolution of 5G

To meet the tremendous demand for data, three approaches can be considered: spectrum expansion, spectrum efficiency enhancement, and network densification.

Spectrum Expansion

To meet the expected explosive growth in data traffic and the diverse QoS requirements, it is necessary to exploit more spectrum by means of spectrum expansion. Specifically, the low spectrum bands with good penetration and signal propagation properties, e.g., the TV whitespace ranging from 572–698 MHz around 700 MHz, can be utilized to improve the building penetration and provide improved coverage for connectivity. The standards IEEE 802.22 and IEEE 802.11af specify how to exploit TV whitespace. In contrast, spectrum in the high frequency bands have larger bandwidths and can support higher data rates, such as spectrum around 3.5 GHz or millimeter wave frequency spectrum around 60 GHz. Moreover, because of the large path loss, high spectrum bands are well suited to small cells with a short communication range.

Spectrum Efficiency Enhancement

Network capacity can also be enhanced by improving spectrum efficiency. In this regard, various wireless technologies have been continuously evolving to improve spectrum efficiency.

Massive MIMO: Massive MIMO is typically comprised of a few hundred inexpensive antenna components, which can focus transmission energy in certain directions and consequently increase throughput and save energy significantly. Moreover, it can also facilitate concurrent transmissions to serve multiple users at the same time. With massive MIMO, capacity can be increased by 10 times or more and the radiated energy efficiency can be improved in the order of 100 times. For implementation, a 100 × 100 massive MIMO testbed based on xTCA standards, namely TitanMIMO, has been developed by a company called Nutaq to enable practical 5G massive MIMO development.

Cognitive Radio: Cognitive radio (CR) has been considered as a powerful technique to increase spectrum efficiency by enabling unlicensed users to access unused spectrum opportunistically [4, 5]. Two main paradigms to efficiently utilize spectrum are spectrum sensing and spectrum database. For the former, unlicensed users sense the spectrum to detect the availability of channels before transmission, and access the channels only when idle. For the latter, unlicensed users can acquire the availability of channels through spectrum databases before accessing the channels. Google and Microsoft have launched spectrum database products for TV white spaces, enabling users to easily obtain the available TV channels for access.

Device-to-Device Communication: The emerging device-to-device (D2D) communication paradigm enables devices in close proximity to communicate with each other directly without sending data to the base station (BS) or the core network [6–8]. Compared with conventional cellular communication, only half of the resources are required in direct communication. As a result, it can significantly improve spectral efficiency. Moreover, transmission power can be saved since the communication is carried out between two adjacent nodes. In addition, D2D communication can reduce delay, which is favorable for latency-sensitive applications.

Network Densification

Compared with spectrum expansion and spectrum efficiency enhancement, network densification is considered to be the dominant approach to boost capacity [9]. It is achieved by densely deploying small cells, such as microcells, picocells, or femtocells, which bring access points closer to users so that network capacity can be dramatically increased and latency can be reduced. Those small cells are the access points with lower transmission power and smaller coverage areas, such as LightRadio from Alcatel-Lucent, LiquidRadio from Nokia Siemens Network, and AtomCell and LampSite from Huawei. Moreover, data traffic originated from indoor users can also be served by Wi-Fi, such as the next generation Wi-Fi 802.11ac, which expects to support multi-gigabit data transmission rates. Such multi-tier networks with a variety of radio access technologies are referred to as heterogeneous networks (HetNets).

For more details on the paths toward 5G, please refer to the [10].

Challenges for Dense Deployment of Small Cells

Although HetNets play a vital role for 5G, the excessive deployment of small cells will pose significant challenges for network operation and management, deployment, and so on.

Network Deployment

Non-technical users may have difficulty deploying their small cells. Furthermore, the inappropriate configuration or installation done by users may cause a negative impact on the existing systems. Additionally, even though users may deploy some small cells for their own service, the network operator still needs to deploy a large number of small cells, which is very challenging due to the following reasons:

- Network operators are on a tight budget since they are buckling under the strain of continuously adding network infrastructures.

- The large-scale deployment of small cells is costly to network operators, in terms of the operating and capital expenditure, such as site lease, installation costs, additional costs for electricity and backhaul[1], and operating expenses. Therefore, the issue of how to deploy the small cells densely in an easy and cost-efficient way needs to be carefully addressed.

Operation and Management

With a large number of small cells, network operation and management becomes very complex as it needs lots of human effort to install, configure, monitor, and maintain the small cells. Additionally, the traffic may change frequently, depending on the location and time, so it is difficult for the network operator to efficiently use the network resources by managing the small cells at different sites to adapt to the changing demands. Further, small cells can be switched On/Off or moved at any time by the users and they are beyond the reach of the network operators, making optimization and management of the network very challenging [11].

Intercell Interference Mitigation

As the network density increases, inter-cell interference is more likely to happen, which significantly limits the gain achieved by the densely deployed small cells. In HetNets, intercell interference can arise in different scenarios. For example, the macrocell may create severe interference to the users in the small cell and the users of macrocells might receive interfering signal from the adjacent small cells. The interference is reflected in a lower signal to interference plus noise ratio (SINR), which degrades the network performance and the user experience. Without mitigating interference among different cells, HetNets cannot be successfully deployed. One way to cope with intercell interference is spectrum splitting, where the whole spectrum is divided into different parts for different cells. However, this leads to inefficient spectrum usage. Intercell interference coordination (ICIC) is introduced in 3rd Generation Partnership Project (3GPP) release 8 to mitigate interference. It coordinates the network resources among different cells by exchanging messages. However, due to the unplanned deployment of small cells by users, it is hard to coordinate the small cells in that manner.

Cloud Assisted HetNets

Cloud computing is becoming increasingly important in today's business due to benefits such as greater flexibility, increased security, scalability, and low cost. To address the challenges stated in the previous section, we introduce a cloud based architecture for HetNets, which utilizes the cloud as the control and management plane of the network[2]. By doing so, the installation, monitoring, management, and upgrade can be easily performed. Furthermore, the centralized management facilitates the round-the-clock optimization of the network, which helps the network operator make efficient use of network resources to satisfy different QoS requirements and adapt to changing demands [13].

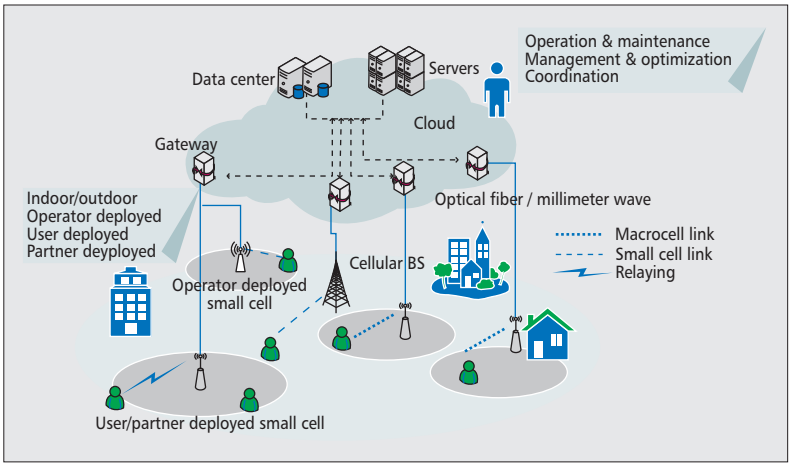


Figure 1. Cloud managed HetNets.

A similar concept is the cloud radio access network (C-RAN), which distributes radio head ends at different locations and pools the baseband processing elements in the cloud. The distributed radio head ends are only responsible for providing an air interface to users, while the cloud processes the users' data from the radio heads. Different from C-RAN, the cloud in the proposed architecture does not process the users' data and the small cells do not to send the users' data to the cloud for processing. Instead, it mainly focuses on operation, maintenance, and management of small cells to provide services by efficiently using network resources.

As shown in Fig. 1, the small cells are deployed densely in the network, either in indoor environments such as homes and offices, or in outdoor scenarios such as road intersections, squares, stadiums, etc. The network operator and the end users or third parties can deploy small cells with open and closed access modes, respectively. Small cells are connected to the cloud through high-speed optical fiber or mmWave. Monitoring, configuration, optimization, and mobility control is centralized in the cloud. The centralized nature also allows the network operators to manage the networks in a more efficient way. Since the traffic demand may dramatically vary, the cloud based architecture helps the network operator to allocate resources on demand and coordinate small cells efficiently to provide services with seamless coverage, high data rates, and low latency. The users or third parties can make some basic configuration changes to their own small cells from a web browser, such as changing the access mode and setting the data rate limit, etc. In summary, the cloud can bring the following advantages to the wireless networks.

Easy maintenance and management: The cloud streamlines provisioning, management, and troubleshooting of multiple sites from a single platform. It can provide services such as remote monitoring, real-time diagnostics, central configuration, and device management. With the cloud, administrators or users can access the network data at any time and anywhere from a web browser.

Scalability: The cloud assisted architecture can simplify the deployment by configuring small

[1] Optical fibers are usually required to connect all the small cell base stations, which is costly. An alternative is mmWave-based high-speed wireless backhaul.

[2] The feasibility of introducing the cloud to manage the wireless network can be verified by Cisco's Meraki, which is a cloud controlled WLAN [12].

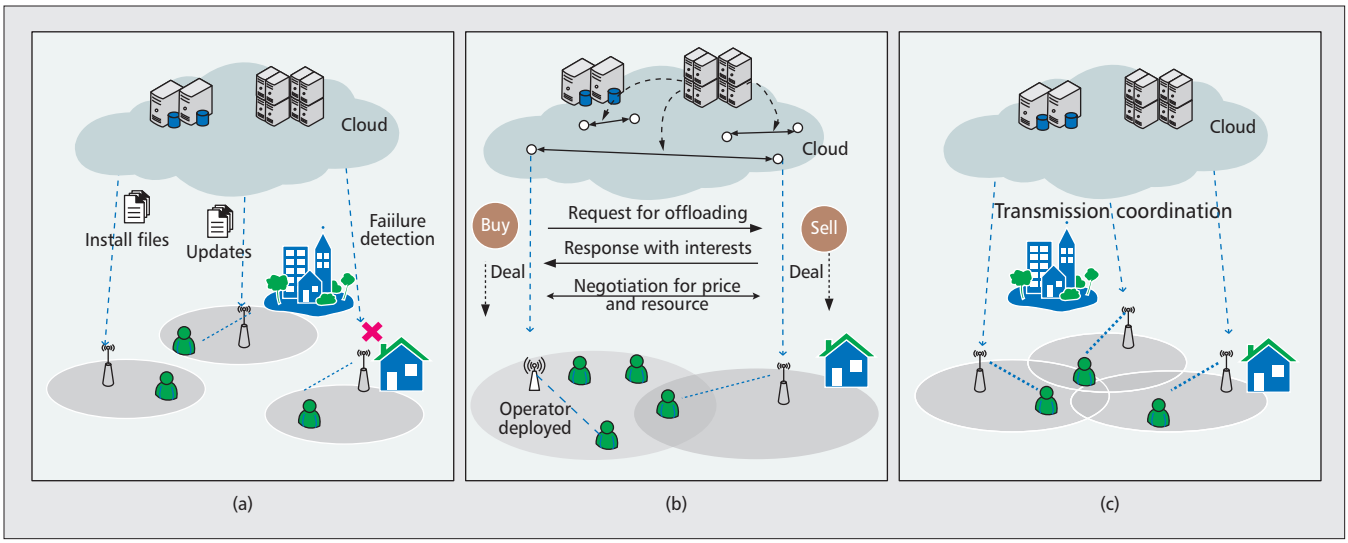


Figure 2. Typical cloud assisted applications in HetNets: a) operation and maintenance; b) cloud assisted business model for deployment; and c) cloud assisted interference mitigation.

cells through the cloud in order to facilitate a simple and quick installation of small cells when network expansions are required.

Efficiency: The cloud enables network operators to rapidly reallocate network resources to meet fluctuating and unpredictable user demands. For instance, in non-peak hours some small cells can be turned off and turned on again during peak hours through the cloud. Furthermore, the cloud helps optimize the performance of the network and intelligently balance the network load.

Low cost: The cloud can reduce the costs and burden to maintain the onsite management. With less need for people resources and more efficient management of network resources, the cloud can reduce the capital expenditure (CAPEX) and operational expenditure (OPEX) for deployment and operation of small cells.

With the cloud, many promising applications can be facilitated. We will now describe in detail three typical applications, as shown in Fig. 2.

Operation and Maintenance

As shown in Fig. 2a, the cloud based architecture provides the function of plug-and-play, where the small cells can be integrated into the existing network with minimal human involvement. When the small cell is initially powered, it automatically connects to the cloud and downloads necessary software and configuration data for installation and configuration. The cloud based architecture can also help monitor the status of small cells at different sites and automatically detect faults to reduce manual efforts. When there are some updates, the cloud can push the notification to the end users or automatically update the small cells. Moreover, it is easier to manage the small cells for the network operator and users or third parties. Users can have accounts in the cloud to change the setting of the small cells through a web browser such as enabling automatic update, setting maximum download/upload speed, changing access modes, etc.

Deployment

To meet traffic demand, the small cells need to be deployed densely. The expense for dense deployment will be huge for network operators. To facilitate the deployment of small cells, a cloud assisted business model is introduced, as shown in Fig. 2b. To encourage users or third parties (e.g., enterprises, facility owners, building proprietors, etc.) to deploy more small cells, the network operator can pay monetary rewards whenever the user-deployed or third party-deployed small cells provide service to the network operator. The users or third parties can set their small cells to be in open access mode for potential monetary rewards, allowing for providing better service to the customers of the network. For instance, those small cells can offload the traffic from the macrocell, or improve the transmission rate of the other users through relaying, etc. Since both the network operator and the users or third parties can receive benefits, it creates a win-win situation. The cloud can coordinate the service trading among different entities and keep the trading data (e.g., price, time, entities). Then the network operator grants the monetary rewards to the contributors based on the service data.

Interference Coordination

When the small cells are connected to the cloud, their transmissions can be coordinated through the cloud. This application can be regarded as an extension of coordinated multipoint transmission/reception (CoMP) included in 3GPP Release 11. Therefore, intercell interference mitigation can be facilitated, where the small cells are coordinated to transmit in different time slots or using different frequency bands, as shown in Fig. 2c. Moreover, the cooperation among the adjacent small cells can be enabled by the cloud, where joint transmission or joint beamforming can be leveraged to mitigate intercell inference. In other words, the small cells can serve a set of users simultaneously without interfering with each other through either precoding or interference cancelation schemes.

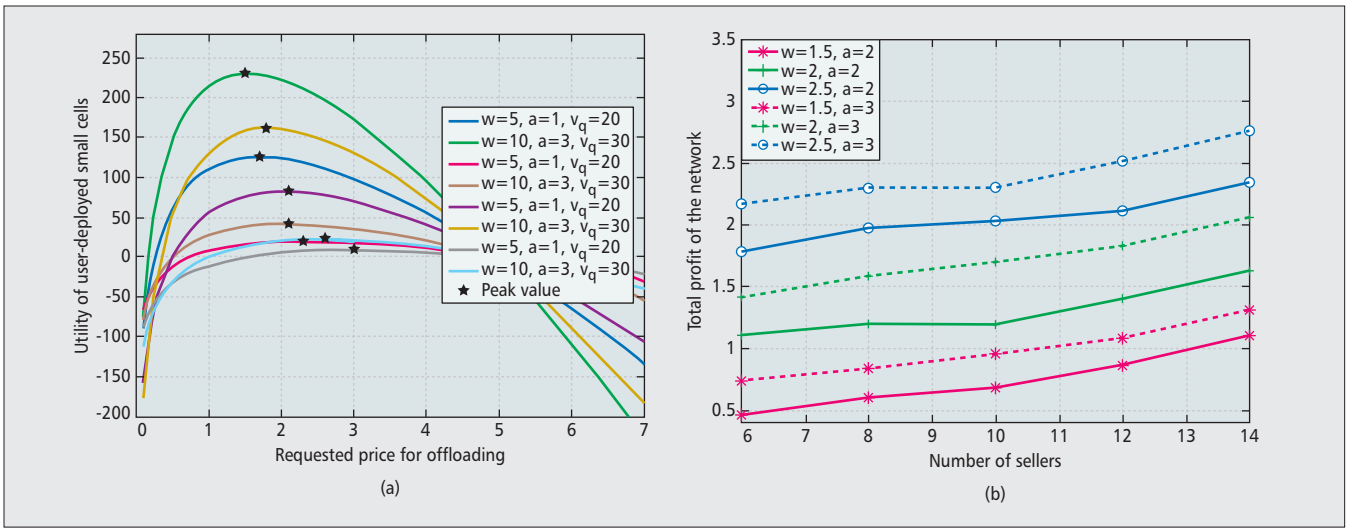


Figure 3. Cloud assisted business model for small cell deployment: a) utility of cooperative small cells versus the requested price; b) total profit of the network versus the number of cooperative small cells.

Case Studies

Cloud Assisted Business Model for Deployment

In what follows, we study a special case in which the network operator deployed cells are overloaded in specific locations or time slots. In such a case, the operator deployed cells intend to recruit some potential user deployed small cells in close proximity for offloading. In return, the network operator pays a certain amount of credits to the contributors. To negotiate the payment and offloading volume, a buyer and seller game is applied. For the operator deployed cell (the buyer), the utility function U_b is given by $U_b = \omega f(v) - p \cdot v$, where $f(v)$ is the satisfaction with respect to the offloading volume v, ω is the equivalent revenue per unit of satisfaction contributed to the overall utility, while p is the price for offloading a unit of traffic (e.g., 10 MB). Note that $f(v) = 1 - e^{-a \cdot v/v_q}$, where $a > 1$ is the satisfactory factor and v_q is the required offloading volume. The utility of the user deployed cell (the seller) is $U_s = p \cdot v - c \cdot v$, where c is the cost to offload a unit of traffic. The strategies for the buyer and seller are the offloading amount v and the unit price p, respectively. In other words, the user deployed small cell selects a price p, while the network deployed cell chooses the offloading volume v to maximize their own utilities.

The buyer-seller game can be analyzed by the backward induction method. First, for a given requested price p, the buyer chooses a suitable offloading volume to maximize its utility U_b, which is a function of p. Based on the result, the seller can then select the best price to maximize its utility U_s. After that, all the trading parameters can be determined, i.e., the price and the offloading volume.

In the network, there might exist multiple operator deployed cells requesting service and multiple user deployed small cells interested in earning credits. Note that the sellers might have different cost coefficients c and the buyers may have different offloading demands v_q. The cloud can facilitate the trading process to maximize the profits of the network, which is the summation of utilities of all sellers and buyers. Initially, the sellers and buyers send the cost coefficients and offloading demands to the cloud. Then the cloud matches the optimal pairs of the seller and buyer by generating a graph, where the sellers and buyers are the vertices and the profit summation of each pair of seller and buyer is the weight of each edge. To find the optimal pairs, the maximum weighted bipartite matching algorithm is applied. Finally, the selected pairs of buyers and sellers perform trading according to the negotiated parameters.

Figure 3a shows the utility of the small cell with respect to the price requested under different system parameters. It can be seen that there exists an optimal price such that the overall utility of the small cell can be maximized. For different values of a, v_q, and w, the optimal prices are adjusted accordingly by the small cell.

Figure 3b shows the total profit of the network versus the number of cooperative small cells. The number of buyers is set to six and the required offloading volumes are equally likely to be 100, 150, and 200 MB, respectively. The costs for the sellers are equally likely to be 0.2, 0.3, and 0.4. It can be seen that the network profit can be increased when more user deployed small cells participate in the trading, because more user deployed small cells can lead to more options to choose the best contributors.

Cloud Assisted Interference Mitigation

With the cloud based architecture, the transmission of small cells can be coordinated, opening the possibility of interference mitigation. In the following, we study the case where the cloud coordinates the transmission of the multiple neighboring small cells to mitigate interference. For instance, a set of small cells denoted by **S** are coordinated to serve the user A while creating no interference to other active users, denoted by the set **U**[3]. Specifically, each small cell broadcasts a weighted version of the message $w_i m$ to the destination user D, where w_i is the weight of small cell i. To maximize the through-

[3] Note that the number of coordinated small cells needs to be greater than the number of potential victim active users.

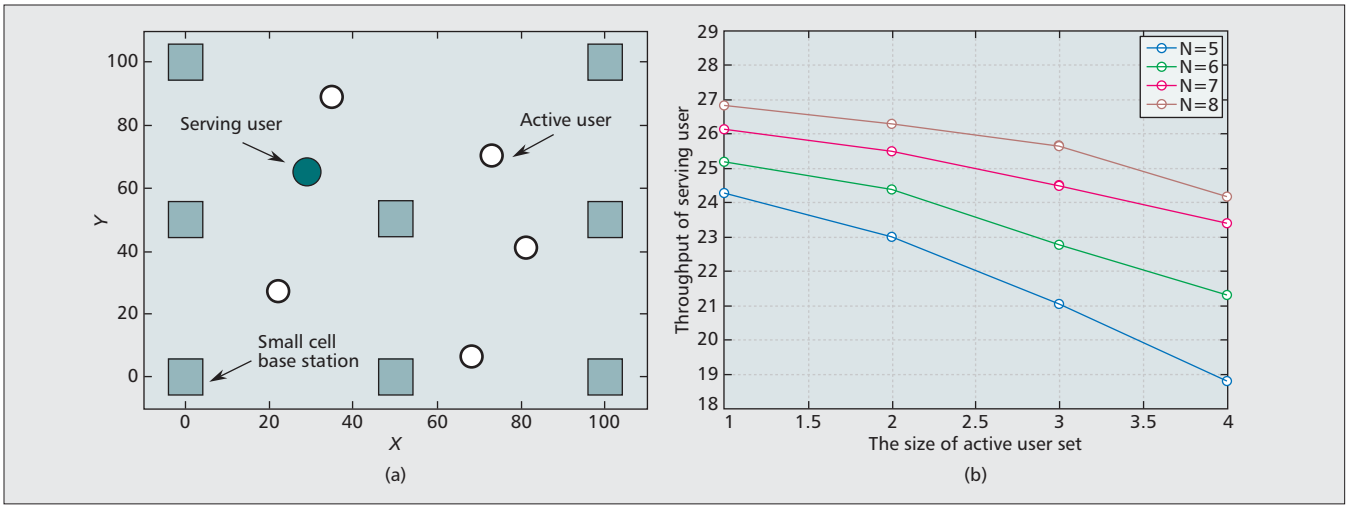


Figure 4. Cloud assisted interference mitigation: a) network topology for simulation; and b) throughput of the serving user versus the number of active users.

put of the serving user while protecting other active users in **U**, the cloud selects suitable weights for small cells. To accomplish this, $\mathbf{w}$ should be in the null space of $\mathbf{h}_{SU}^{\dagger}$ such that $\mathbf{h}_{SU}^{\dagger}\mathbf{w} = 0$, where $\mathbf{h}_{SU}$ is the channel matrix from small cells to the users in **U**.

The optimal $\mathbf{w}^{\star}$ can be selected to maximize the transmission rate of the destination user, under the constraint of no interference at the users in **U**. Therefore, $\mathbf{w}^{\star}$ should maximize $|\mathbf{h}_{SU}^{\dagger}\mathbf{w}|^2$ under the conditions that $\mathbf{h}_{SU}^{\dagger}\mathbf{w} = 0$ and $\mathbf{w}^{\dagger}\mathbf{w}^{\star} \le P_{\max}$, where $P_{\max}$ is the total power. It can be seen that $\mathbf{w}$ is orthogonal to $\mathbf{h}_{SU}$, which means $\mathbf{w}$ belongs to the subspace of $\mathbf{h}_{SU}^{\perp}$, i.e., the null space of $\mathbf{h}_{SU}$. As a result, $\mathbf{w}^{\star}$ should be selected in the direction of the orthogonal projection of $\mathbf{h}_{SD}$ onto $\mathbf{h}_{SU}^{\perp}$.

Figure 4a shows the network scenario for simulation, where the squares represent the possible locations for small cells and the users are randomly distributed in the area. The number of coordinated small cells $N = 5, 6, 7, 8$, while the total power constraint is set to 2 W. The simulation results are obtained by a Monte Carlo simulation consisting of 5000 trials. Figure 4b shows the throughput of the targeted user versus the size of **U**. It can be seen that as the number of the protected active users increases, the throughput of the targeted user drops. Moreover, as the number of coordinated small cells increases, the throughput of the targeted user increases.

Related Standardization Activities

The Next Generation Mobile Networks (NGMN) Alliance is a mobile telecommunications association of mobile operators, vendors, manufacturers, and research institutes. The alliance's project results have been acknowledged by 3GPP and IEEE. In Feb. 2014, the NGMN Alliance announced the launch of a global initiative for 5G, with the objective of guiding the development of technologies and standards to satisfy the needs of the future. In this beginning phase, the NGMN Alliance has defined the requirements for 5G in terms of user experienced data rate, latency, mobility and so on.

Since HetNets play a key role in the evolution of 5G, 3GPP standards aim to guide the operation and management of small cells. For instance, inter-cell interference coordination (ICIC) was introduced in 3GPP Release 8 where BSs can communicate with each other via the X2 interface. Enhanced ICIC (eICIC) was introduced in 3GPP Release 10, which integrated almost blank subframes (ABS) to mitigate interference in the time domain. In 3GPP Release 11, ICIC has evolved to further enhanced ICIC (feICIC). In 3GPP Release 12, mechanisms for efficient operation of the small cell layer were introduced, such as interference mitigation through optimally powering On/Off small cells. Moreover, HetNet mobility, Wi-Fi cellular interworking, self-optimizing network, M2M application, etc, are included in Release 12.

For the cloud, the IEEE Intercloud Working Group is working on the project P2302-Standard for Intercloud Interoperability and Federation. This standard aims to create an economy among cloud providers that is transparent to users and supports evolving business models. It defines the topology and functions, and guides the interoperability of clouds.

Conclusion and Future Work

In this article we have introduced a cloud assisted HetNet architecture to realize 5G, which can simplify the complexity in terms of operation, maintenance, and deployment, caused by large-scale small cells. Meanwhile, it provides the centralized management to efficiently use network resources by coordinating the transmissions of small cells. Case studies on cloud assisted deployment and interference mitigation have been provided, which demonstrate the benefits of the cloud based architecture. It is anticipated that the cloud will accelerate the pace of 5G development, and further diversify wireless applications and services. In the future, the following research topics can be studied:

Security and Privacy: Since the cloud is involved in the operation, maintenance, and management of small cells, the security of the

cloud is of significance. For instance, malicious insiders can manipulate the small cells, which threatens the normal operation of the network. Besides, account or service traffic hijacking is another great security risk. Moreover, since small cells can be deployed by third parties or users, malicious users can easily deploy their own small cells to compromise the security of the users in the coverage, by performing man-in-the-middle attacks. In addition, considering that the network is split into a large number of small cells, the user's privacy, such as location privacy, can be easily revealed.

Energy Efficiency: With large scale deployment of small cells, energy consumption will be a critical issue. To make efficient use of energy while satisfying performance requirements, a cloud assisted small cell coordination mechanism is necessary, which can determine whether the small cell stays active or inactive and how the traffic can be balanced in real time [14].

Mobility Management: Mobility management is performed to maintain the users' connectivity when moving across different cells. Due to the densely deployed small cells, it is expected that users may handoff frequently. It is very critical to provide smooth handoff to the moving users for seamless connectivity [15]. To enable mobile users to have seamless connectivity anywhere and anytime, a cloud based intelligent handoff and location management scheme for HetNets is necessary.

References

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

[2] N. Lu *et al.*, "Vehicles Meet Infrastructure: Toward Capacity-Cost Tradeoffs for Vehicular Access Networks," *IEEE Trans. Intelligent Transportation Systems*, vol. 14, 2013, pp. 1266–77.

[3] X. Zhang, W. Cheng, and H. Zhang, "Heterogeneous Statistical QoS Provisioning over 5G Mobile Wireless Networks," *IEEE Network*, vol. 28, 2014, pp. 46–53.

[4] 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.

[5] N. Zhang *et al.*, "Cooperative Heterogeneous Framework for Spectrum Harvesting in Cognitive Cellular Network," *IEEE Commun. Mag.*, to apprear.

[6] A. T. Gamage *et al.*, "Device-to-Device Communication Underlaying Converged Heterogeneous Networks," *IEEE Wireless Commun.*, vol. 21, 2014, pp. 98–107.

[7] L. Lei *et al.*, "Operator Controlled Device-to-Device Communications in LTE-Advanced Networks," *IEEE Wireless Commun.*, vol. 19, no. 3, 2012, p. 96.

[8] J. Liu *et al.*, "Device-to-Device Communications Achieve Efficient Load Balancing in LTE-Advanced Networks," *IEEE Wireless Commun.*, vol. 21, no. 2, 2014, pp. 57–65.

[9] Z. Zhang *et al.*, "Coalitional Games with Overlapping Coalitions for Interference Management in Small Cell Networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, 2014, pp. 2659–69.

[10] F. Boccardi *et al.*, "Five Disruptive Technology Directions for 5G," *IEEE Commun. Mag.*, vol. 52, 2014, pp. 74–80.

[11] X. Zhou *et al.*, "Towards 5G: When Explosive Bursts Meet Soft Cloud," *IEEE Network*, vol. 28, 2014, pp. 12–17.

[12] Cisco Meraki, https://meraki.cisco.com/.

[13] T. Taleb, "Toward Carrier Cloud: Potential, Challenges, and Solutions," *IEEE Wireless Commun.*, vol. 21, no. 3, 2014, pp. 80–91.

[14] S. Zhang, Y. Wu, and Z. Niu, "Traffic-Aware Network Planning and Green Operation with BS Sleeping and Cell Zooming," *IEICE Trans. Commun.*, vol. 97, no. 11, 2014, pp. 2337–46.

[15] D. Lopez-Perez, I. Guvenc, and X. Chu, "Mobility Management Challenges in 3GPP Heterogeneous Networks," *IEEE Commun. Mag.*, vol. 50, no. 12, 2012, pp. 70–78.

Biographies

Ning Zhang [S'12] earned the Ph.D degree from the University of Waterloo in 2015. He received his B.Sc. degree from Beijing Jiaotong University and the M.Sc. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2007 and 2010, respectively. His current research interests include dynamic spectrum access, 5G, physical layer security, and vehicular networks.

Nan Cheng [S'13] is currently a Ph.D. candidate in the department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. He received his B.S. degree and M.S. degree 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.

Amila Tharaperiya Gamage [S'07] received his B.E. degree in electronics and telecommunications engineering from Multimedia University, Malaysia, in 2008, and his M.E. degree in telecommunications engineering from the Asian Institute of Technology, Thailand, in 2011. He is currently working toward his Ph.D. degree in the Department of Electrical and Computer Engineering, University of Waterloo, Canada. From 2008 to 2009 he was a solutions architect with Dialog Telekom PLC, Sri Lanka. His research interests include resource management for interworking heterogeneous networks, cooperative communication, and cloud computing. He is a corecipient of Best Paper Awards at IEEE ICC 2014.

Kuan Zhang [S'13] received the B.Sc. degree in electrical and computer engineering and the M.Sc. degree in computer science from Northeastern University, Shenyang, China, in 2009 and 2011, respectively, and is working toward the Ph.D. degree at the University of Waterloo, Waterloo, ON, Canada. He is currently with the Broadband Communications Research (BBCR) Group, Department of Electrical and Computer Engineering, University of Waterloo. His research interests include security and privacy for mobile social networks, cloud computing, and e-healthcare.

Jon W. Mark [M'62, SM'80, F'88, LF'03] received the 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, Waterloo, Ontario, 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 had been on sabbatical leave at the following places: IBM Thomas J. Watson Research Center, Yorktown Heights, NY, as a visiting research scientist (1976–77); AT&T Bell Laboratories, Murray Hill, NJ, as a resident consultant (1982–83); Laboratoire MASI, universit pierre et marie curie, Paris France, as an invited professor (1990–91); and the Department of Electrical Engineering, National University of Singapore, as a visiting professor (1994–95). 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 was an editor of *IEEE Transactions on Communications* (1983–1990), a member of the Inter-Society Steering Committee of the *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).

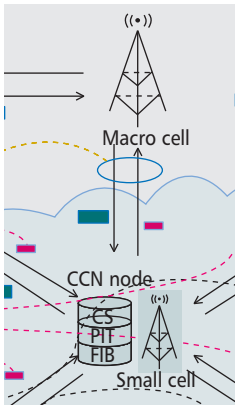
Xuemin (Sherman) Shen [M'97, SM'02, F'09] received the B.Sc. (1982) degree from Dalian Maritime University (China) and the M.Sc. (1987) and Ph.D. degrees (1990) from Rutgers University, New Jersey (USA), all in electrical engineering. He is a professor and university research chair, Department of Electrical and Computer Engineering, University of Waterloo, Canada. He was the 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 on wireless communications and networks, control and filtering. He served as the technical program committee chair/co-chair for IEEE Infocom'14, IEEE VTC'10 Fall, the symposia chair for IEEE ICC'10, the tutorial chair for IEEE VTC'11 Spring and IEEE ICC'08, the technical program committee chair for IEEE Globecom'07, the general co-chair for Chinacom'07 and QShine'06, the chair for IEEE Communications Society Technical Committee on Wireless Communications, and P2P Communications and Networking. He also serves/has served as the editor-in-chief for *IEEE Network*, *Peer-to-Peer Networking and Applications*, and *IET Communications*; he is a founding area editor for *IEEE Transactions on Wireless Communications*; an associate editor for *IEEE Transactions on Vehicular Technology*, *Computer Networks*, and *ACM/Wireless Networks*, among others; and a guest editor for *IEEE JSAC*, *IEEE Wireless Communications*, *IEEE Communications Magazine*, and *ACM Mobile Networks and Applications*, among 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 IEEE Fellow, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, and a Distinguished Lecturer for the IEEE Vehicular Technology Society and the Communications Society.

Mobility management is performed to maintain the users' connectivity when moving across different cells. Due to the densely deployed small cells, it is expected that users may handoff frequently. It is very critical to provide smooth handoff to the moving users for seamless connectivity.

Content Distribution over Content Centric Mobile Social Networks in 5G

Due to the rapid growth of network traffic and new demands from mobile users, the current mobile social networks face challenges in dealing with a huge amount of content requests, high operating costs, and mobility control. How to design mobile social networks efficiently for the upcoming 5G era has become an important issue. The authors propose and outline a framework of content centric mobile social networks for 5G.

Zhou Su and Qichao Xu



Abstract

Due to the rapid growth of network traffic and new demands from mobile users, the current mobile social networks face challenges in dealing with a huge amount of content requests, high operating costs, and mobility control. How to design mobile social networks efficiently for the upcoming 5G era has become an important issue. In this article we propose and outline a framework of content centric mobile social networks for 5G. First, we present a content centric based network architecture consisting of mobile users, communities, content centric nodes, small cells, and macro cells. Next, we discuss the detailed process to deliver mobile content based on the interests of content and mobile users. In addition, a novel caching scheme is presented to store replicas of mobile content, and the related experiment results are also given.

COMMUNICATIONS STANDARDS

Introduction

With the advance of network technologies and the innovation of mobile services, many efforts have been given by both academia and industry to design the fifth generation (5G) mobile networks [1, 2]. Some related projects such as 5GNOW [3] and 20BAH [4] are being carried out in academia, and there are also many standardization activities in industry.

Among these research activities, mobile social networks (MSNs) keep attracting much attention, where the MSNs have been developed rapidly and millions of mobile users can interact with each other to exchange content [5, 6]. Besides, since 5G can make it possible for mobile users to experience more emerging multimedia services including wearable mobile communications, augmented reality applications, etc, it can be predicted that MSNs will be one of the most important network paradigms in 5G.

However, compared with other conventional networks, there are some new challenges to develop MSNs in 5G. First, the amount of content requests in MSNs is not of the same order of magnitude as others. It always takes an extra delay, so it is hard to satisfy the requirements in 5G. Next, in MSNs the replicas of one content may be stored on different sites. As these replicas are managed according to their different locations in the current networks, the overhead to manage these replicas on different sites incurs huge operating costs. In addition, with the rapid development of the Internet of things (IoT), connected vehicles, etc., various types of content need to be delivered efficiently while mobile users are in motion [7]. New consideration to mobility should be given when designing MSNs in 5G.

In this article we propose content centric based mobile social networks to resolve the above problems. First, in the content centric networks (CCNs) [8], content is delivered based on the interest in it instead of sending the conventional requesting message. In MSNs the communities are organized by mobile users with common interests. If the content is delivered by interest, delivering one content in a community may satisfy multiple mobile users who have the same interest. Therefore, the total number of requesting messages in MSNs can be reduced. Second, in CCNs replicas are not controlled by their locations specified by IP addresses. Instead, the content is recognized by its content ID. This makes it possible to manage different replicas of the same content with the ID of the original content, and so the overhead to control these replicas can be reduced at the same time. Third, there are some CCN nodes in CCNs that cache some replicas of frequently used content within the coverage area of small cells. Therefore, when mobile users keep moving through the coverage area of different cells, they can obtain the wanted content from the replicas cached in CCN nodes in small cells, without contacting the far away server.

To realize the above content centric mobile social networks, both content distribution using virtualization and related standardization are needed. On one hand, the construction of virtual CCNs has been studied for content distribution [9]. However, the network architecture and process of mobile social content delivery have not been mentioned. On the other hand, the IETF has carried out standardization activities to discuss the protocol for CCNs [10]. Although the above shows some standardized approaches such as the structure of the identifier and the status of the path, the related analytical models and algorithms need to be studied further, especially for content centric mobile social networks.

Therefore, in this article we outline the delivery of mobile content based on content centric mechanism in MSNs. First we present a network structure of content centric MSNs, which consist of mobile users, communities, CCN nodes, small cells, and macro cells. Then the process of content delivery based on the interest in content among mobile users in communities is shown. Next, according to the availability of cached replicas in the content store of CCN nodes, we propose a novel caching scheme to replace replicas to efficiently use the cache capacity of the content store. In addition, experiment results are given to verify the proposal.

Zhou Su is with Shanghai University.

Qichao Xu is with Shanghai University.

 0163-6804/15/$25.00 © 2015 IEEE

Mobile Content Delivery in 5G

Mobile Communication in 5G

The rapid increase in mobile traffic has placed a huge burden on current mobile networks [11], where the volume of content, the population of users, and types of mobile devices all keep growing quickly. Due to limited resources including power, spectrum, etc., the related analysis shows that performance expectations cannot always be satisfied by various mobile network applications [12]. Besides, recent mobile networks are expected to support new emerging applications, such as tactile Internet, augmented reality, wearable devices, and smart mobile social communication, etc. To cope with these new demands, the existing networks could not be reliable enough to meet performance demands. To satisfy the emerging demands, the design of 5G mobile networks is needed.

According to the standards of 5G mobile networks, there are some detailed requirements relating to capacity, data rate, latency, connectivity, operating cost, and QoE provisioning, respectively [1]. For example, the aggregate data rate needs to be increased by 1000 times above the rate in 4G; the edge rate should be improved from 100 Mbps to 1Gbps; the round time latency is expected to be 1 ms while it is 15 ms in 4G [2].

Nowadays there are many efforts to design 5G mobile networks to satisfy the above requirements for the year 2020 and beyond. Among these efforts, as one of the most popular and practical network paradigms to provide content services, the design of MSNs is an important issue in 5G.

Mobile Social Networks

Different from the conventional network paradigms such as the client-server based structure, mobile users in MSNs need not always contact far away servers to request content. Instead, they can directly communicate with each other to share content by short range wireless interfaces with peer to peer opportunistic links.

One of the most important features in MSNs is that mobile users may have different interests in different content, and mobile users with common interests can form a community based on their social ties. In the community, the content can be shared and distributed among these mobile users. Small cells, which are low power cell towers installed by the operator, are responsible for providing communication within the community. Compared with small cells, a macro cell is made up of the traditional base stations. The macro cell can provide communication within a large coverage area about a few kilometers. With cooperation among the macro cell, small cells, communities, and mobile users, mobile content can be delivered in MSNs in 5G, with the main structure shown in Fig. 1. However, there are also new challenges to deliver mobile content in MSNs in 5G.

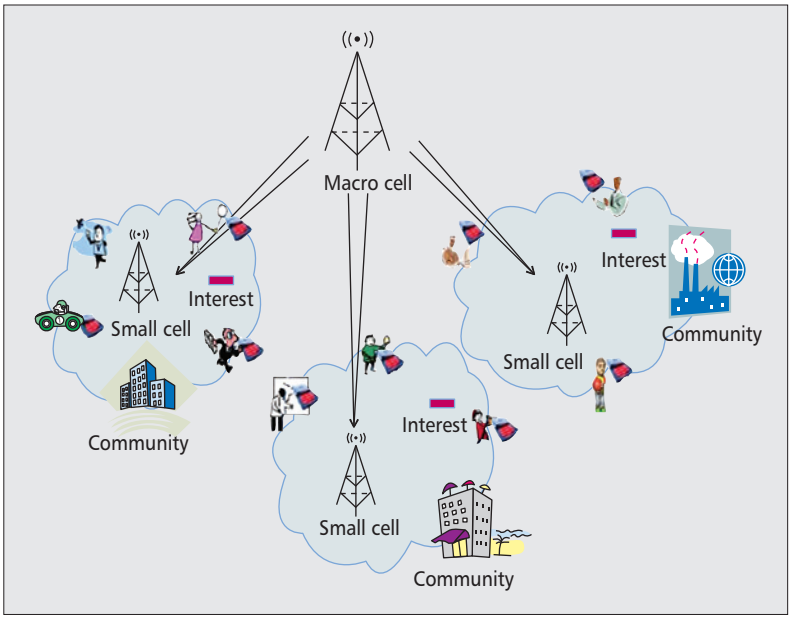


Figure 1. Mobile content delivery in 5G.

Challenges in Mobile Content Delivery in MSNs in 5G

How to Control a Huge Amount of Content Requests? In MSNs, mobile users cannot only produce their own content, but also request content from others. For example, a mobile user may update and publish his own content frequently, and can also subscribe to other people's content according to his interests and social ties. Compared with conventional networks, the amount of content requests in MSNs is not of the same order of magnitude. In particular, most requests are sent according to the interests of mobile users. Since there is an obvious relationship based on interests among content, mobile users, and communities, the requests should be controlled with a consideration of interests.

How to Control a Large Number of Replicas of Content? When a mobile user publishes his own content, this content may be subscribed to by other mobile users and then stored on different sites. Therefore, for the same content there will be a large number of its replcias, which incurs a high operating cost to manage these replicas. For example, when the producer of original content wants to update this content or publish some related information to this content, it is hard to recognize which replicas of his content are on which sites. Futhermore, due to mobility, it becomes even harder to control these replicas. Different from the conventional way of controlling replicas based on the location specified by an IP address, a new approach is required to manage replicas based on their own identity, e.g. its ID or name.

How to Control Load Balance? Compared with wired networks, both macro cells and small cells have their own limited capacities to deliver content. With the increase in the amount of content and the population of mobile users, it is expected that network traffic can be distributed. Instead of forwading all of the requests to the site that stores the original content, some content should be cached and placed within the small cell by different strategies to provide content with a possbile load balance.

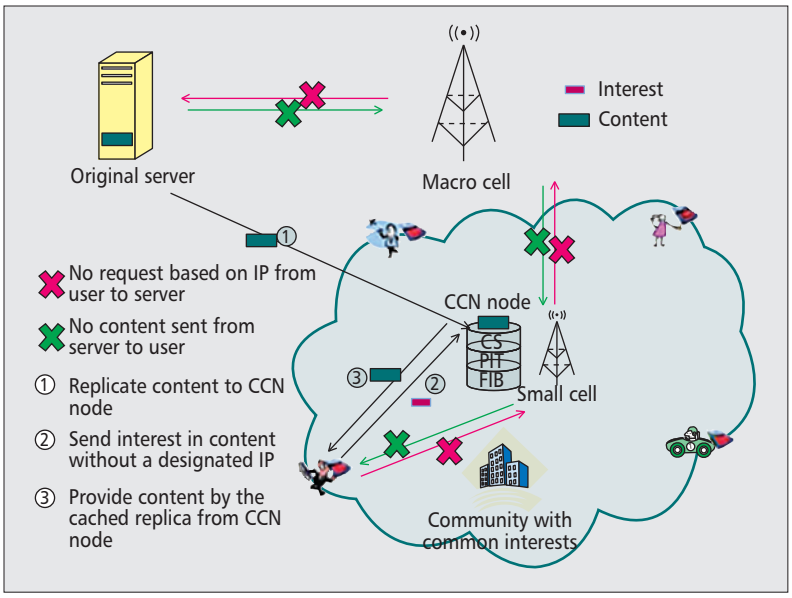


Figure 2. Network architecture of content centric networks.

Content Centric Mobile Social Networks

Content Centric Networks

To face the challenges when developing MSNs in 5G, a fundamental change to design a new network structure is needed, because the existing network architecture was designed based on the assumption that content delivery must rely on the IP address of the content. However, with the rapid spread of multimedia content, in addition to the conventional server with a designated IP address storing the content, different mobile users moving on different sites may also produce and consume content. Nowadays, content has less and less relation with where it is stored.

For example, a mobile user may publish the same content to two different communities located in different small cells. Although this content is the same, due to the different location, the existing networks manage it as two different contents. Therefore, the traditional approach to recognize and manage content based on the location specified by IP address will not be efficient now. In fact, current Internet users do not care where the content is from, they care more about what this content is. To face this new challenge, content centric networks (CCNs) [8–10, 13, 14] have been proposed as a new architecture for future networks to replace current networks.

As shown in Fig. 2, in CCNs there are many CCN nodes to store replicas of original content. The request for content is not sent to the IP address where this content is originally stored. Instead, an interest for the content is published over the CCNs; if a CCN node has the content where this interest refers, the replica of the content in this CCN node will be sent directly to the user who sent the interest. The interest in this content can be shown by a hierarchical name such as content ID. By sending the interest to the CCN node, the current IP address becomes unimportant, as the content requester need not know where the content was originally stored.

In the above CCN node, the content store (CS) is working with a pending interest table (PIT) and a forwarding information base (FIB). The content store caches the replicas of content. If a user sends an interest and the replica of this content is available in the content store, this replica will be provided to the user directly. Otherwise, the CCN node will check whether there is a pending interest for this content in the pending interest table. If there is still no matching interest, the CCN node will check the forwarding information base to wait for the wanted content to be fetched by a suggested delivering path. With cooperation among the content store, the pending interest table, and the forwarding information base, the content can be delivered according to the interest.

Content Centric Mobile Social Networks

We propose the content centric mechanism as an efficient solution for mobile content delivery in MSNs in 5G for the following reasons.

Community with Common Interest vs. Interest Based Content: In MSNs, the community consists of mobile users with common interests. As the content in CCNs is also delivered based on the interest, it is natural that multiple mobile users can be satisfied with one interest. As shown in Fig. 3a, when mobile users in the community subscribe to content with the same interest, the original server only needs to send the content once to the small cell. Then the small cell will distribute this content to these multiple mobile users who have interest in this content. That is to say: as multiple mobile users in the community may have the same interest in the content, content centric delivery of one content can satisfy the interest of multiple mobile users, and the total traffic of content delivery between the original server and the small cell can be reduced.

Mobile User ID vs. Interest Based Content ID: In MSNs, the content is produced by a mobile user who has a social ID. The content ID can be a hierarchical name containing the mobile user ID. If multiple replicas of this content are delivered over different communities by using the content ID, these replicas can be recognized as the replicas of the same content. When a mobile user wants to update content or publish some new information related to this content, the existing networks manage the replicas in different IP addresses as different content and update these replicas with a high overhead. But as shown in Fig. 3b, with content ID these replicas can be recognized as the replicas of the same content. When the original content is updated, different replicas of this content in the same small cell can be updated on time by the command from the small cell, with a reduction of overhead.

Mobility vs. Content Store: A mobile user may keep moving during the request of one content. If this mobile user can only get the interested content from the original server, when he leaves the coverage area of one small cell, the connection between the original server and the small cell may be interrupted. As shown in Fig. 3c, the content store in the CCNs nodes in small cells keeps some replicas of content. If the interested content is available in the content store, this replica can be provided to the mobile user.

Therefore, if the mobile user keeps moving through different small cells, he can get the interested content in the small cell without contacting to the original server, resulting in a reduction of delay.

Framework of Content Centric Mobile Content Delivery in 5G

Based on the above introduction, we show a framework of content centric mobile content delivery in 5G in this section.

Network Architecture

Mobile Users with Community: Mobile users take mobile devices to join MSNs, where the community consists of different mobile users with common interests. Mobile users may deliver their own content to others who have social ties with them. They can also request content stored outside of the communities.

Small Cells: Small cells are low power cell towers that are installed by the operator. Compared with macro cells, although the backhaul and access features are the same, the transmission power and coverage are lower. Generally, the small cell is directly responsible for communication among mobile users who want to share content with each other based on their social ties in the community.

CCN Nodes: CCN nodes are placed in the small cell for content centric delivery. Each CCN node has the following three parts with different data structures: content store, pending interest table, and forwarding information base.

- **Content Store (CS).** In each CCN node there is a caching space called the content store to keep replicas of some content. In fact, the content store has a limited capacity and cannot store all replicas of all content. The content store is similar to the buffer memory of conventional routers, but it is intended to cache content based on interest and has no direct relation with where this content is from.
- **Pending Interest Table (PIT).** The PIT is intended to control the pending interests that are currently waiting for responses. The track of forwarded interests to the destination can be kept, so the returned content can be sent to the content requester later. The PIT may keep a group of entries, and each entry will be erased as long as this entry is used to forward a matching content.
- **Forwarding Information Base (FIB).** The FIB is intended to forward interest to potential sources of matching content. Note that it can have multiple potential sources, so the query process can be parallel distributed. It is different from conventional routers, which only allow a single source.

Macro Cells: The macro cell consists of the traditional operator-installed base stations. Because it covers a wide area, it may provide access on the order of kilometers. Generally, the macro cell may provide the guaranteed minimum data rate for mobile content delivery and can serve thousands of mobile users with backhaul within its coverage area. The macro cell provides communications to the

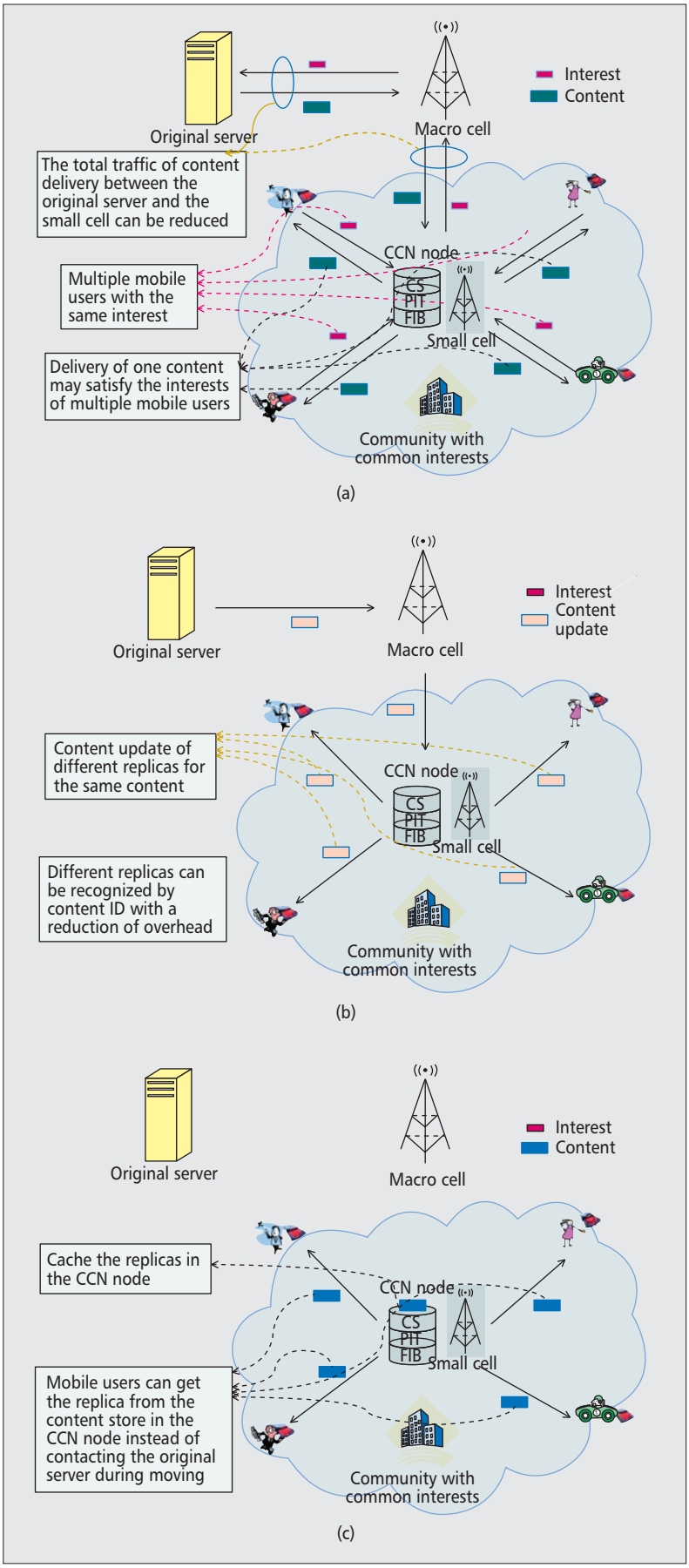


Figure 3. Content centric delivery as an efficient solution for mobile content in MSNs in 5G.

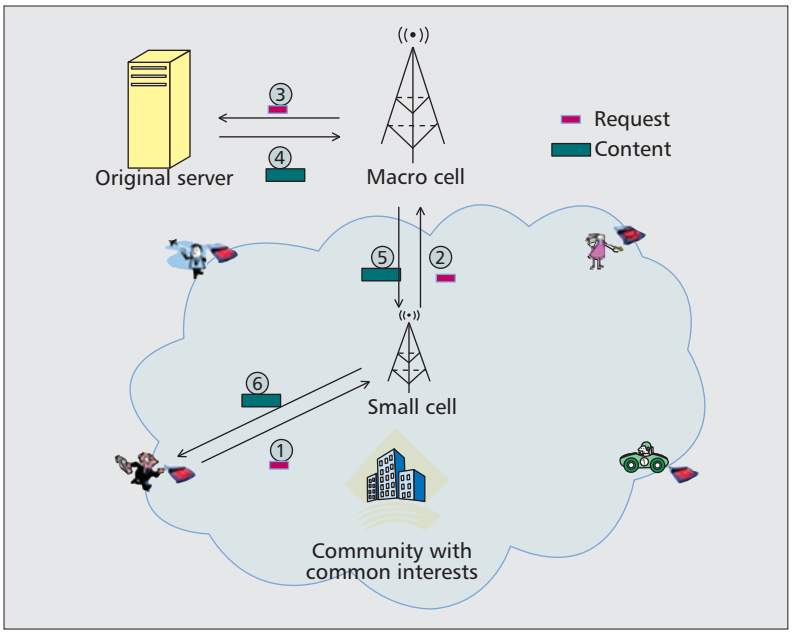


Figure 4. Conventional mobile content delivery.

small cell within its coverage area, and the small cell is responsible for the communities that are located in its area and have different mobile users.

Process of Mobile Interest and Content Delivery

Conventional Mobile Content Delivery: In existing MSNs, the process of content delivery is shown in Fig. 4. First, a mobile user within the community sends a request to the small cell where the community of this mobile user is in its coverage area. Then the small cell forwards this request to its connecting macro cell. Next, the macro cell contacts the original server of this content. After that, the original server sends the requested content to the macro cell. As the next step, the macro cell distributes the requested content through its small cell. At last, the small cell sends the requested content to this user. In this conventional mechanism, the mobile user needs to wait for a long time to let the small cell fetch content from the original server. Besides, as the content itself is forwarded from the original server to the requester, the traffic along the delivery path places a heavy burden on the networks.

Content Centric Mobile Content Delivery: Different from the existing method, content centric MSNs deliver content based on the interest. Mobile users request content by broadcasting the interest in content. Any mobile users who hear the interest and have the content can respond with the content. Therefore, during the process of content query, only the interest is forwarded.

According to the availability of replicas of the interested content in CCN node, there are two cases as follows.

Case 1: A Replica of the Content is Available in the Content Store of the CCN Node. As shown in Fig. 5a, when a mobile user sends an interest of content, the interest will be sent to the content store of the CCN node in the small cell, where the community of this mobile user is in its coverage. If the content store of the CCN node in the small cell keeps the replica of the content that the interest refers to, the replica of this content in the content store will be directly sent to the mobile user.

Case 2: A Replica of the Content is Unavailable in the Content Store of the CCN Node: In Fig. 5b, if there is no replica of content in the content store that matches the interest, the CCN node will check whether there is a matching entry of the interest in the PIT. If there is such an exact-match entry in the PIT, it means that this is a pending interest. If such an entry is not available in the PIT either, the CCN node will check if this interest has an exact-match entry in the FIB. Then the CCN node will wait for the content to be fetched from a delivery path suggested by the FIB.

Based on the above two cases, if a mobile user's interested content can be cached in the content store, both the response delay and the network traffic can be reduced. However, because of the limited capacity, not all of the replicas of content can be cached in the content store. If a newly arriving content needs to be cached in the content store, another content that is being cached in the content store must be removed in order to make room for this newly arriving content. How to cache the replicas of different content in the content store becomes very important for the performance of content centric MSNs. This so-called caching scheme becomes a new challenge.

Caching Scheme for a Content Store in a CCN Node in a Small Cell

Due to the importance of caching, we present a social centric caching (SCC) scheme for a content store in the CCN node within small cell as follows.

Suppose that the total number of contents to be distributed in MSNs is Q. For content q ($q = 1, \ldots, Q$), the probability that mobile users have interest in this content can be defined by P_q. For mobile user i ($i = 1, \ldots, I$) in community j ($j = 1, \ldots, J$), $P_{i,j}(w)$ denotes the probability that the degree of this mobile user is w. Here, the degree of mobile user i in MSNs means the number of friends he may have in community j. If a mobile user has more friends, this user may have more influence over content delivery. Therefore, the degree should be taken into consideration. Besides, if the rate of content delivery from mobile user i to one of his friends is λ_i, the accumulative rates from mobile user i to his friends in community j can be denoted by $\lambda_i \cdot P_{i,j}(w)$.

According to the social mobility in MSNs, mobile users may usually access networks through a few small cells. For example, students at a university may always access networks from the small cells in the dormitory and classroom. For mobile user i, if he has connected to a small cell k ($k = 1, \ldots, K$) by M times during the access of MSNs, as for the m-th time ($m = 1, \ldots, M$), the connection time is denoted by $S_{i,m,k}$. We call $S_{i,m,k}$ the social stay of mobile user i during the m-th connection with small cell k. Then the probability that mobile user i stays in the cover-

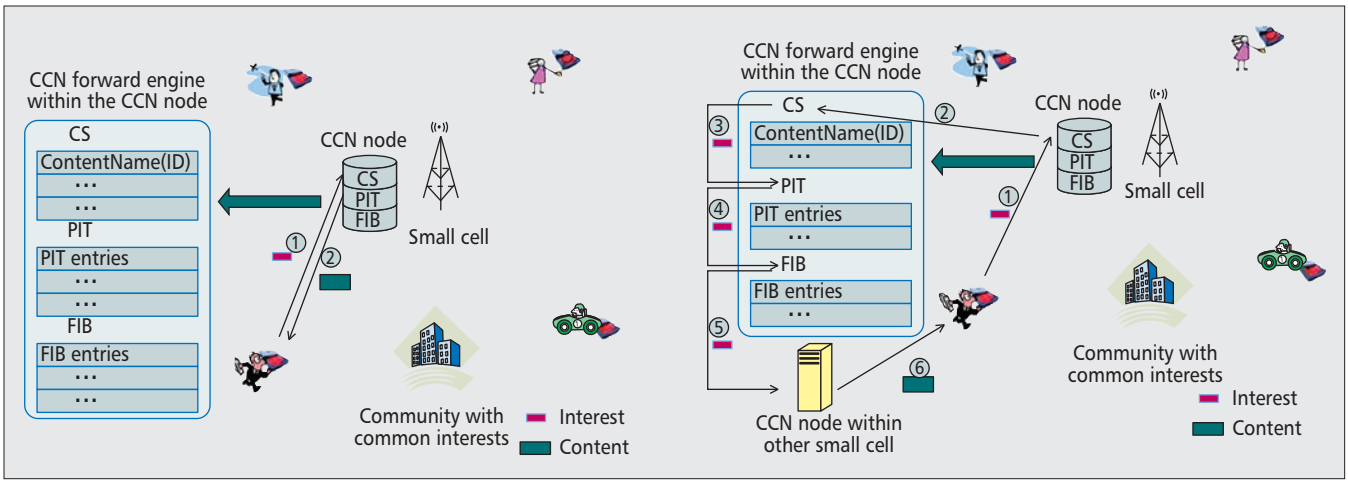


Figure 5. a) Replica of content is available in the content store of CCN node; and b) replica of content is unavailable in the content store of CCN node.

age area of small cell k becomes $T_{stay}^{i,k} = \Sigma_m s_{i,m,k}/T$, where T is a watching period.

If an interest in content q arises from mobile user i, when the replica of content q is being cached in the content store of a CCN node within the small cell k, the cost in delay to fetch the content q from this content store to mobile user i can be defined by $d_{i,k,q}$. Otherwise, when the replica of content q is unavailable in the content store of a CCN node within the small cell k, the CCN node needs to check both the PIT and the FIB to fetch the replica from another CCN node that is not located in small cell k, resulting in a cost denoted by $d'_{i,k,q}$.

Then we can obtain the relative cost if content q is not being cached in the content store of a CCN node within small cell k by $cost_{q,k} = \Sigma_i \Sigma_j (P_q \cdot \lambda_i \cdot P_{i,j}(w) \cdot T_{stay}^{i,k} \cdot (d'_{i,k,q} - d_{i,k,q}))$.

Performance Evaluation

We compare the hit ratio of the proposed SCC with the random method in Fig. 6. In the simulation, the distribution of interests follows the Zip-f distribution [15]. The rate of content delivery is determined by the Poison distribution. The degree of mobile users is decided by the power law distribution. Here, the random method means that the priorities of replicas to cache in the content store are determined at random. We carried out experiments to test the performance under different simulation times. In our scheme, the priorities of replicas to cache in the content store are determined by $cost_{q,k}$. As $cost_{q,k}$ takes into consideration the popularity of content, features of the social community, and delay, it can be looked upon as the cost to remove cached content from the content store. From Fig. 6, it can be obtained that the proposed scheme can result in a better hit ratio than the random method.

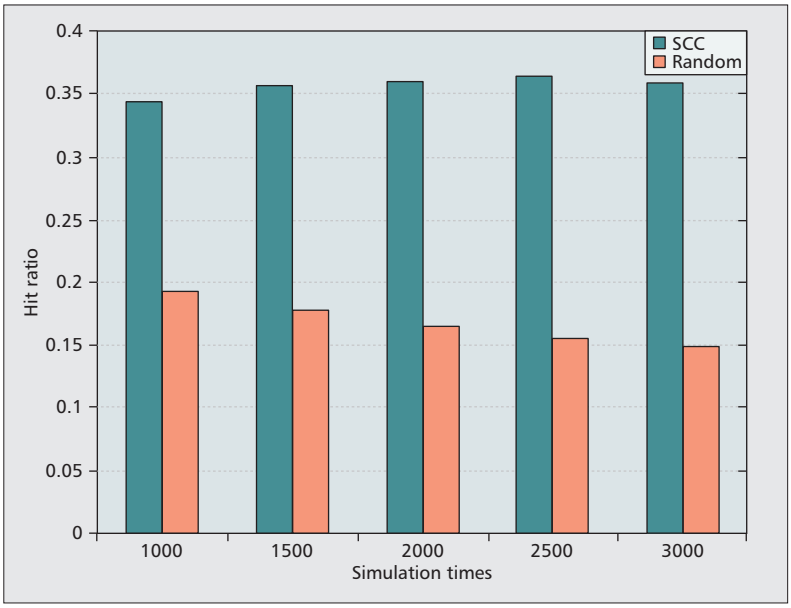


Figure 6. Hit ratio with different simulation times.

Conclusion and Future Work

In this article a novel framework to deliver content over content centric MSNs in 5G has been presented. We have shown the system structure consisting of mobile users, CCN nodes, small cells, and macro cells. The process of mobile content delivery based on the interest of content and mobile users in communities has been discussed. As caching in the content store of a CCN node plays an important role in the performance of content delivery, we also propose a caching scheme to determine which replicas should be stored in the content store. Experiment results were presented to show the efficiency of our scheme. There are some future research issues as follows:

- As there is a tradeoff between the cost of caching and caching performances, the related analysis of the cost model should be studied.
- The interests of mobile users may change during the delivery of content. How to develop a model to mimic the change of mobile users' interests should be discussed.
- Because multiple replicas of content are stored within the network, how to protect these replicas should be considered.

ACKNOWLEDGMENT

This work was supported in part by the fundamental key research project of the Shanghai Municipal Science and Technology Commission under grant 12JC1404201.

REFERENCES

[1] P. K. Agyapong *et al.*, "Design Considerations for a 5G Network Architecture," *IEEE Commun. Mag.*, vol. 52, no. 11, Nov, 2014, pp. 65–75.
[2] J.G.Andrews *et al.*, "What Will 5G Be?," *IEEE JSAC*, vol. 32, no. 6, 2014, pp. 1065–82.
[3] 5GNOW: 5th Generation Non-Orthogonal Waveforms for Asynchronous Signalling, www.5gnow.eu
[4] 2OBAH:2020 and Beyond AdHoc, www.arib.or.jp/ADWICS/2020bah-J.pdf
[5] Q. Xu *et al.*, "Epidemic Information Dissemination in Mobile Social Networks with Opportunistic Links," *IEEE Trans. Emerging Topics in Computing*, DOI: 10.1109/TETC.2015.2414792.
[6] Q. Xu *et al.*, "Analytical Model with a Novel Selfishness Division of Mobile Nodes to Participate Cooperation," *Peer-to-Peer Networking and Applications*, DOI:10.1007/s12083-015-0330-6
[7] K. Zheng *et al.*, "Radio Resource Allocation in LTE-Advanced Cellular Networks with M2M Communications," *IEEE Commun. Mag.*, vol. 50, no. 7, 2012, pp. 184–92.
[8] V. Jacobson *et al.*, "Networking Named Content," *Proc. CoNEXT 2009*, Rome, Italy, Dec. 2009.
[9] M.Ohtani *et al.*, "VCCN: Virtual Content-Centric Networking for Realizing Group-Based Communication," *Proc. IEEE ICC2013*, Budapest, Hungary, June 2013, pp. 3476–80.
[10] A. Detti *et al.*,"IP Protocol Suite Extensions to Support CONET Information Centric Networking," IETF, draft-detti-conet-ip-option, http://tools.ietf.org/html/draft-detti-conet.
[11] Z. Su *et al.*,"A Novel Design for Content Delivery over Software Defined Mobile Social Networks," *IEEE Network*, vol. 29, no. 3, 2015 (in press).
[12] L. Lei *et al.*, "Performance Analysis of Device-to-Device Communications with Dynamic Interference using Stochastic Petri Nets," *IEEE Trans. Wireless Commun.*, vol. 12, no. 12, Dec. 2013, pp. 6121–41.
[13] X. Jiang *et al.*, "Interest Set Mechanism to Improve the Transport of Named Data," *ACM SIGCOMM Comp. Commun. Rev.*, vol. 43, no. 4, pp. 515–16, 2013.
[14] S. Wang *et al.*, "On Performance of Cache Policy in Information-Centric Networking," *Proc. ICCCN12*, Munich, Germany, 2012.
[15] L. Breslao *et al.*, "Web Caching and Zip-like Distributions: Evidence and Implications," *Proc. IEEE INFOCOM 1999*, New York, Apr. 1999.

BIOGRAPHIES

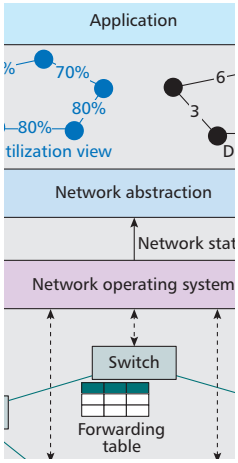
ZHOU SU (zhousu@ieee.org) received the B.E and M.E degrees from Xian Jiaotong University, Xi'an, China, in 1997 and 2000 respectively, and the Ph.D degree from Waseda University,Tokyo, Japan, in 2003. His research interests include multimedia communication, web performance, and network traffic. He received the best paper award at the International Conference CHINACOM2008, and the Funai Information Technology Award for Young Researchers in 2009.

QICHAO XU (xqc690926910@shu.edu.cn) is currently a master student at the school of Mechatronic Engineering and Automation of Shanghai University, Shanghai, P. R. China. His research interests are in the general area of wireless network architecture and mobile social network.

Software-Defined Networking Security: Pros and Cons

SDN is a new networking paradigm that decouples the forwarding and control planes —traditionally coupled with one another —while adopting a logically centralized architecture aiming to increase network agility and programability. While many efforts are currently being made to standardize this emerging paradigm, careful attentions need to be paid to security at this early design stage too, rather than waiting until the technology becomes mature, thereby potentially avoiding previous pitfalls made when designing the Internet in the 80' s.

Mehiar Dabbagh, Bechir Hamdaoui, Mohsen Guizani, and Ammar Rayes



Abstract

Software-defined networking (SDN) is a new networking paradigm that decouples the forwarding and control planes, traditionally coupled with one another, while adopting a logically centralized architecture aiming to increase network agility and programability. While many efforts are currently being made to standardize this emerging paradigm, careful attention needs to be paid to security at this early design stage too, rather than waiting until the technology becomes mature, thereby potentially avoiding previous pitfalls made when designing the Internet in the 1980s. This article focuses on the security aspects of SDN networks. We begin by discussing the new security advantages that SDN brings and by showing how some of the long-lasting issues in network security can be addressed by exploiting SDN capabilities. Then we describe the new security threats that SDN is faced with and discuss possible techniques that can be used to prevent and mitigate such threats.

COMMUNICATIONS STANDARDS

What is SDN?

Just when you start getting used to the IT buzzwords, such as mobile clouds [1], Internet of Things [2], and network virtualization [3], the IT people hit you with a new one. Software-defined networking (SDN) has recently received a great deal of attention from academia and industry, and has quickly become the new buzzword. This new networking paradigm (illustrated in Fig. 1) is based on the idea of decoupling the network's control plane from the forwarding plane, which results in turning the traditional network's complicated routing devices into simple switches whose job is merely to follow the policy that is depicted by an intelligent and programmable logically centralized controller. This is different from traditional networks shown in Fig. 2, where routing devices are performing both forwarding and control functions in a distributed fashion using static protocols.

In addition to those core differences that distinguish SDN networks from traditional networks, there are also implementation details that are specific to SDN networks. Unlike traditional networks where networking devices decide how an incoming packet should be handled based solely on its IP destination address, SDN follows a *flow-based* forwarding scheme where multiple header fields depict how the incoming packet should be handled. Furthermore, all network devices in SDN networks are recording traffic statistics, something that was performed by only a few devices (if any) in traditional networks.

SDN brings promising opportunities to network management in terms of simplicity, programability, and elasticity. These opportunities were quickly recognized by big industrial corporations, which started at an early stage funding research projects that aim at developing SDN. Today, the major networking vendors such as Cisco are releasing network infrastructure that supports SDN [4]. Furthermore, the paradigm turned out to be just what is needed for managing cloud and data centers. In fact, Google has revealed the use of SDN for managing the networking infrastructure of their data centers [5]. While many efforts are currently being made to improve and standardize SDN, we believe that further attention should be paid to security.

In this article we analyze SDN from a security perspective with the objective of shedding light on the new security capabilities and the new security threats that are brought by this new paradigm. The rest of this article is organized as follows. We briefly explain how the SDN paradigm works. We analyze the security advantages that SDN brings while highlighting what characteristics differentiate SDN from traditional networks, and how these characteristics can be exploited to improve network security. We discuss the new security issues that SDN faces, and describe which techniques could be used to prevent, mitigate, or recover from some of these issues. We describe the current state of the SDN standards with respect to security. Finally, we summarize and conclude the article.

How Does SDN Work?

Since the forwarding and the control planes are separated in the SDN paradigm, we explain next how each plane works. Throughout this section we refer to Fig. 1 for illustration.

Forwarding Plane: This plane is made up of simple switches that are interconnected to form the physical network. The role of these switches is to forward packets based on the control plane's routing policy. In order to achieve this role, each switch maintains a *forwarding table* whose entries are basically forwarding rules that are installed by the control plane. Each rule in the table is made up of three fields: pattern field, counter field, and action field. The pattern field defines a *flow* which is basically a set of packet header field values. Upon reception of a packet, the switch searches its forwarding table trying to find a rule whose pattern field matches the packet's header values. Once such a rule is found, the rule's counter field gets incremented and the rule's action field gets executed, which could be:

Mehiar Dabbagh and Bechir Hamdaoui are with Oregon State University.

Mohsen Guizani is with Qatar University.

Ammar Rayes is with Cisco Systems.

 0163-6804/15/$25.00

- Forward the packet as it is or after modifying some of its header fields to a specific port.
- Drop the packet.
- Report the packet back to the controller.

Control Plane: This plane represents the network's brain and is responsible for monitoring the network, making routing decisions, and programming the physical network how to behave. It is made up of three main layers: the network operating system layer, the network abstraction layer, and the application layer. As for the network operating system, it is connected to each switch in the network using a duplex link. This layer collects state information from the network switches such as: connectivity states, link delays, link utilization, etc., and passes this state information to the network abstraction layer. The role of the abstraction layer is to extract suitable network representations called views out of these collected data as it has a global view of the entire network. An example of a network representation would be a graph where vertices represent switches and edges represent links, where the edge's weight represents the link's delay or utilization. The application layer takes one or multiple representations as input, runs a certain algorithm that finds a routing policy that achieves certain objectives such as minimizing delay or avoiding link over-utilization, and returns a set of forwarding rules that are passed back to the network operating system, which distributes these rules to the network switches.

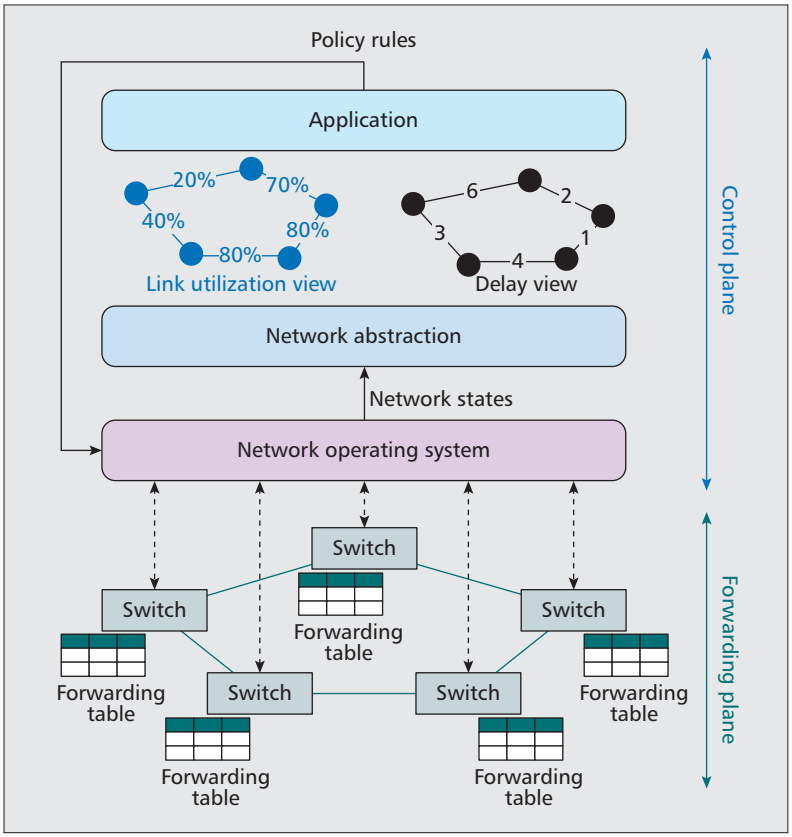


Figure 1. An SDN network.

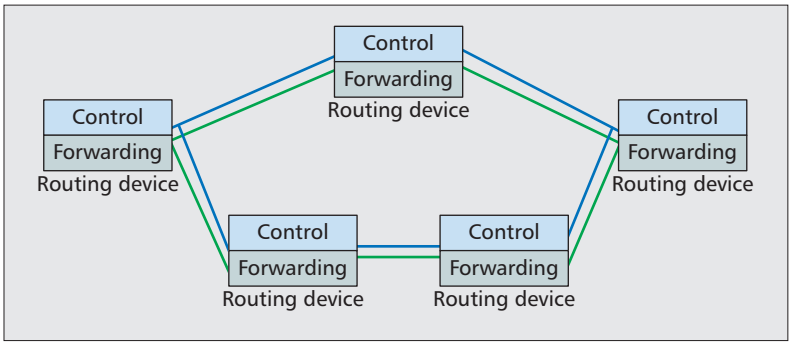


Figure 2. A traditional network.

SDN Security Advantages

The new structure of the SDN paradigm brings great benefits to the networking scheme. The way the functionalities are abstracted in the SDN paradigm allows writing high-level software applications to manage the network without worrying how to configure the underlying physical network. This is only one of numerous advantages that SDN brings. Our aim is not to discuss those numerous general benefits (e.g., see [6] for further details), but rather to focus on those related to security. We identify three core characteristics that differentiate SDN networks from traditional networks from a security perspective. Next we discuss each of those characteristics, explaining why traditional networks lack these characteristics, and how each one can be exploited to improve network security.

Global Network View

The fact that the controller in the SDN paradigm has a global network view is perhaps SDN's greatest security advantage over traditional networks. This global network view is attributed to centralization and to the fact that all the elements in the network are collecting and reporting traffic statistics. This is different from traditional distributed networks whose devices require exchanging lots of information and waiting for a convergence time in order to infer partially the state of the remaining parts of the network, and where only a few devices (if any) are logging traffic statistics. The security advantages that the global network view brings to the SDN paradigm are the following:

Network-Wide Intrusion Detection. The global network view allows the SDN controller to run a network-wide intrusion detection system (IDS) that analyzes the traffic statistics collected from all the network switches in order to detect malicious traffic. This is different from traditional networks, where the IDS is a device that is usually installed on a certain part of the network and thus provides limited detection capabilities as it has limited visibility. After collecting traffic statistics from the SDN switches, there are two ways for an IDS to operate:

- Misuse Detection: This approach is based on the idea of building profiles, called signatures, for known attacks. The behavior of the network is monitored constantly, and an intrusion is reported if the monitored behavior matches any of those signatures.

- Anomaly Detection: The network traffic is profiled when no malicious activity is going on by basically capturing the general characteristics of the packets originating from trusted hosts within the network while running trusted applications. An intrusion is reported if the network behavior deviates significantly from those profiles.

Each one of those two detection mechanisms has its pros and cons. Anomaly detection has the ability to identify new attacks and requires less in-depth knowledge about malicious behavior. However, this method produces more false positive alarms compared to misuse detection, as a deviation from the profiled traffic may not necessarily be a malicious behavior but rather a new normal behavior different from the profiled one. Researchers are still investigating improvements for the IDS framework in SDN with special focus on reducing processing overhead, increasing detection accuracy, and making the framework adaptive where new normal or malicious behavior can be learned automatically.

Detection of Malicious Switch Behavior. The global network view not only supports more efficient detection of intrusions originated from malicious traffic, but also helps in detecting network switches' malicious behavior. Consider for example the case where a networking device is dropping some or all incoming packets, creating a *black hole*. Identifying which routing device is responsible for that is not a trivial task in traditional networks. Probing techniques that are usually used in those networks, such as Traceroute [7], may not be effective for detection as the malicious switch may forward properly only those probes to hide its malicious behavior. Identifying such misbehaving behavior is easier in the SDN paradigm as switches report periodically to the control plane the number of received, forwarded, and dropped packets. By analyzing these reports, the misbehaving switch can be identified, or at least the list of suspected switches would be narrowed down. This is done even when the malicious switch claims in its report that all its incoming packets are forwarded properly to its neighbors, as the neighboring switches' reports will indicate correctly how many packets were actually received from that switch. However, it is still hard to differentiate dropped packets caused by link errors from those originating from malicious behavior. This is even more challenging when multiple malicious switches collude together to hide their malicious activity.

Network Forensics. Finally, the fact that the control plane logs the global view of the network over time facilitates performing forensic analysis. One can return back to the logged traffic in order to understand how an undetected attack was performed, something that is very helpful for developing effective defensive mechanisms against future instances of those attacks. Furthermore, this helps in identifying and isolating compromised hosts in addition to tracing back the person/organization behind those attacks.

Self-Healing Mechanisms

Another characteristic that distinguishes SDN from traditional networks is the fact that it is supplied with self-healing mechanisms. *Conditional rules* are an example of such mechanisms that were introduced to the SDN paradigm in [8]. These rules are installed on the switches by the control plane and are activated once a certain condition is met. The condition is usually related to the switch's collected statistics, such as when the number of packets belonging to a certain flow received during a certain period of time exceeds a predefined threshold. The activated rule specifies how the switch should respond when the specified condition is met. These reactions provide automated resiliency against attacks. The reaction specified by the conditional rule could be to drop the packets specified by the rule or to forward those packets through different paths to mitigate the load on certain parts of the network. This provides resiliency against denial of service (DoS) attacks targeting network hosts or network links. The reaction could be to modify the destination address of certain packets so they are delivered to a *honeypot*, which is an isolated and monitored host that is used as a trap to collect further information about malicious activities. This redirection is done in a stealthy way without having the originator of the traffic observe that. This is very useful for detecting *botnets*, which are a set of connected compromised hosts controlled by an attacker and used as a platform to launch distributed attacks against other parts of the network.

SDN characteristic	Attributed to	Security use
Global network view	• Centralization • Traffic statistics collection	• Network-wide intrusion detection • Detection of switch's malicious behavior • Network forensics
Self-healing mechanisms	• Conditional rules • Traffic statistics collection	• Reactive packet dropping • Reactive packet redirection
Increased control capabilities	• Flow-based forwarding scheme	• Access control

Table 1. SDN security pros over traditional networks.

Increased Control Capabilities

SDN networks have more control capabilities compared to traditional networks. This is attributed to the adoption of the flow-based scheme in SDN, where multiple header fields define how packets should be handled in the network rather than relying merely on the destination address, as in traditional networks.

Thus, the SDN controller can have better access control by specifying which types of packets should be carried within the network based on the payload type, the source address, or any other header field value. The rules installed by the controller can, for example, allow only TCP packets that originate from a certain host to be routed through the network. This helps in limiting malicious traffic from entering to or from originating from any switch in the SDN network. This is different from traditional networks, where networking devices forward packets blindly, leaving access decisions to the receiving end where a firewall usually sits and is required to inspect the payloads of all delivered packets.

SDN networks have more control capabilities compared to traditional networks. This is attributed to the adoption of the flow-based scheme in SDN, where multiple header fields define how packets should be handled in the network rather than relying merely on the destination address as in traditional networks.

SDN Security Threats and Countermeasures

Having explained the security advantages that SDN brings, which are summarized in Table 1, we now explain the new security threats and attacks that the SDN paradigm is exposed to. This section divides those attacks and threats into three categories based on which part of the SDN paradigm they target, i.e., the forwarding plane, the control plane, or the links connecting the two planes. Countermeasure techniques that could be used to prevent, mitigate, or recover from some of those attacks are also described while highlighting the challenges encountered when developing these defensive mechanisms.

Forwarding Plane Attacks and Countermeasures

We describe first some of the security threats the SDN's forwarding plane faces, as well as some possible countermeasures.

Switch DoS: Given that current switches have limited storage capacity and that the rules produced by the controller should cover all flows (all header fields possibilities), it becomes clear that it is next to impossible to store all these rules in current switches. A reactive caching mechanism is thus adopted instead in current SDN implementations where whenever a switch does not find a matching rule for the flow of one of its incoming packets, the packet is stored temporarily on a switch buffer, and a query is sent to the controller asking for the missing rule. Once that rule is received, the packet is processed based on that rule and the rule is cached in the switch's forwarding table so that the following packets of that flow are processed directly.

This reactive caching mechanism makes switches vulnerable to a DoS attack where a malicious user floods the switch with packets of large payloads that belong to different flows. The rules of some of these flows may not be cached in the forwarding table, which requires sending queries to the control plane. As a result, the switch ends up storing some of those large packets in its buffer waiting for the control responses. This buffer can be filled up quickly, especially if those packets have large payloads, causing legitimate packets belonging to new flows to be dropped as there is not enough space on the buffer to store those packets.

Multiple solutions were proposed to address this attack. One solution is proactive caching, where switches do not wait to receive new packets to request rules, but rather cache *a priori* as many rules as their table can fit. This technique turns out to be very efficient, especially when combined with using aggregate rules, where the installed rules cover ranges of header fields rather than single values, which compresses the number of needed rules. Finally, having a low delay on the links that connect switches with the control plane helps in processing the buffered packets quickly, making this attack harder to succeed. We will discuss later how to maintain low switch-controller communication delay.

Packet Encryption and Tunnel Bypassing: The adoption of the flow-based forwarding scheme allows SDN networks to customize how packets with different payloads are to be treated. This could be used for access control purposes, where the controller can specify which types of packets are allowed to flow inside and through the network. Packets with different payloads must be dropped or sent to the controller for further inspection.

Although the flow-based scheme brings new network management capabilities, it is not clear how this scheme should deal with encrypted packets where not all the packet's header fields are visible to the network switches. In fact, entire packets (headers and payloads) can be concealed from the network switches by creating tunnels that encapsulate an encryption of those packets within other packets. Once those packets are received at the other end of the tunnel, the inner packet is decrypted, decapsulated, and routed based on the policy of the new network that the other end of the tunnel belongs to. These encrypted packets and tunnel connections allow malicious users to skip their network border and bypass their network's access control policy. This can be addressed as in [9] by constructing models to identify the payload type of the encrypted packets based on analyzing traffic statistics such as message length, inter-packet arrival times, etc.

Control Plane Attacks and Countermeasures

We now discuss some potential attacks and their countermeasures, arising from the centralized nature of SDN's control plane.

DDoS Attack: The control plane is susceptible to distributed denial of service (DDoS) attacks where multiple compromised hosts distributed in the network may synchronously flood the network switches with packets. Since not all rules will already be available in the switches' tables, many queries will be generated and sent to the controller, which ends up utilizing the controller's processing power causing legitimate queries to be delayed or dropped.

One solution to such attacks is *replication*, where multiple physical controllers manage the network rather than a single one. However, the forwarding plane should continue to operate as if a single controller is programming the entire network. This is referred to as *logical centralization* and requires these multiple controllers to be connected to each other via secure links in order to maintain consistent rules all the time and to arrange how they should split up managing the network.

When replication is adopted, each switch will be connected to multiple controllers, and one of those controllers will be selected to be the master for that switch. Queries are directed to the master controller in the general case or to one of the remaining connected controllers when the master fails. The amount of queries generated by the switches, the processing capabilities of the controllers, and the switch-controller link delays are all taken into account when assigning which controller should be the master for each switch in the network. This is done with the objective of balancing the load on the controllers while also minimizing the switch-controller communication delays. These assignments need not be static but rather can be

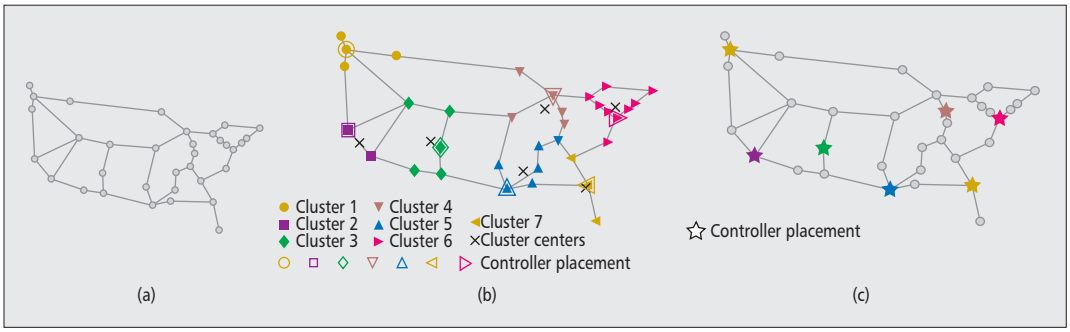


Figure 3. Illustration of the controller placement decisions for $k = 7$ where the objective is to minimize the average switch-controller distance: a) Internet2 OS3E topology; b) clustering-based controller placement; and c) global optimal controller placement.

dynamic, where a controller with high load can ask one of the remaining light loaded controllers to become the master for some of its assigned switches. This helps keep the load balanced among the controllers and provides resiliency against DDoS attacks.

When multiple controllers manage the network, deciding where to deploy these controllers is an important design question from a performance and resiliency perspective. The distance separating the switch from its master controller is a key factor to be taken into account when making those placements. Keeping this distance short guarantees low delay on the switch-controller communication link, which improves network responsiveness and makes it harder for switch DoS attacks to take place.

The controller placement problem is illustrated using the Internet2 OS3E topology, which is made up of a set of interconnected switches as shown in Fig. 3a, and where k replicated controllers need to be deployed on some network nodes. We are searching for a placement that minimizes the average switch-master delay, where each switch selects the controller reached with the shortest path to be its master. The problem is known to be NP-hard [10], and thus finding global optimal placements becomes computationally infeasible as the size of the network and/or the number of controllers grow. However, suboptimal solutions can still be obtained by clustering the switches into k clusters, with each cluster containing a set of switches that are geographically close to one another. Figure 3b shows the resulting clusters, when using k-Means [11], for the OS3E topology when $k = 7$, where switches belonging to different clusters are marked by different shapes/colors, and cluster centers are marked by 'x'. Next, for each cluster, the node that has the least average shortest-path distance to the remaining cluster members is selected for controller placement as illustrated in Fig. 3b. We also show in Fig. 3c the global optimal controllers placement which was obtained by a brute-force search. Observe that five out of seven locations are optimal locations, whereas the other two are adjacent to the optimal locations. In fact, the average switch-master distance based on the clustering-based placements is only 1.9 percent more than the optimal solution.

In addition to switch-controller delays, there are other factors that should be taken into account when making controller placements. For example, the ability to maintain switch-controller connectivity when some links are compromised is an important factor that needs to be accounted for when deciding on where to place controllers. This requires further investigation.

Compromised Controller Attacks: We discussed previously how to make the controller robust against DDoS failures. However, we did not discuss the scenario in which the attacker somehow gains access to the controller, putting all switches controlled by the compromised controller at the mercy of the attacker. The compromised controller can program those switches to drop all incoming traffic, or use them as a platform to launch serious attacks against other targets, such as directing all received packets to a victim so as to deplete its resources.

SDN must have some sort of resiliency against compromised controller attacks. Control replication could be used to resist such attacks. However, this solution will not be efficient if all the controllers are installed on similar platforms as they would all share the same vulnerabilities, allowing the attacker to break into all of them once it succeeds to break into one of them. Diversity among the platforms hosting those controllers is thus an essential condition for this resiliency technique to work. Furthermore, switches no longer query and receive responses from a master controller but rather communicate with all of the controllers that they are connected to. A majority voting technique can be used to decide which rule to follow in case different rules were received from the connected controllers. A switch connected to $2m + 1$ controller in that case would be resilient to up to m compromised controllers. Observe that the larger the number of controllers the switch is connected to, the higher it is resilient to be controlled by a compromised controller. Deciding how many controllers each switch should be connected to can be formulated as an optimization problem as in [12] while taking the processing capacity of those controllers as a constraint and where each switch in the network has a different security requirement (e.g., switches on the edge of the network have lower security requirements compared to those in the core of the network).

The distance separating the switch from its master controller is a key factor to be taken into account when making those placements. Keeping this distance short guarantees low delay on the switch-controller communication link, which improves network responsiveness and makes it harder for switch DoS attacks to take place.

Targeted level	Malicious behavior	Caused by	Possible countermeasures
Forwarding plane	Switch DoS	• Limited forwarding table storage capacity • Enormous number of flows • Limited switch's buffering capacity	• Proactive rule caching • Rule aggregation • Increasing switch' s buffering capacity • Decreasing switch-controller communication delay
	Packet encryption and tunnel bypassing	• Invisible header fields	• Packet type classification based on traffic analysis
Control plane	DDoS attack	• Centralization • Limited forwarding table storage capacity • Enormous number of flows	• Controller replication • Dynamic master controller assignment • Efficient controller placement
	Compromised controller attacks	• Centralization	• Controller replication with diversity • Efficient controller assignments
Forwarding-control link	Man-in-the-middle attacks	• Communication messages sent in clear • Lack of authentication	• Encryption • Use of digital signatures
	Replay attacks	• Communication messages sent in clear • Lack of timestamping	• Encryption • Timestamp inclusion in encrypted messages

Table 2. Summary of the new security issues that SDN networks are exposed to along with possible countermeasures.

Forwarding-Control Link Attacks and Countermeasures

Sending unencrypted communication messages on the link connecting the control and forward planes clearly makes the link susceptible to a *man-in-the-middle attack*. The attacker in that case can infer the control policy by eavesdropping the communication exchanged on the link. Even worse, the rules sent from the controller can be tampered with or new rules can be fabricated giving the attacker full control over the switch. Thus it is clear that the link layer must be secured against those attacks. Encryption must be used to prevent eavesdropping while the encrypted message needs to include some proof of the entity who originated those messages. A timestamp also needs to be included in the encrypted messages in order to prevent *replay attacks*, where a malicious user collects the encrypted rules sent by the controller, and sends them back at a later point of time, causing the switch to return back to use old policy rules.

For convenience, Table 2 presents a summary of the security issues that SDN networks are exposed to along with possible countermeasures we discussed in this section.

SDN Standards: OpenFlow's Robustness to Security Threats

We explained in previous sections the pros and cons of the SDN paradigm from a security perspective. In this section we discuss the current state of the SDN standards in terms of exploiting SDN's security capabilities and of adopting countermeasures against the newly exposed security threats. We focus on OpenFlow as it was the first released SDN standard and as it has gone through multiple revisions before becoming now widely deployed by networking vendors [13]. Our discussion is based on the standard's latest specifications released in December, 2014 [14].

Many of SDN's security advantages are exploited well in OpenFlow as the standard's specifications require switches to collect traffic statistics and to adopt the flow-based forwarding scheme. However, OpenFlow neither enforces switches to support conditional rules nor specifies how such rules should be handled. This limits SDN's self-healing ability as reactions to malicious traffic are not triggered directly by switches, but could be triggered by generating new rules by the controller later after analyzing the traffic statistics collected by the switches.

As for OpenFlow's countermeasures, the standard supports proactive rule caching though reactive rule caching is more widely used. However, the standard completely ignores how the incoming packets should be handled when some of their headers are hidden due to encryption. Although controller replication is supported in OpenFlow, no specifications were released regarding where to place the replicated controllers or how to make master controller selection [14]. Furthermore, dynamic master control assignment is not directly supported by OpenFlow controllers such as NOX [15]. Finally, one of the major security concerns with OpenFlow is the fact that it leaves it optional whether or not to encrypt the controller-switch communication channel [14], which makes this channel completely vulnerable to the threats discussed earlier. We anticipate that future releases of SDN standards will address these raised concerns.

Conclusion

We explained in this article how SDN works and analyzed its pros and cons from a security perspective. In summary, three key characteristics give SDN great security advantages over traditional networks: the global network view, the self-healing mechanisms, and the additional con-

trol capabilities. While SDN provides promising solutions to many security problems, it is also exposed to new threats and attacks targeting the forwarding plane, the control plane, or the links connecting the two planes. Several preventive and mitigation techniques were also described to address some of those security issues. The current state of the SDN standards was also analyzed with respect to security. We hope that SDN networks will further exploit the paradigm's security advantages while also addressing the new security concerns in the future.

Acknowledgment

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

References

[1] T. Taleb and A. Ksentini, "Follow Me Cloud: Interworking Federated Clouds and Distributed Mobile Networks," *IEEE Network*, vol. 27, no. 5, 2013, pp. 12–19.
[2] L. Atzori, A. Iera, and G. Morabito, "From "Smart Objects" to "Social Objects": The Next Evolutionary Step of the Internet of Things," *IEEE Commun. Mag.*, vol. 52, no. 1, 2014, pp. 97–105.
[3] T. Taleb, "Toward Carrier Cloud: Potential, Challenges, and Solutions," *IEEE Wireless Commun.*, vol. 21, no. 3, 2014, pp. 80–91.
[4] "Software-Defined Networking: Why We Like It and How We Are Building on It," Cisco Inc., White Paper, 2013.
[5] S. Jain *et al.*, "B4: Experience with a Globally-Deployed Software Defined WAN," *Proc. ACM SIGCOMM Conf.*, 2013, pp. 3–14.
[6] M.R. Sama *et al.*, "Software-Defined Control of the Virtualized Mobile Packet Core," *IEEE Commun. Mag.*, vol. 53, no. 2, 2015, pp. 107–15.
[7] M. Dabbagh *et al.*, "Fast Dynamic Internet Mapping," *Future Generation Computer Systems*, 2014.
[8] S. Shin *et al.*, "AVANT-GUARD: Scalable and Vigilant Switch Flow Management in Software-Defined Networks," *Proc. ACM SIGSAC Conf. Computer & Commun. Security*, 2013, pp. 413–24.
[9] Z. Fadlullah *et al.*, "DTRAB: Combating Against Attacks on Encrypted Protocols Through Traffic-Feature Analysis," *IEEE/ACM Trans. Net.*, vol. 18, no. 4, 2010, pp. 1234–47.
[10] B. Heller, R. Sherwood, and N. McKeown, "The Controller Placement Problem," *Proc. 1st ACM Wksp. Hot Topics in Software Defined Networks*, 2012, pp. 7–12.
[11] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," 2006, Morgan Kaufmann.
[12] H. Li *et al.*, "Byzantine-Resilient Secure Software-Defined Networks with Multiple Controllers in Cloud," *IEEE Trans. Cloud Computing*, no. 99, 2014, pp. 1–1.
[13] R. Jain and S. Paul, "Network Virtualization and Software Defined Networking for Cloud Computing: A Survey," *IEEE Commun. Mag.*, vol. 51, no. 11, 2013, pp. 24–31.
[14] "OpenFlow Switch Specifications Version 1.5.0," *Open Networking Foundation*, Dec. 2014.
[15] A. Dixit *et al.*, "Towards an Elastic Distributed SDN Controller," *Proc. ACM SIGCOMM Wksp. Hot Topics in Software Defined Networking*, 2013, pp. 7–12.

Biographies

Mehiar Dabbagh (dabbaghm@eecs.oregonstate.edu) received his B.S. degree in telecommunication engineering from the University of Aleppo, Syria, in 2010, and the M.S. degree in electrical and computer engineering from the American University of Beirut (AUB), Lebanon, in 2012. During his master's studies, he worked as a research assistant in the Intel-KACST Middle East Energy Efficiency Research Center (MER) at the American University of Beirut (AUB), where he developed techniques for software energy profiling and software energy-awareness. Currently he is a Ph.D. student in electrical engineering and computer science at Oregon State University (OSU), where his research focus is on how to make cloud centers more energy efficient. His research interests also include: cloud computing, energy-aware computing, networking, security, and data mining.

Bechir Hamdaoui [S'02, M'05, SM'12] is an associate professor at the School of EECS at Oregon State University. He received the Diploma of Graduate Engineer (1997) from the National School of Engineers at Tunis, Tunisia. He also received M.S. degrees in both ECE (2002) and CS (2004), and the 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 *Wireless Communications and Computing Journal* (2009–present). He also served as an associate editor for *IEEE Transactions on Vehicular Technology* (2009–2014) and for the *Journal of Computer Systems, Networks, and Communications* (2007–2009). He served as the program chair for SRC in ACM MobiCom 2011 and 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 Senior Member of the IEEE Computer Society, IEEE Communications Society, and IEEE Vehicular Technology Society.

Mohsen Guizani [S'85, M'89, SM'99, F'09] is currently a professor in the Computer Science & Engineering Department, Qatar University, Qatar. He 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, Syracuse, 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 EiC of the *Wireless Communications and Mobile Computing* journal 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 has received two best research awards. He is a member of IEEE Communication Society, and a Senior Member of ACM.

Ammar Rayes is a distinguished engineer at Cisco Systems and the Founding President of The International Society of Service Innovation Professionals (www.issip.org). He is currently chairing Cisco's Services Research Program. His research areas include: smart services, Internet of Everything (IoE), machine-to-machine, smart analytics, and IP strategy. He has authored/co-authored over a hundred papers and patents on advances in communications-related technologies, including a book on network modeling and simulation and another on ATM switching and network design. He is the editor-in-chief of *Advances of Internet of Things* journal, and has served as an associate editor of *ACM Transactions on Internet Technology* and the *Journal of Wireless Communications and Mobile Computing*. He received his BS and MS Degrees in EE from the University of Illinois at Urbana and his Doctor of Science degree in EE from Washington University in St. Louis, Missouri, where he received the Outstanding Graduate Student Award in Telecommunications.

One of the major security concerns with OpenFlow is the fact that it leaves it optional of whether or not to encrypt the controller-switch communication channel [14], which makes this channel completely vulnerable to the threats discussed earlier. We anticipate that future releases of SDN standards will address these raised concerns.

CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE
BIO-INSPIRED CYBER SECURITY FOR COMMUNICATIONS AND NETWORKING

BACKGROUND

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

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

The goal of this Feature Topic is twofold: (1) to survey the current academic and industry research in bio-inspired cyber security for communications and networking so that the ComSoc community can understand the current evolutionary state of cyber threats, defenses, and intelligence, and can plan for future transitions of the research into practical implementations; and (2) to survey current academic and industry system projects, prototypes, and deployed products and services (including threat intelligence services) that implement the next generation of bio-inspired methods. Please note that we recognize that in some cases, details may be limited or obscured for security reasons. Topics of interests include, but are not limited to:

- Bio-inspired anomaly and intrusion detection
- Adaptation algorithms for cyber security and networking
- Biometrics related to cyber security and networking
- Bio-inspired security and networking algorithms and technologies
- Biomimetics related to cyber security and networking
- Bio-inspired cyber threat intelligence methods and systems
- Moving target techniques
- Network artificial immune systems
- Adaptive and evolvable systems
- Neural networks, evolutionary algorithms, and genetic algorithms for cyber security and networking
- Prediction techniques for cyber security and networking
- Information hiding solutions (steganography, watermarking) and detection for network traffic
- Cooperative defense systems
- Bio-inspired algorithms for dependable networks

SUBMISSIONS

Articles should be tutorial in nature and written in a style comprehensible and accessible to readers outside the specialty of the article. Authors must follow the *IEEE Communications Magazine*'s guidelines for preparation of the manuscript. Complete guidelines for prospective authors can be found at http://www.comsoc.org/commag/paper-submission-guidelines.

It is important to note that *IEEE Communications Magazine* strongly limits mathematical content, and the number of figures and tables. Paper length should not exceed 4500 words. All articles to be considered for publication must be submitted through the IEEE Manuscript Central site (http://mc.manuscriptcentral.com/commag-ieee) by the deadline. Submit articles to the "June 2016 / Bio-inspired cyber security for communication and networking" category.

SCHEDULE FOR SUBMISSIONS

- Submission Deadline: November 1, 2015
- Notification Due Date: February 1, 2016
- Final Version Due Date: April 1, 2016
- Feature Topic Publication Date: June 2016

GUEST EDITORS

Wojciech Mazurczyk
Warsaw University of Technology
Poland
wmazurczyk@tele.pw.edu.pl

Sean Moore
Centripetal Networks
USA
smoorephd@gmail.com

Errin W. Fulp
Wake Forest University
USA
fulp@wfu.edu

Hiroshi Wada
Unitrends
Australia
hiroshi.wada@nicta.com.au

Kenji Leibnitz
National Institute of Information and Communications Technology
Japan
leibnitz@nict.go.jp