November 2015, Vol. 53, No. 11

IEEE Communications MAGAZINE

www.comsoc.org

- Communications Education and Training: Ethics and Professionalism
- Underwater Wireless Communications and Networks
- Software Defined Wireless Networks
- Green Communications and Computing

IEEE

IEEE COMMUNICATIONS SOCIETY

A Publication of the IEEE Communications Society

November 2015, Vol. 53, No. 11

IEEE Communications MAGAZINE

www.comsoc.org

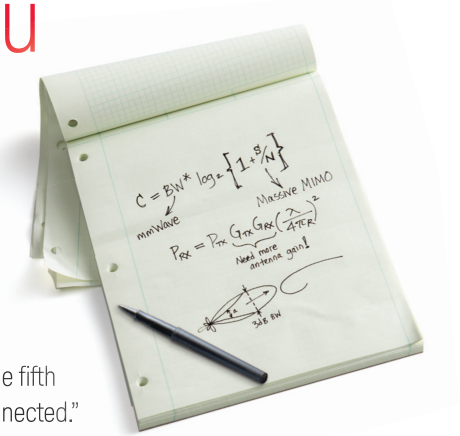
·Communications Education and Training: Ethics and Professionalism

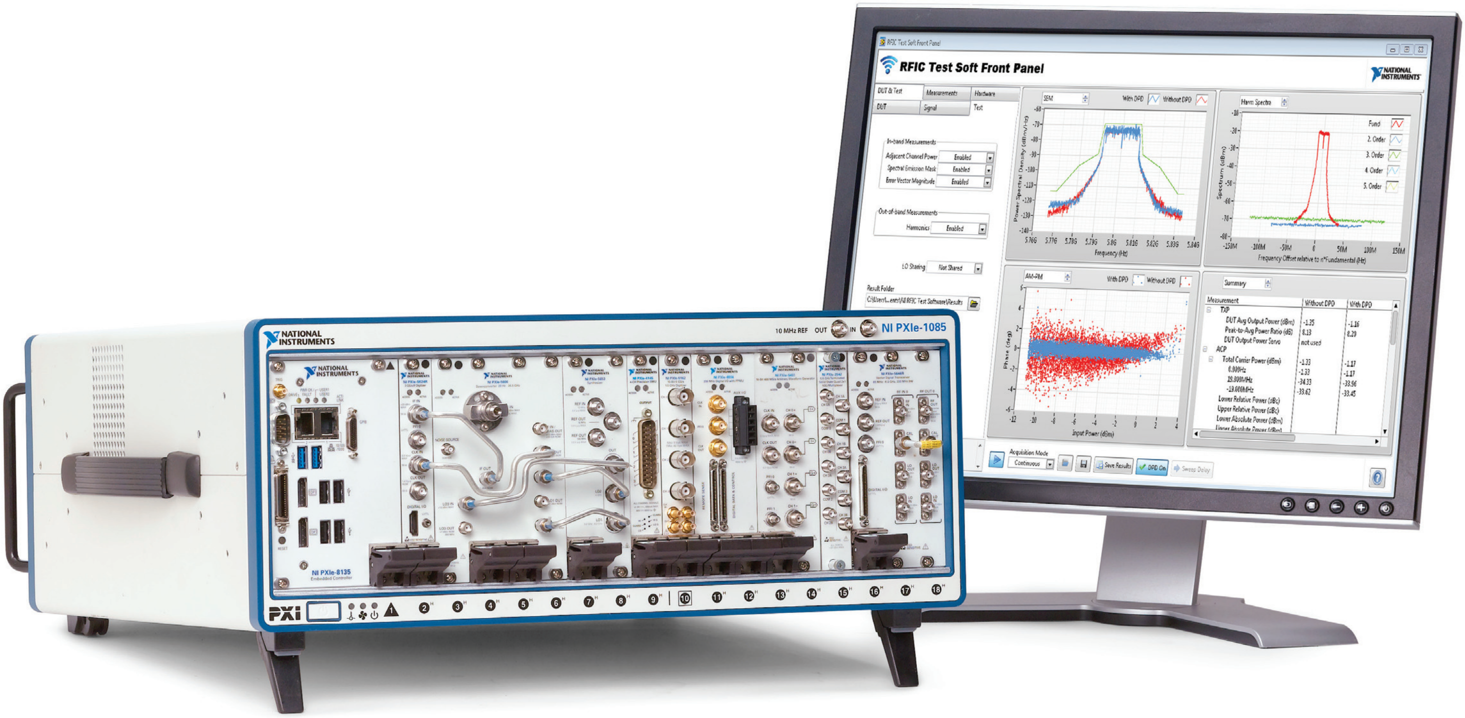
·Underwater Wireless Communications and Networks

·Software Defined Wireless Networks

·Green Communications and Computing

IEEE
IEEE COMMUNICATIONS SOCIETY
A Publication of the IEEE Communications Society


$C = BW^* \log_2 \{1 + S/N\}$
mmWave
Massive MIMO
$P_{RX} = P_{TX} \underbrace{G_{TX} G_{RX}}_{} \left(\frac{\lambda}{4\pi R}\right)^2$
Need more antenna gain!
α
3dB BW


RFIC Test Soft Front Panel
NATIONAL INSTRUMENTS
NI PXIe-1085
NI PXIe-8135
PXi

IEEE Communications MAGAZINE

NOVEMBER 2015, Vol. 53, No. 11

www.comsoc.org/commag

Communications Education and Training: Ethics and Professionalism

Guest Editors: David G. Michelson, Wen Tong, and Barry L. Shoop

Underwater Wireless Communications and Networks — Theory and Application: Part 1

Guest Editors: Xi Zhang, Jun-Hong Cui, Santanu Das, Mario Gerla, and Mandar Chitre

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.

Software Defined Wireless Networks (SDWN): Part 1

Guest Editors: Honglin Hu, Hsiao-Hwa Chen, Peter Mueller, Rose Qingyang Hu, and Yun Rui

Green Communications and Computing Networks

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

Accepted from Open Call

Currently Scheduled Topics

	Publication Date	Manuscript Due Date
LTE Evolution	June 2016	November 30, 2015
Wireless Technologies for Development (W4D)	July 2016	December 1, 2015
Recent Advances in Green Industrial Networking	October 2016	December 15, 2015
Communications, Caching, and Computing for Content-Centric Mobile Networks	August 2016	January 1, 2016

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

IEEE BlackSeaCom™
– 6-9 June 2016 // Varna, Bulgaria –

IEEE
IEEE COMMUNICATIONS SOCIETY

www.ieee-blackseacom.org

4th International Black Sea Conference on Communications and Networking

CALL FOR PAPERS

The IEEE BlackSeaCom series of conferences are held in the countries surrounding the Black Sea. The goal of the IEEE BlackSeaCom is to bring together visionaries in academia, research labs and industry from all over the world to the shores of the Black Sea. Here they will address many of the outstanding grand challenges that exist in the areas of communications and networking while having an opportunity to explore this exciting and dynamic region that has a rich history.

Following the first three editions of the conference in Batumi, Georgia in 2013, in Chisinau, Moldova in 2014 (originally scheduled for Odessa, Ukraine), in Constanta, Romania, in 2015, and the next edition of the conference will take place on 6-9 June 2016 in Varna, Bulgaria - a beautiful Black Sea resort.

We seek original, completed and unpublished technical papers not currently under review by any other journal, magazine or conference. Besides the regular technical papers, the conference welcomes the submission of poster papers which present work in progress, with preliminary results, tutorials and demos.

TECHNICAL SCOPE

- OpenFlow and Software Defined Networks
- Cloud Communications and Data-center networks
- Internet of Things and Sensor Networks
- Smart grids, M2M and Vehicular Networks
- Social Networking, eHealth, and Multimedia Applications
- Future Internet and testbeds
- Network Management
- Communication and Information theory
- Mobile and Wireless Communications and Networking
- Optical Networks and Systems, Radio over Fiber
- Cognitive Radio and Networking
- Security, Privacy and Trust
- Green Communications and Networks
- Signal Processing for Communications

IMPORTANT DATES

Regular Technical Papers:
Submission deadline: **Jan. 22, 2016**
Notification of acceptance: **March 11, 2016**
Camera-ready papers: **March 25, 2016**
(and authors registered)

Poster Papers:
Submission deadline: **March 25, 2016**
Notification of acceptance: **April 12, 2016**
Camera-ready poster-papers: **April 19, 2016**
(and authors registered)

General Co-Chairs
Alexander Gelman (adg@comsoc.org)
Erdal Panayirci (eepanay@khas.edu.tr)

TPC Co-Chairs
Albena Mihovska (albena@es.aau.dk)
Malathi Veeraraghavan (mv@ieee.org)

Conference Operations (incl. Registration)
Vladimir Poulkov (vkp@tu-sofia.bg)
Rozalina Dimova (rdim@abv.bg)

HOSTING UNIVERSITIES:

Technical University - Sofia

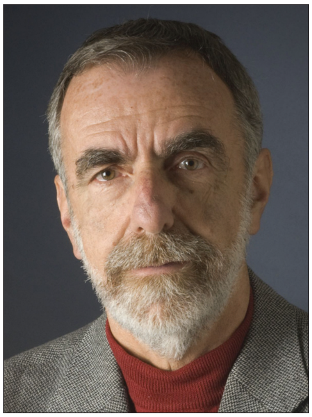
Technical University of Varna

PLEASE NOTE: To be published in the IEEE BlackSeaCom 2015 Conference Proceedings (both regular technical papers and poster papers) and to be eligible for publication in the IEEE Xplore (only for regular technical papers), an author of an accepted paper is required to register for the conference at the full (IEEE member or non-member) rate and the paper must be presented by an author of that paper at the conference unless the TPC Chair grants permission for a substitute presenter arranged in advance of the event and who is qualified both to present and answer questions. Non-refundable registration fees must be paid prior to uploading the final IEEE formatted, publication-ready version of the paper. For authors with multiple accepted papers, one full registration is valid for up to 3 papers. Accepted and presented regular technical papers and poster papers will be published in the IEEE BlackSeaCom 2015 Conference Proceedings, while only accepted and presented regular technical papers will submitted to IEEE Xplore.

www.ieee-blackseacom.org

STANDARDS ACTIVITIES

The President Pages from September to December 2015 will be devoted to a description of the activities and related achievements of the leadership of the IEEE Communication Society during my term as ComSoc President (2014-2015). The third page, in the November 2015 issue, is coauthored by Rob Fish and myself, and summarizes the activities in the areas of Standards Activities.


SERGIO BENEDETTO

ROBERT S. FISH

Robert S. Fish received his Ph.D. from Stanford University. Currently, Dr. Fish is President of NETovations LLC, a consulting company focused on the creation of communications and networking technology innovation. He is also on the faculty of the Computer Science Department at Princeton University. From 2007 to 2010 he was Chief Product Officer and Senior VP at Mformation, Inc., specializing in carrier software for mobile device management. From 1997 to 2007 he was Vice President and Managing Director of Panasonic U.S. R&D laboratories, working on the embedding of networking in consumer devices. Prior to this he was Executive Director, Multimedia Communications Research at Bellcore after starting his career at Bell Laboratories. Dr. Fish has more than 30 publications and 17 patents.

During his career, Dr. Fish and his organizations have initiated and managed standards development activities in IEEE, ISO/IEC JTC1, 3GPP, OMA, IETF, ATSC, CableLabs, OSGi, and many others.

Rob is the Vice President of Standards Activities of the IEEE Communications Society. He is also a member of the Board of Governors of the IEEE Standards Association, and was Chair of IEEE-SA's Global Committee, and a founding member of the IEEE-SA Corporate Advisory Group. He co-edited a series in *IEEE Communications Magazine* on IEEE standards in communications and networking. He is Co-founder and Steering Committee Chair of ComSoc's Consumer Communications and Networking Conference, and a member of the IEEE Conferences Committee. He is the former chair of ComSoc's GLOBECOM and ICC Management and Strategy Committee. For his leadership and contributions to the Multimedia Communications Technical Committee, Dr. Fish was the recipient of MMTC's Distinguished Service Award.

INTRODUCTION

The mission of ComSoc Standards Activities is to develop IEEE communications and networking standards as well as to create and maintain ComSoc standards-related activities and products. Standards Activities are governed by the Standards Activities Council that is headed by the VP-Standards Activities. During the 2014- 2015 period the VP-Standards Activities was Dr. Robert S. Fish, assisted by the members of the council and other senior volunteers and staff. Over the past two years the Council's priorities were to increase the quantity, quality, and timeliness of ComSoc standardization efforts as well as to increase participation by all ComSoc communities, both academic and industrial, in standards-related programs.

STANDARDS ACTIVITIES COUNCIL

The Council consists of two Boards: the Standards Development Board (COM/SDB), chaired by the Director of the ComSoc Standards Development, Dr. Mehmet Ulema; and the Standardization Programs Development Board, chaired by ComSoc Director of Standardization Programs Development, Dr. Alex Gelman. The Council is formed by ComSoc as the ComSoc structure for Standards Activities governance and as a partner to IEEE-SA with peer-relations to its governance organizations. According to ComSoc Bylaws the ComSoc VP-Standards Activities is ComSoc's representative to the IEEE-Standards Association (IEEE-SA) Board of Governors, the ComSoc Standards Development Board Chair is the official ComSoc liaison to the IEEE-SA Standards Board, and the ComSoc Standardization Programs Development Board Chair is the liaison to IEEE-SA Industry Connections Program.

ComSoc Standards Activities was created to serve as ComSoc's presence in the global Standards community. It provides industry a platform for the development of communications and networking standards as well as a way to build partnerships between academic and industrial researchers and standards developers. We believe that these partnerships help to improve the technical outcomes of consensus-based standardization processes. In ComSoc, the idea is to create an ecosystem where the standardization process involves not only materially interested parties, but also includes, at least in the early stages, but preferably throughout the entire standardization cycle, technology researchers with only a professed interest in standardization. The thought behind this inclusive ecosystem is that a balance of material and professed interests will result in technically superior standards.

Such an ecosystem needs to offer a value proposition to all stakeholders. Particularly challenging is to align the interests of industrial and academic researchers so as to induce them to collaborate. As a way to bring academics into Standards Activities, the ComSoc Standards Activities Council strives to organize Conference and Publication activities that feature standards-related contributions. Incentives that stimulate contributions to standards can also include an ability to cite standardization work and be rewarded for it, including promotion to IEEE fellow grade. But perhaps most important is to let researchers develop what we call scholarly standards. Those are standards in technologies that are still in early, often in precompetitive, stages of evolution but that utilize the very latest cutting edge research results.

Of course, not every standardization activity in ComSoc is like this. We also provide a platform for competitive-stage standards within the scope of communications and networking technology. In fact, some of our standards involve taking already-commercialized, but proprietary technologies, and enhancing them through an open, consensus-based process that results in a widely adopted global standard.

With this in mind, let's take a look at some of the accomplishments of ComSoc's Standards Activities Council over the last couple of years.

STANDARDS DEVELOPMENT

Standards Working Groups: Standards development in ComSoc is undertaken in working groups directly sponsored and supervised by COM/SDB or in the Standards Committees under COM/SDB that are organized around a theme and that can sponsor their own standards working groups.

In the latter class is the the Dynamic Spectrum Standards Committee (DySPAN-SC), which manages all of the IEEE 1900.x series of standards. These standards concern themselves with the definition and operation of cognitive radio systems. In the past two years revised versions of both 1900.1 and 1900.6b have been released to both update definitions and concepts as well as provide for interfaces to spectrum databases.

Similarly, the Powerline Communications Standards Committee (PLC-SC) manages the IEEE 1901.x series of powerline communications standards. Recently, its working groups have released IEEE 1901.2a-2015 IEEE Standard for Low-Frequency (less than 500 kHz) Narrowband Power Line Communications for Smart Grid Applications — Amendment 1, which enhances the 1901.2 narrowband powerline communications standard. PLC-SC adopted the Smart Energy Profile 2.0 application protocol, contributed by Zigbee Alliance, as the IEEE 2030.5 standard. This standard defines an application protocol to enable utility management of the end user energy environment, including things like demand response, load control, time of day pricing, management of distributed generation, electric vehicles, etc. Also under PLC-SC was the IEEE 1909.1 standard, which recommends practices for the testing and implementation of smart grid communication with regard to safety, electromagnetic compatibility, and environmental and mechanical requirements.

Also working under PLC-SC is the P1905 group, which concerns itself with Convergent Digital Home Networks. Last year they released a new standard, 1905.1a-2014 — IEEE Standard for a Convergent Digital Home Network for Heterogeneous Technologies Amendment 1: Support of New MAC/PHYs and Enhancements. This standard now includes IEEE 1901™ over power lines, IEEE 802.11™ for wireless, Ethernet over twisted pair cable, and MoCA 1.1 over coax. Additional network technologies are supported by an extensible mechanism.

A variety of projects managed directly by working groups under COM/SDB have also made significant progress in the last two years.

The P1904 working group on Service Interoperability for Ethernet Passive Optical Networks (SIEPON) has released a series of three standards that facilitate conformance testing with the IEEE 1904.1 SIEPON standard. In addition, two new projects were started: P1904.3 — Standard for Radio Over Ethernet Encapsulations and Mappings and P1904.2 — Management Channel for Customer-Premises Equipment Connected to Ethernet-based Subscriber Access Networks. COM/SDB also approved the change of the working group's name to "Access Networks" in view of the wider scope of its activities.

The P1906.1 working group on Nanoscale and Molecular Communications released IEEE P1906.1/Draft 2.0 Recommended Practice for Nanoscale and Molecular Communication Framework, which provides a precise, common definition of nanoscale communications and a general framework in order to enable diverse disciplines to have a common language and reference. A version of the 1906.1 ns-3 interface is freely available as a simulation tool to universities for students and researchers to more easily explore new directions in small-scale communication in a consistent and interoperable manner (https://github.com/ieee-p1906-1-reference-code).

The P1911 working group adopted from the HDBaseT Alliance specifications their new standards, 1911.1 and 1911.2. In addition, P1911.3 started work on the HDBaseT 5Play standard. These three standards enable distribution of ultra-high definition, uncompressed digital media. P1911.3 aims to add wireless compatibility, increased power throughput, Internet Protocol harmonization, and added security features.

The P2410 working group completed the IEEE 2410-2015 Biometric Open Protocol Standard. This standard provides a biometric-based access management solution that includes Identity Assertion, Role Gathering, Multilevel Access Control, Assurance, and Auditing.

INNOVATIONS IN THE STANDARDS PROCESS

One of the significant achievements of Standards Activities over the past two years has been to expand and refine our standards processes. Although our core processes remain compliant with the World Trade Organization and IEEE-SA processes, we have added two significant refinements to ComSoc's standards platform. First is the creation of the Rapid Reaction Standardization (RRS) methodology. This methodology brings together industry and academic experts in a short workshop format that requires them to provide position statements on the topic at hand before the workshop and then work together in real time to identify technology standardization gaps. Participants are recruited through public announcements and

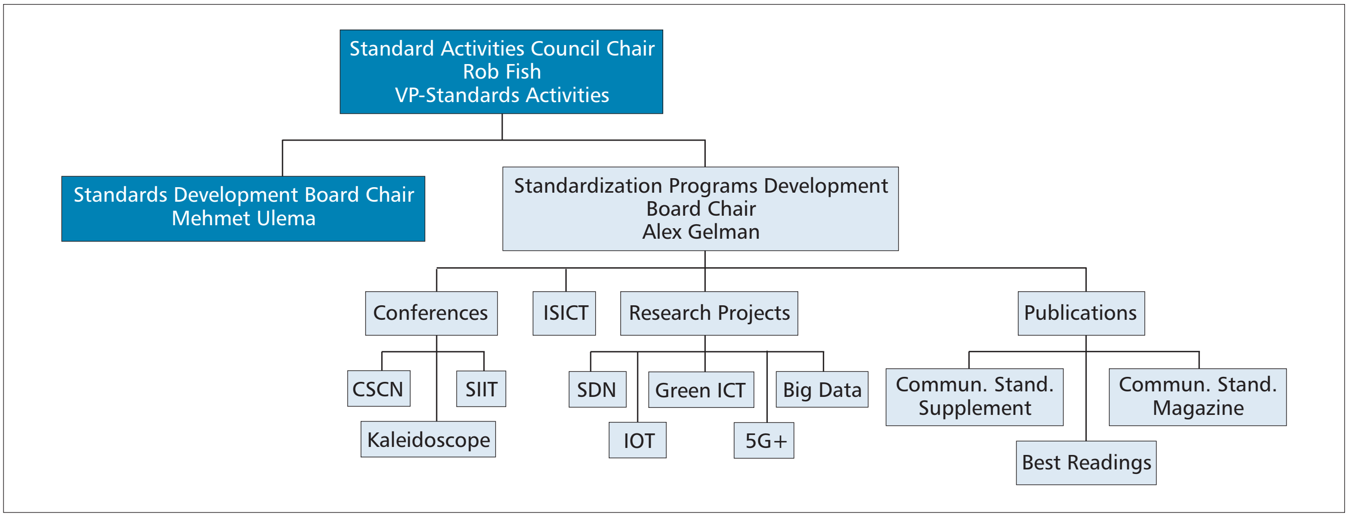


Figure 1. CSPDB activities chart.

ComSoc Technical Committees. Out of these sessions, working groups and study groups may be formed, or, and this is our second platform addition, we may form a new type of group that we call a "Research Group." A research group is formed when the technology gap identified requires more research before an official study group or working group can be formed. These research groups are open to all interested experts in the field. The output of a research group is often a white paper or a publication of some sort. After the research group has finished its work, it may morph into a regular study or working group.

Over the past two years, RRS activities have been held in the areas of Software Defined Networks/Network Function Virtualization (SDN/NFV), Internet of Things (IoT), and Fifth Generation Wireless Communications (5G). Currently in the planning stages are two additional RRS activities around the themes of Big Data and Green Communications.

NEW INITIATIVES

Out of these RRS activities, as well as through bottom-up requests, have come a number of new initiatives that may be of interest to ComSoc members. These include the working groups, study groups, and research groups detailed below.

New Working Groups: The IEEE P1912 Privacy and Security Architecture for Consumer Wireless Devices project was initiated by a group of security and privacy experts, and relates to IoT applications. The PAR for this standard describes a common communication architecture for diverse wireless communication devices such as, but not limited to, devices equipped with near field communication (NFC), home area network (HAN), wireless area network (WAN), wireless personal area network (WPAN), or radio frequency identification technology (RFID). The PAR also specifies approaches for end user security through device discovery/recognition, simplification of user authentication, tracking items/people under user control/responsibility, and alerting; while supporting privacy through user controlled sharing of information independent of the underlying wireless networking technology used by the devices.

The IEEE P1914.1 project targets a standard for Packet-based Fronthaul Transport Networks, and developed out of our 5G RRS activity. The Fronthaul Packet Transport project will enable the implementation of critical 5G technologies, such as massive Multiple-Input-Multiple-Output (massive MIMO), Coordinated Multi-Point (CoMP) transmission and reception, and scalable Centralized/Virtual Radio Access Network (C-RAN/V-RAN) functions.

Recently, the Com/SDB approved three new PARs that developed out of the SDN/NFV RRS activity. The PARs, P1914.1, P1915.1, and P1916.1, deal respectively with SDN/NFV security, performance, and reliability.

New Study Groups: Along with the above working groups, there have been a number of Study Groups (SG) formed as a result of the RRS activity. These include a SG on Service Virtualization. The objective of this SG is to investigate various scenarios in virtualization of Physical Network Functions (PNF) investments by service providers; to study the creation and orchestration of Service Composition Flows (the VNF Forwarding Graph); and to investigate service continuity of an offered network service (messaging) by building a context aware forwarding graph.

Another study group formed as a result of the RRS activity is one on IoT APIs. The objective of this SG is to identify primary standards development opportunities in the API/interfaces aspects of the mobile Healthcare, smart phone as a gateway, and related areas.

New Research Groups: In the SDN/NFV area, the following research projects have been formed:

- Research Group on Software Defined and Virtualized Wireless. Chair: Fabrizio Granelli, University of Trento.
- Research Group on SDN/NFV Structured Network Objects. Chair: Kenneth Kerpez, Assia.
- Research Group on SDN/NFV SLAs for Virtualized Environments. Chair: Mohammad Asad Rehman Chaudhry, University of Toronto.

An additional two project proposals are also being explored:

- SDN Testbed Project
- SDN Outage Database Project

In the Internet of Things area, the following research projects have been formed:

- Research Group on IoT Architectures. Chair: Jaeseung Song, Sejong University.
- Research Group on IOT Services. Chair: Yacine Ghamri-Doudane, University of La Rochelle.
- Research Group on IOT Communications and Networking Infrastructure. Chair: Stefano Giordano, University of Pisa.

In the area of 5G and beyond, an RRS activity produced significant interest in a Research Group on 5G Channel Modeling. Chair: Kevin Lu.

Two additional research projects are in formation:

- Research Group on Cloud-based Mobile Core Networking for 5G.
- Research Group on Radio Analytics for 5G.

Next New Initiatives: ComSoc participated in the IEEE Big Data Standardization Workshop that took place at the end of October 2015. Also planned for November of 2015 is a Rapid Reaction Standardization Activity workshop under the IEEE Green ICT initiative. Subject matter experts have been identified for these workshops in collaboration with the relevant ComSoc Technical Committees

OTHER STANDARDS PROGRAMS

As we can see from Figure 1, the ComSoc Standardization Programs Development Board, besides managing research groups, is also responsible for standards-related publications as well as meetings and conferences. Several notable achievements in this area have occurred in the last two years.

Communications Standards Supplement and Magazine: ComSoc Standards Activities initiated the creation of an IEEE magazine on communications standards in cooperation with ComSoc's VP of Publications, Katie Wilson and her team. The new magazine is currently being incubated as an *IEEE Communications Magazine* supplement. The inaugural issue was published in December 2014, and since then the supplement issues have been published quarterly. The supplement issues are organized around feature topics and also include news from IEEE and other global Standards Development Organizations, status reports from IEEE, e.g. ComSoc, standards development and research projects, and columns from notable standards leaders. Proposals for standards-related Feature Topic issues are solicited. They should be addressed to the supplement editor, Glenn Parsons.

IEEE-Conference on Standards in Communications and Networking (IEEE-CSCN): The inaugural IEEE Conference on Standards for Communication and Networking took place in Tokyo at the end of October 2015. This conference was created in cooperation with ComSoc's VP of Conferences, Hikmet Sari and his team. The support of the conference from industry and academia is clear from an impressive paper submission volume and in a successful patronage program. The intent is to bring the best papers from the conference into approriate dedicated Feature Topics of the *Communications Standards Supplement* in *IEEE Communications Magazine*. The conference General Chair is Tarik Taleb (Aalto University) who is also a member of CSPDB.

LOOKING TOWARD THE FUTURE OF COMSOC STANDARDS ACTIVITIES

Our plans for the future are in synch with the ComSoc 2020 strategy: grow the number of Standards Committees with the long-term goal of having Standards Committees harmonized with the Technical Committees in ComSoc, thus covering the entire technical scope of the Communications Society through related Standards Activities. If this is achieved, there will be a natural growth in standards-related programs in ComSoc, including in publications, conferences, and in education and training.

ComSoc is in the process of reorganizing its governance structure by combining Standards Activities with Industry Relations. The logic behind this is that standards activities are already of great interest to industry, and that this interest can be a part of bringing more industry participation in all ComSoc Industry-related activities. The next couple of years will be challenging, and a significant volunteer effort will be needed to help in this reorganization.

All in all, this has been an exciting couple of years for standards activities in ComSoc, and I look forward to these activities continuing to grow and prosper for the benefit of our members and all of society.

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)

REMEMBERING DAVID DAUT

DAVID DAUT

It is with great sadness that we report that David G. Daut, who served the *IEEE Transactions on Communications* (TCOM) for 25 years as Publications Editor, died on January 24, 2015.

David earned his Bachelor's degree from the New Jersey Institute of Technology, and his Master's and Ph.D. degrees from the Rensselaer Polytechnic Institute. From 1980 until his death, he was a faculty member in the Electrical and Computer Engineering Department at Rutgers University, where he served as Chair from 1986 to 1988 and again from 1997 to 2006. David was a prolific researcher who published numerous journal and conference papers on communications during his career. His work was recognized by the 1984 IEEE Communication Society Rice Prize Paper Award. David was also an enthusiastic IEEE volunteer. For example, he served as an elected member of the IEEE Board of Directors (Division III) during 1998 and 1999.

In addition to these accomplishments, David will be remembered most by our community as the Publications Editor of TCOM. David served in this capacity from 1988 to 2013. A dedicated and responsive Publications Editor was crucial for the timely processing of accepted papers and the success of a journal. However, the tasks to be performed by a Publications Editor are many and time-consuming. These duties included: requesting and verifying the final accepted papers and shaping each issue. To ensure a well-rounded print issue, he would delay or advance papers in the queue. In the age of IEEEx-plore and rapid posting the reason for shaping an issue may not be readily apparent. However, for most of David's tenure as Publications Editor, the print copies of TCOM were the preferred (if not only) option to access the papers. Therefore, organizing the material into well-rounded and appealing issues was an important task. Once an issue was complete, David sent the files of all papers of the issue to IEEE for printing. If the material submitted by the authors was incomplete, David would follow-up with the authors requesting the missing files. He also made sure that no paper was lost. Finally, he created and changed templates such as the overlength charge form, author submittal form, etc. David performed these demanding tasks for 25 years with politeness, professionalism, patience, great attention to detail, and commitment.

While the basic tasks to be performed by the Publications Editor remained the same over the 25 years of David's tenure, the means of communication and the way the papers were processed changed significantly. In the beginning, all communication was by mail and the processing was paper-based. Later, the communication moved to email and the processing was based on different file formats. Finally, TCOM moved to Manuscript Central and now the entire process from paper submission to exporting the final files to IEEE is done online. David was instrumental in implementing these changes and was always looking for ways to improve the service to the authors and the community, even if this meant more work for him. In 2012, paper publication was automated through the use of Manuscript Central. David was instrumental in this process. He worked closely with the TCOM Administrative Assistant and the Editor-in-Chief (EiC) to implement the online processing. The move to Manuscript Central was very smooth, which was largely thanks to David's assistance and vast knowledge of the involved processes.

Publications Editor is a position that involves a lot of work but receives very little credit. Nevertheless, David filled this position with great skill and commitment for 25 years. To illustrate the level of commitment that David demonstrated on a daily basis, we provide an excerpt from an email that the TCOM EiC, Robert Schober, received from David in the aftermath of Hurricane Sandy on November 7, 2012:

"I have successfully received the email below from you today. You are correct, we have been hit quite badly by the hurricane. Most of central, northern and coastal New Jersey, as well as the New York City area, have been ravaged by high winds, floods and relentless pounding by the surf. Rutgers University was closed from October 29 to November 5. I was at home without power for six days, no reliable phone service and no access to the Internet. Fortunately, in my neighbourhood power had been restored over the weekend, and we are slowly getting back to some sense of normal. I was at the University on Monday to lecture; however, today (Tuesday) I had to be home to oversee the repair of the roof on my house. The high winds had blown off many shingles. There is some urgency to get the repair finished since another storm, a Nor'easter, is due on Wednesday and Thursday which promises to bring high winds and a lot of rain. Predictions are that the coming storm will not be as severe as the hurricane. Everyone hopes the predictions are correct.

On Sunday, October 28, I did go into the office to catch up on TCOM activities along with other tasks in anticipation of the approaching hurricane. I was able to get all of the correspondence and rapid posting submissions up to date at that time. Upon checking my email yesterday, I noticed that there are only a handful of new acceptances and supporting material submissions that had come to me during the past week. I plan to go to the office tomorrow, Wednesday, and bring everything back up to date. Also, I will take some time to prepare an update for you on what papers I have already tentatively scheduled for the upcoming months of January, February, and March. Although the papers are already in, or being prepared by the Publisher for rapid posting, I have them assigned to monthly issues awaiting to be finalized and announced to the Publisher. In fact, this week I will be sending Joe the list of papers, that is the Table of Contents, for the January 2013 issue of the Transactions on Communications.

I will take a look at the email you sent dated October 30 and the attachments when I can access them from my work computer tomorrow. Afterward, I will be in a better position to comment on those items."

Even after 25 years of service, not even a hurricane that severely damaged his house could stop David from his duties as Publications Editor of TCOM. In 2013, David's exemplary service to our community was recognized with the IEEE Communication Society Joseph LoCicero Award for Exemplary Service to Publications for "outstanding service and dedication to the IEEE Communication Society as Publications Editor of the *IEEE Transactions on Communications*.

We will miss David's advice, company, and friendship dearly.

Desmond Taylor (EiC 1996–1999), Norman Beaulieu (EiC 2000–2003), Ender Ayanoglu (EiC 2004–2007), Michele Zorzi (EiC 2008–2011), Robert Schober (EiC since 2012), Len Cimini (ComSoc Director of Journals), Sarah Kate Wilson (ComSoc VP Publications), and Sergio Benedetto (ComSoc President)

GCN GLOBAL COMMUNICATIONS NEWSLETTER

November 2015
ISSN 2374-1082

CONFERENCE REPORT

IEEE ConTEL 2015

By Drazen Lucic, Hakom, Croatia

The International Conference on Telecommunications (ConTEL) 2015 was organized by the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, and the Institute of Microwave and Photonic Engineering, Graz University of Technology, Austria. ConTEL 2015 was technically co-sponsored by the IEEE. This year ConTEL celebrates the 13th issue of bi-annual events, and after 2011 as an alternate to Zagreb in Croatia, it was hosted again by Graz University of Technology 13–15 July. ConTEL 2015's program is a platform for scientists, industry, operators, and service providers to share ideas, discuss recent advances, to exchange their R&D experience, and present state-of-the-art techniques. The ConTEL 2015 technical program comprised keynote talks, a general track, special sessions on smart cities and multimedia services, as well as a workshop on future regulatory challenges.

The special sessions were on "Design and Evaluation of Interactive Multimedia Services and Applications" and "Smart Cities: Crowdsourcing and M2M Communication for a Connected Society". The special session included 12 papers, whereas the general track comprised eight sessions with 37 papers to be presented, both selected from more than 100 valid submissions. Each paper was evaluated by at least three independent reviewers with respect to their technical content, novelty, originality, and presentation. Overall, the review process involved 148 reviewers and 331 reviews. Keynote and invited talks were presented within plenary sessions. The first keynote, entitled "Overcoming the Optical Networks Capacity Crunch", was given by Peter J. Winzer, head of the Optical Transmission Systems and Networks Research Department at Bell Labs, Holmdel, NJ, United States. The second keynote, "Discovering the NFC Air Interface", was presented by Michael Gebhart, System Expert at NXP Semiconductors, Austria. The third keynote speaker was Magdy A. Bayoumi from the University of Louisiana at Lafayette, United States.

In the context of the Fifth Workshop on Regulatory Challenges in the Electronic Communications Market, Fatima Barros, Professor at the Lisbon School of Business and Economics and 2015 Chair of the Body of European Regulators for Electronic Communications, spoke about the "Regulatory Challenges in a New Digital Eco-system". Her speech concerned the anticipation of the new regulatory requirements and challenges at both the national and EU levels to create new opportunities for growth and innovation. In contrast to the papers submitted for the general track, or special sessions, the 12 contributors to the regulatory workshop are based on extended abstracts.

The research papers session on "Regulatory Challenges in the Electronic Communications Market" was a joint session with the main conference. After the session two round tables were performed, the first one with the topic "Digital Single Market", and the second one with the topic "Broadband Networks, Services and Users". Among others, the contributors, as well as the participants of the round tables, were the heads and of the national regulatory authorities of Albania, Austria, Croatia, Macedonia, Portugal, and Serbia. The conclusion of the debate at the round tables was that the development in technology, in the electronic communications and related markets, as well as the continuous change in consumers' needs, expectations and behavior, is affecting all sectors of the economy and society, resulting in a new digital economy. New opportunities for growth and innovation are emerging in Europe, and it is crucial that the new regulatory requirements and challenges are anticipated and addressed at both the national and European Union levels, in a coordinated way. Therefore, to make the most of the European digital economy, it is necessary to break down existing barriers, and use a holistic approach to promote the required cross-sector measures. The Digital Single Market Strategy for Europe, presented by the European Commission, goes in this direction and identifies the key role of telecommunications and the importance of appropriate regulation. Therefore, the main regulatory challenges are related to the fair treatment of players on new and cross-sector markets, demand take-up, and the promotion of competition as the main driver for investment in new infrastructures. In order to meet the long term connectivity needs of the European Union, exploit innovation, and capitalize on the new digital economy, European regulators will need to have an active role and regulate and deregulate as and when needed.

The next ConTEL conference is scheduled for 2017 in Zagreb, Croatia.

Participants of ConTEL 2015 in Graz.

ConTEL 2015 Session.

DISTINGUISHED LECTURER TOUR

IEEE ComSoc DLT of Pradeep Ray to China

By Pradeep Ray, Australia

I have been serving IEEE, particularly ComSoc, for more than 20 years as an author, a reviewer, an editor, a conference organizer, and a TC Chair, but this is my first time serving as a ComSoc Distinguished Lecturer (DL), for the term 2014-2015. I made my Distinguished Lecture Tour (DLT) to China during this past summer (22 June – 6 July). Before this I delivered a DL in 2014 in India, Bangladesh, UAE, and the Philippines. It was indeed a fantastic experience. I shared my research findings with our IEEE members, and also got to know more researchers to support multi-disciplinary collaboration across the engineering, public health, and business sectors, and enrich my understanding of local cultures.

I prepared a list of lectures on various aspects of eHealth and telemedicine, including talks on "Cooperative Service Management in Healthcare Sector: Emerging Trends and Future Challenges", and "Towards an Intelligent and Ubiquitous Healthcare Infrastructure". These talks were delivered at the IEEE Chapters in Beijing, Xian, Harbin, and Shanghai. The hosting chapters provided me with the necessary logistic arrangements, i.e. accommodations, local transportation, English speaking guides (essential in China), and meals, as required during the DLT program. My first lecture was in Beijing, which was hosted at the Beijing University of Post and Telecommunications (BUPT) on June 22, that happened to be a public holiday. In spite of its being a public holiday, there were approximately a dozen attendees during my talk. Most of them were Ph.D. students or researchers (some from overseas). It seemed that most of the attendees did not have much background in this emerging area of communications, although they were quite interested, based on their questions. After the talk at BUPT, the host took me to the Beijing West railway station, from where I took the overnight train to Xian. I was able to visit the Great Wall and the Summer Palace in Beijing on Sunday 21 June.

Pit 1 of Terracotta Warriors Museum near Xian.

Night view of European architecture at the Bund in Shanghai.

Pradeep Ray lecturing at Harbin Insititute of Technology at Harbin.

Xidian University in Xian was the host of my second talk on 24 June. They also provided me with pick-up arrangements from the Xian Railway Station on 23 June and took us to the Xian Airport on 25 June for the trip to Tianjin, where we spent the weekend before travelling to Harbin. This talk was longer, as requested by the hosts, and it seemed this group, led by Prof. Gang Yang, were already working on telemedicine. There were approximately 20 attendees, and we had a long discussion with the team of Prof. Gang Yang after my talk on 24 June. I was most impressed with the historical sites that have been preserved the 5000 years of Chinese history in Xian, which was the capital of 13 dynasties out of the 19 in Chinese history. The most prominent museum (the largest on-site archeological museum in the world) is the Terracotta warriors, which features full size colored statues of different divisions of armed forces built to protect the soul of the buried emperor. These statues were buried in 600 pits in that area. Only three of them have been excavated so far. The rest have not been touched because the statues lose color as soon as they come in contact with the outside atmosphere. Xian was considered the safest capital in China, as it is surrounded by mountains on three sides and the ferocious yellow river on the fourth side. This city has been the home of many Chinese intellectuals and politicians, including Chairman Mao Tse Tung.

My lecture in Harbin was conducted at Harbin Institute of Technology (HIT). The Chapter arranged to meet me at Harbin station on 29 June and take me to the Harbin airport on 5 July. It seemed that a number of schools of HIT were interested in multi-disciplinary research in eHealth, and hence I met with more than 50 people at different times during my week-long stay there. Prof. Weixiao Meng was my host from the HIT School of Information and Electronics Engineering. I had long discussions on future collaborations with Prof. Yongbin Yan, Prof. Doug Vogel, and Prof. Xitang Guo of HIT eHealth Research Insititute, and I delivered three more lectures on different aspects of eHealth services and research. Prof. Meng's students took my photo, shown above, during the IEEE DL. Of course, I did not miss the opportunity to visit the famous Siberian Tiger Reserve that hosts 1000 tigers of different types and age groups. We were driven in a caged vehicle while the tigers roamed free inside the reserve.

The Joint Institute of Shanghai Jiao Tong and Michigan University were my hosts for IEEE DL at Shanghai on 6 July. There were approximately 20 attendees. Prof. Xinwan Li was my host, and we discussed possible collaborations. I took this opportunity to take some photos of the building architectures around the famous Bund in Shanghai:

Thanks to the excellent organization by the IEEE Communications Society Asia Pacific Office, ComSoc DL organizers, and the hosts in Beijing, Xian, Harbin, and Shanghai, this IEEE ComSoc DLT

(Continued on Newsletter page 4)

REGIONAL REPORT

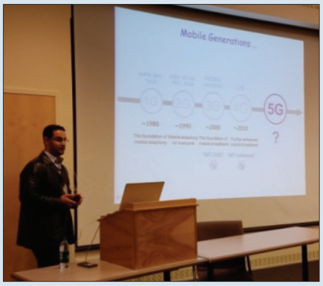
Recent Activities in the ComSoc Northeastern USA Region 1

By Dr. Ali Abedi, Chair of Maine Chapter and Region 6 Representative on the North America Region Board, and Dr. Ronald O. Brown, Vice-Chair of Maine Chapter, USA

The Northeastern USA Region 1 consists of the following chapters: Binghamton, Mohawk Valley, North Jersey, Boston, New Hampshire, Princeton Central Jersey, Buffalo, New Jersey Coast, Rochester, Connecticut, New Jersey Coast (Joint), Syracuse, Long Island, New York, Worcester County, and Maine.

The IEEE Maine Communications Society Chapter, in collaboration with several other chapters on the east coast (Rochester, Boston, North Jersey, Princeton Central Jersey, Jersey Coast, Columbus, Atlanta, and Palm Beach) organized a Distinguished Lecture Tour for Dr. Tarik Taleb on the East Coast.

Dr. Tarik Taleb is currently a professor at the School of Engineering, Aalto University, Finland. He has been working as senior researcher and 3GPP Standards Expert at NEC Europe Ltd, Heidelberg, Germany. His talk, titled "Towards 5G: Carrier-Grade Programmable Virtual Mobile Networks," was delivered on 19 March 2015 at the University of Maine. The talk highlighted the challenges that current and future mobile systems do and will face. He then showcased how programmable virtual mobile networks can be used as an efficient solution to revolutionize the congestion management concept and deal with the ever-growing volume of mobile traffic. His talk was well received by students, faculty, and industry professionals in attendance. Dr. Taleb's talk was unique in that it provided the audience with a multi-faceted perspective from the academic, industry, and standards points of view. The talk was followed by a tour of NASA's Inflatable Lunar Habitat (see photo below left), which is at the University of Maine's Wireless Sensing laboratory, and includes a 124-node wireless network of passive and active sensors.


Dr. Taleb.

On 24 April the Maine Chapter hosted Sebastian Ventrone, IBM Senior Technical Staff Member and IBM Life Time Master Inventor, to speak about the Art of Design through the Use of Innovation and the Harvesting of Patents. His talk was directed at students to complement their classroom training with this unique professional development opportunity. The student chapter was also involved in organizing this talk, and because of the high level of interest from students, they are planning to repeat this event in the Fall.

Another interesting event hosted by the Maine chapter in collaboration with the Boston and New Hampshire chapters was a DLT talk on 1 May by Dr. Hamid Jaafarkhani on the topic "Distributed Beamforming in Wireless Relay-Interference Networks", which was paired with a second talk on "Decoding of Binary Codes in the

(Continued on Newsletter page 4)

At NASA's Inflatable Lunar Habitat.

Profs. Jaafarkhani and Szczecinski at the Maine Chapter.

REGIONAL REPORT

Recent Activities in the ComSoc Western USA Region 6

By Zhensheng Zhang, Vice Chair of the San Diego Chapter and Region 6 Representative on the North America Region Board, USA

Within the North America region of the Communications Society (ComSoc), there are seven local regions (http://www.comsoc.org/about/chapters/NARegion), with the Western USA making up Region 6. In this article we will provide some details about the region and report on some of the recent activities within the region.

The Western USA Region 6 has approximately 3,715 ComSoc members and consists of the following 16 chapters: Buenaventura, Coastal Los Angeles, Foothill, Hawaii, Oakland East Bay, Orange County, Oregon, Phoenix, Sacramento, San Fernando Valley, San Francisco, San Diego, Santa Clara Valley, Seattle, Tucson, and Utah

This region is one of the most active regions within North America. Most of the chapters actively organize technical, advanced, and tutorial-type or hands-on seminars, host distinguished lecture tours, and plan social events to provide our ComSoc members with opportunities to network with their peers and add value to their ComSoc membership. Below are a few sample activities within the region, just to provide a glimpse within the region's activities. More detailed descriptions of the activities can be found in each chapter's website.

The Santa Clara Valley Chapter organized a technical talk on 13 May 2015 on the topic "NFV, SDN, and the Need for Optimized Server and Processor Architectures", given by Gopal Hegde, VP/GM, and Bharat Mota. Software defined networks (SDN) and network function virtualization (NFV), which advance the evolution of cloud computing, are technologies that will play a critical role in next-generation data centers and enterprise networks. Driven by continued demand for low-cost computing, the proliferation of NFV and SDN will place additional burdens on both computing and communications infrastructures of data centers. On one end, computing resources will be expected to shoulder additional demands of network workloads, whereas on the other end, there will be higher demands on scalability, lower latency, and increased I/O bandwidth demands, to mention just a few. Two experts from Cavium and Freescale provided a review of these issues, and presented new optimized server architectures to respond to the higher demands enabled by NFV and SDN. There were approximately 60 attendees. The Santa Clara Valley Chapter also hosted an annual society mixer event on 13 August 2015, providing networking opportunities for our members.

A Distinguished Lecture Tour was hosted by the Seattle, Oregon, and Vancouver chapters between 19 May and 23 May 2015. The Distinguished Lecturer was Dr. Rath Vannithamby, covering the topic "5G Evolution and Candidate Technologies". As 4G standards have been completed and networks are beginning to be deployed, the attention of the mobile research community is shifting toward what will be the next set of innovations

(Continued on Newsletter page 4)

Region 6 Report/*Continued from page 3*

in wireless communication technologies. Given a historical 10-year cycle for every generation of cellular advancement, it is expected that networks with 5G technologies will be deployed around the year 2020. Technologies for future cellular wireless networks and devices are expected to meet the needs of an increasingly diverse set of devices and services in 5G. This presentation discussed the usages and technologies that will comprise the next set of cellular advancements in 5G, in particular: the applications and usages for future 5G communications; a set of key metrics for these usages and their corresponding target requirements; and the potential network architectures and enabling technologies to meet 5G requirements. It is expected that some of the new technologies comprising 5G will be evolutionary, covering gaps and enhancements from 4G systems, while some other technologies will be disruptive. These technologies will encompass the end-to-end wireless system, from wireless network infrastructure to spectrum availability to device innovations. The presentation also provided an overview of 5G activities around the world to provide a better understanding of the vision and research direction of various teams as they tackle the challenging problems of capacity, massive numbers of IoT devices, ultra-low latency, ultra-low power efficiency, etc., that wireless networks are expected to face beyond 2020.

The San Diego Chapter, which won the ComSoc Best Chapter Award in 2013, hosted many technical talks and is actively co-sponsoring the ComSoc flagship conference IEEE GLOBECOM 2015 to be held 6–10 December 2015 in San Diego.

On 18 June 2015 the San Diego chapter hosted a technical talk on "Noninvasive Detection of Emotional Contagion in Online Social Networks", given by Dr. Lorenzo Coviello. We can summarize the talk as follows: Does semantic expression spread online from person to person? And if so, what kinds of expression are most likely to spread? To address these questions, the speaker proposed a non-experimental, noninvasive method to detect and quantify contagion of semantic expression in massive online social networks. Using only observational data, the method avoids performing emotional experiments on users of online social networks, a research practice that recently became the object of criticism and concern. The model combines geographic aggregation and instrumental variables regression to measure the effect of an exogenous variable on an individual's expression and the influence of this change on the expression of others to whom that individual is socially connected. The method is applied to the emotional content of posts generated by a large sample of Facebook users over a period of three years. Those results suggest that each post expressing a positive or negative emotion can cause friends to generate one to two additional posts expressing the same emotion, and it also inhibits their use of the opposite emotion. The method can be applied to contexts different than emotional expression and to different forms of content generated by the users of online platforms. The method makes it possible to determine the usage of words in the same semantic category spread, and to estimate a signed relationship between different semantic categories, showing that an increase in the usage of one category alters the usage of another category in one's social contacts. Finally, it also allows one to estimate the total cumulative effect that a person has on all of their social contacts. Approximately 50 people attended the talk.

As mentioned earlier, the Western USA Region 6 is one of the most active local regions within North America. ComSoc members are encouraged to check each chapter's website for upcoming events and to get involved in their chapter's activities.

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

MILAN JANKOVIC (MILAN.JANKOVIC@RATEL.RS)
EWELL TAN, SINGAPORE (EWELL.TAN@IEEE.ORG)

A publication of the
IEEE Communications Society

www.comsoc.org/gcn
ISSN 2374-1082

Region 1 Report/*Continued from page 4*

Presence of Channel Mismatch: Correction of the LLRs", delivered by Dr. Leszek Szczecinski. Hamid Jafarkhani is a Chancellor's Professor in the Department of Electrical Engineering & Computer Science at the University of California, Irvine, where he is also the Director of the Center for Pervasive Communications & Computing and the Conexant-Broadcom Endowed Chair. Hamid Jafarkhani is one of the top 10 most-cited researchers in the field of "computer science" in the period 1997–2007. Leszek Szczecinski is an associate professor at INRS-EMT, University of Quebec, Canada, and adjunct professor at in the Electrical and Computer Engineering Department at McGill University. These two talks attracted both a technical and general audience to the event, making the long trips to the talk venue worthwhile.

Because of the large geographic dispersion of members in the Maine section, some attendees had to travel three hours each way to attend these talks. The attendance in these events ranged from 30 to 60 attendees, which is comparable to the membership count in this chapter.

About the authors: Dr. Ali Abedi is the Region-1 representative on the COMSOC North American Region Board and Chair of the IEEE Maine COMSOC Chapter. Dr. Ronald O. Brown is an independent consultant and Vice Chair of the IEEE Maine COMSOC Chapter.

Distinguished Lecturer Tour/*Continued from page 2*

on eHealth was a great success. I am a great admirer of the IEEE DL program that benefits lecturers as well as the host chapters. I would like to thank all those that made it possible, including Ewell Tan of the Asia Pacific Office and hosts Prof. Xiaofeng Tao, Chair-Beijing Chapter, Prof. Jiandong Li, Chair-Xian Chapter, Prof. Weixiao Meng, Chair-Harbin Chapter, and Prof Xinwan Li, Chair-Shanghai Chapter.

CONFERENCE CALENDAR

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

2016

JANUARY

COMSNETS 2016 — 8th Int'l. Conference on Communication Systems & Networks, 5–9 Jan.
Bangalore, India
http://www.comsnets.org/index.html

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

WONS 2016 — 12th Annual Conference on Wireless On-Demand Network Systems and Services, 20–22 Jan.
Cortina d'Ampezzo, Italy
http://2016.wons-conference.org/

ICACT 2016 — 18th Int'l. Conference on Advanced Communication Technology, 31 Jan.–2 Feb.
Phoenix Park, Pyeongchang, Korea
http://www.icact.org/

FEBRUARY

IEEE BHI 2016 — IEEE Int'l. Conference on Biomedical and Health Informatics, 24–27 Feb.
Las Vegas, NV
http://bhi.embs.org/2016/

MARCH

DRCN 2016 — 12th Int'l. Workshop on Design of Reliable Communication Networks, 14–17 March
Paris, France
https://drcn2016.lip6.fr/

ICBDSC 2016 — 3rd MEC Int'l. Conference on Big Data and Smart City, 15–16 Mar.
Muscat, Oman
http://www.mec.edu.om/conf2016/index.html

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

IEEE CogSIMA 2016 — IEEE Int'l. Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, 21–25 Mar.
San Diego, CA
http://www.cogsima2016.org/

WD 2016 — Wireless Days 2016, 23–25 Mar.
Toulouse, France
http://wd2015.sciencesconf.org/

IEEE ISPLC 2016 — 2016 IEEEInt'l.Symposium onPower Line Communications and Its Applications, 29 Mar.–1 Apr.
Bottrop, Germany.
http://www.ieee-isplc.org/

APRIL

IEEE WCNC 2016 — IEEE Wireless Communications and Networking Conference, 3–6 Apr.
Doha, Qatar
http://wcnc2016.ieee-wcnc.org/

IEEE INFOCOM 2016 — IEEE Int'l. Conference on Computer Communications, 10–15 April
San Francisco, CA
http://infocom2016.ieee-infocom.org/

WTS 2016 — Wireless Telecommunications Symposium, 18–20 Apr.
London, U.K.
http://www.cpp.edu/~wtsi/

IEEE/IFIP NOMS 2016 — IEEE/IFIP Network Operations and Management Symposium, 25–29 Apr.
Istanbul, Turkey
http://noms2016.ieee-noms.org/

MAY

IEEE ICC 2016 — IEEE International Conference on Communications, 23–27 May
Kuala Lampur, Malaysia
http://icc2016.ieee-icc.org/

JUNE

IEEE BlackSeaCom 2016 — 4th Int'l. Black Sea Conference on Communications and Networking, 6–9 June
Varna, Bulgaria
http://www.ieee-blackseacom.org/

IEEE NETSOFT — IEEE Conference on Network Softwarization, 6–10 June
Seoul, Korea
http://sites.ieee.org/netsoft/

IEEE HPSR 2016 — IEEE 17th Int'l. Conference on High Performance Switching and Routing, 14–17 June
Yokohama, Japan
http://www.ieee-hpsr.org/

EUCNC 2016 — European Conference on Networks and Communications, 27–30 June
Athens, Greece
http://eucnc.eu/

JULY

IEEE ICME 2016 — IEEE Int'l. Conference on Multimedia and Expo, 11–15 July
Seattle, WA
http://www.icme2016.org/

TEMU 2016 — Int'l. Conference on Telecommunications and Multimedia, 25–27 July
Heraklion, Greece
http://www.temu.gr/

AUGUST

EUSIPCO 2016, 29 Aug.–2 Sept.
Budapest, Hungary
http://www.eusipco2016.org/

SEPTEMBER

IEEE PIMRC 2016 — IEEE Int'l. Symposium on Personal, Mobile, and Indoor Radio Communications, 4–7 Sept.
Valencia, Spain
http://www.ieee-pimrc.org/

–Communications Society portfolio events appear in bold colored print.

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

–Sister Society conferences appear in plain black 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.

COMMUNICATIONS EDUCATION AND TRAINING: ETHICS AND PROFESSIONALISM

David G. Michelson

Wen Tong

Barry L. Shoop

Since antiquity, the difficulty of fulfilling multiple and sometimes conflicting moral obligations to different parties has been well recognized. By the 19th century, the emergence of engineering as a distinct profession was accompanied by a need to clarify the relationship between the self interest that practitioners of engineering have in advancing their careers and business interests, and their moral obligations to society, to their employers and/or clients, and to their profession.

During the late 19th and early 20th centuries, U.S. engineers formed societies to foster the development of professionalism within the major disciplines through exchange of technical information and standardization of practices. After a series of high profile bridge failures that led to significant loss of life during this period, governments began to regulate the engineering profession through formal licensing procedures and oversight conducted by independent associations and boards.

Both the societies, which promoted the interests of the profession, and the associations and boards, which represented the interests of the public, began to develop formal Codes of Ethics to guide, protect and set expectations for both their members, their employers or clients, and the public. The challenge in developing and following such codes, of course, lies in the conflicting nature of an engineer's moral obligations to each of these interested groups.

Engineering school curricula have traditionally focused on the so-called hard skills related to math, physics, computing, signal processing and other technical subjects. As the media has placed an ever brighter spotlight on the causes and consequences of errors in engineering judgment and decision making, engineering accreditation boards have mandated increases in the amount of time and effort devoted to developing learning outcomes involving the so-called soft skills related to communication, teamwork, project management skills and, importantly, responsibility to society and the public.

For engineering schools, formulating an effective approach for developing learning outcomes that help students anticipate and prepare for the ethical and professional challenges that they will face during their career has itself been a challenge. Just as traditional lectures are increasingly seen as inadequate for helping students develop technical skills and acumen, so they are considered inadequate for helping students develop the necessary ethical and professional insights.

Development of innovative or novel approaches for exposing students to ethical issues and dilemmas is not sufficient, however. It also necessary to assess the quality of the learning outcomes that are achieved in order to permit subsequent offerings to be improved. The articles that comprise this Feature Topic on Ethics and Professionalism in Communications Education and Training provides important insights.

In "Engineering Ethics Education: Aligning Practices and Outcomes," Diana Bairaktova and Anna Woodcock suggest that it is not sufficient to impart and assess ethical awareness via students' responses to vignettes concerning ethical dilemmas. Instead, they present results that suggest that current practice be extended to include motivational variables that may influence students' ethical awareness and predict their ethical behaviour.

In "Integration of Ethical Training into Undergraduate Senior Design Projects on Wireless Communications," Wilmer Arrellano, Ismail Guvenc and Nezih Pala share the ethical training framework that they have integrated into the two-semester senior design project course at Florida International University and the success that they have achieved by encouraging students to make decisions based upon ethical theories.

In "The Effect of a Stand-Alone Ethics Course in Chilean Engineering Students' Attitudes," Ruth Murrugarra and William Wallace share their experience in asking students to model and simulate ethical issues using computer-based agent simulation techniques and, importantly, assessing the impact on the student's perspectives and attitudes. The results highlight the value of diversity in the backgrounds and perspectives of participants in discussions and problem solving concerning ethical issues.

The final article in this Feature Topic focuses on Professionalism and the emergence of Telecommunications Engineering as a distinct engineering discipline. Historically, telecommunications has been classified as a sub discipline of Electrical Engineering. Tarek El-Bawab and his colleagues have recently convinced ABET that the two fields are distinct and should be treated as peers. This exciting development, which has been supported by the ComSoc Education & Training Board during the past five years, represents an historic milestone for our profession and is described in detail in the article.

Recognizing the increased interest in Engineering Ethics, IEEE recently inaugurated a biannual conference that brings together practitioners, regulators, and researchers to share recent experiences and new insights in this area. Readers are encouraged to participate in the next edition of the IEEE Ethics conference to be held in Vancouver, Canada in May 2016. Details can be found at http://sites.ieee.org/ethics-conference/.

BIOGRAPHIES

DAVID G. MICHELSON (davem@ece.ubc.ca) is with the Department of Electrical and Computer Engineering at the University of British Columbia. His research interests include antenna design, channel modeling, radar remote sensing and the scholarship of teaching and learning. He is a member of the Boards of Governors of the IEEE Communications and Vehicular Technology Societies, and is past Director of Education and Training for ComSoc. He serves as General Co-Chair of the 2016 IEEE Ethics conference.

WEN TONG (tongwen@huawei.com) is the vice president and head of wireless research at Huawei Technologies and a Huawei Fellow. He leads one of the largest wireless research organizations in the industry with more than 700 research experts. During a twenty-year career, he has pioneered fundamental technologies in wireless with 210 granted US patents. He also serves on the Board of Directors of the WiFi Alliance and the Board of Directors of the Green Touch Consortium.

BARRY L SHOOP (Barry.Shoop@usma.edu) is the 2016 IEEE President and Head of the Department of Electrical Engineering and Computer Science at the U.S. Military Academy at West Point. He is responsible for an undergraduate academic department with over 79 faculty and staff that engage over 1800 students in electrical engineering, computer science, and information technology each year. His research interests include optical information processing, neural networks, image processing, disruptive innovations and educational pedagogy.

Engineering Ethics Education: Aligning Practice and Outcomes

Diana Bairaktarova and Anna Woodcock

Diana Bairaktarova is with Virginia Polytechnic Institute and State University.

Anna Woodcock is with California State University San Marcos.

ABSTRACT

Analysis of the strengths and weaknesses of current efforts indicate that engineering programs lack consistent, accurate, and reliable methods of teaching professional ethics and measuring their outcomes. This raises two equally important issues: how we teach ethics and which student outcomes we are assessing. Engineering students are performing poorly on the Ethics part of the Fundamental Engineering exam, so clearly there is a misalignment between teaching practice and outcomes. Engineering ethics instruction is often focused on the instruction of moral judgment and assessing ethical awareness via students' responses to vignettes describing ethical dilemmas. In this study, we propose extending current practice from a focus on teaching moral reasoning to also considering students' ethical awareness and future behavior. We introduce motivational variables that engineering educators should consider when designing ethics curricula. The study findings suggest that these motivational factors may influence students' ethical awareness and predict their ethical behavior.

INTRODUCTION

In a comprehensive review with diverse samples across United States engineering schools, Colby and Sullivan (2008) describe the current status of how undergraduate engineering education supports students' ethical development. Their analysis of the strengths and weaknesses of current efforts indicate that engineering programs lack accurate and reliable methods of teaching professional ethics and measuring outcomes. Engineering ethics education is generally conducted by introducing the Code of Ethics of the profession and reviewing case studies. The discussion of case studies is brief and mainly done so students can apply the different cannons of the Code of Ethics. Some of the case studies are real and depict major failure events. In the presentation of such cases, educators hope that the moral learned will leave a lasting impression on students. The other commonly used instructional techniques involve presenting short abstracts and more ambiguous vignettes that are explicitly tied to the Code of Ethics. These vignettes present ethical dilemmas in more ambiguous ways, but educators hope that when students are introduced to and become familiar with the Code of Ethics they will easily recognize a breach of ethics. The same type of vignettes are presented in the professional ethics section of the Fundamental Engineering (FE) exam as multiple-choice questions, yet students' performance on that part of the exam is not optimistic.

In addition, although the importance of engineering ethics education and its significance are highlighted by the ABET and National Academy of Engineering, ethics is included sparsely in the undergraduate curricula and is still not a required stand-alone course. In many College of Engineering courses, ethics are introduced in a few lectures in other engineering courses or professional development classes, and often the ethics section assessment is not even graded. Despite its importance, students do not take ethics seriously. Instead of developing their professional identity at school, students generally gain such an identity from family and friends. Engineers' professional ethics tend to mostly be an extension of their personal ethics. Instruction on ethics generally serves only to encourage students to act ethically based on their personal beliefs. This instructional approach introduces a problem that we are dealing with in engineering ethics education, namely, the misalignment between the importance, content, pedagogy, and assessment used in teaching ethics.

Given this state of affairs, we argue that we should take a step back and question what we are and should be actually assessing: Students' ability to remember the cannons of the Code of Ethics? Their ability to think logically? Their moral reasoning or ethical awareness? Their ethical behavior? We claim that our engineering students are ill-prepared to act professionally in the workplace, yet we may be using the wrong assessment to make such a claim.

COMPLEXITY OF ETHICS

Ethics dilemmas are complex and the motivation to act ethically is different for everyone. Ethics are influenced by factors such as culture, family, level of maturity, social status, and personality traits. Although psychologists have long studied the development of moral identity (Rest 1986a, Kohlberg 1984), there is sparse research focused on engineering students' moral development, or subsequent ethical behavior. Knowing what is

0163-6804/15/$25.00 © 2015 IEEE

appropriate and ethical does not guarantee ethical behavior. When teaching engineering ethics, if we are instructing students to be moral members of society, then instead of assessing their ability to exercise moral judgment, we need to find a way to access their ethical awareness and to gauge their future ethical behavior. We teach engineering ethics using pedagogical methods that focus on abstract ethical frameworks and moral justification instead of engaging students in real-life engineering activities and situations where they would have the opportunity to practice ethical awareness and ethical behavior. Literature suggests that in engineering ethics education we are more concerned about justifying the inclusion of ethics in the engineering curricula, than whether a program or course has a positive impact on students' ethical behavior. Measurement and instruction of engineering ethics is heavily influenced by the notion of moral reasoning; however, the ability to exercise moral judgment about what is right and wrong does not guarantee that students are ethically educated or likely to behave in an ethical manner.

Rest posits that moral reasoning entails at least four distinct psychological processes: *moral sensitivity*, *moral judgment*, *moral motivation*, and *moral character and implementation*. However, there is a dearth of research linking moral reasoning with ethical or moral behavior. One study reported assessment of moral reasoning using the Rest Define Issues Test (DIT). The DIT was associated with cheating behavior whereby individuals low on moral judgment cheated more and sooner than those with higher moral judgment. However, the study also revealed that even those with high moral judgment cheated given sufficient temptation (Malinowski and Smith, 1985). Although moral reasoning may facilitate ethical awareness, it seems insufficient to predict ethical behavior. We argue in the spirit of Rest's four processes of moral reasoning, the measurement and instruction of ethical awareness, ethical intentions, and ethical behavior are what is needed in engineering ethics education, rather than only instruction and assessment of moral reasoning. Some work has been conducted to address ethics assessment in engineering instruction; however, this work focuses on measuring moral reasoning, not measuring ethical intentions and behavior. For example, Shuman and colleagues (2005) presented a scoring rubric to assess students' ability to recognize and resolve complex, open-ended ethical dilemmas. The rubric is based on five components: recognition of dilemma; recognition and appropriate use of facts; analysis; consideration of multiple points of view; consideration of risk and solution.

A decade ago, researchers started to consider other psychological variables in the measurement of engineering ethics. Recently, Harding and colleagues (2012) proposed a new model of ethical decision-making that combined students' demographic, academic, and motivational variables with moral reasoning to predict self-reported cheating behavior on college tests. We propose that this line of thinking be extended, and that understanding which factors influence ethical awareness and subsequent decisions to behave ethically must be a vital part of engineering ethics education. We claim measuring moral reasoning is insufficient for assessing outcomes of engineering ethics education because moral awareness has not been demonstrated to translate into ethical behavior. We argue that classes with real life scenarios and open-ended questions promoting discussion will provide students with a more complete exposure to engineering ethics, while raising the profile of engineering ethics at the same time. We argue that when teaching ethics we need to consider who our students are in regards to their individual differences and how these personality traits influence ethical decision-making and behavior under both real and hypothetical conditions. In particular, we are interested in understanding how students' perceived volitional control over their ability to act ethically and students' orientation toward the people in their environment may influence their ethical awareness.

Spheres of Control

Spheres of control reflect the degree to which people feel they have agency or control over specific areas of their lives. Paulhus and Van Selst (1990) describe three spheres or domains for control: personal efficacy, interpersonal control, and sociopolitical control. Individuals differ in the degree to which they feel they have volitional control in different domains. The domain closest to the core self is personal efficacy, which concerns motivation to have mastery or control over objects, events, and tasks. The next sphere is interpersonal control, which concerns regulating and controlling relationships with other people in dyads and groups. The outermost layer is sociopolitical control, which concerns controlling institutions and organizations. Ethical dilemmas vary in terms of the sphere or spheres in which volitional control is necessary to act in an ethical manner. We predict the level of control individuals have in each of the spheres of control should influence ethical awareness and intentions to behave ethically when faced with dilemmas in those domains.

Ethical dilemmas vary in terms of the sphere or spheres in which volitional control is necessary to act in an ethical manner. We predict the level of control individuals have in each of the spheres of control should influence ethical awareness and intentions to behave ethically when faced with dilemmas in those domains.

Person and Thing Orientations

Individuals selectively orient to both the people and things in their environment. The extent to which students are interested in the people and things in their environment is strongly associated with the college major they select and the career they choose to pursue (Graziano *et al.*, 2012; Su and Rounds, 2015). Not surprisingly, engineers and engineering students are typically highly thing oriented, but many are also quite highly person oriented (Woodcock *et al.*, 2013). Engineers vary quite widely in their orientation to people, which may be an important point in ethical awareness, which often involves an understanding of the motives and intentions of other people. The degree to which engineers-in-training orient to the people and interpersonal interactions in their environment should be predictive of their ethical awareness, particularly in ambiguous ethical dilemmas.

Research

To test the relationship between ethical awareness, person orientation, and the three spheres of control, we conducted a study with 190 undergraduate engineering majors enrolled in an upper-level undergraduate course. The participants were primarily male (73 percent), in either their third or fourth undergraduate year (73 percent), majoring in civil, electrical, or mechanical engineering, and reported they had previously taken an ethics class (72 percent). The students completed validated measures of person orientation and the spheres of control (Paulhus and Van Selst, 1990), read two vignettes of ethical dilemmas (adapted from the Lockheed Martin Ethics Challenge Game, which is commonly used in engineering ethics classes), and indicated how they would respond to each situation by selecting one of four possible responses. Each vignette had one answer that was ethically correct.

Sixty percent of the students correctly responded to at least one of the ethical dilemmas; only 10 percent responded correctly to both. We predicted students who had previously taken an ethics class would perform better on the ethical dilemma tasks than those students who had never taken an ethics class. However, we found no impact of having previously taken an ethics class, ($\chi^2 = 3.76$, $p = .15$). We predicted the degree to which students were oriented toward the people in their environment (person orientation) would positively predict ethical awareness. However, our data did not support this hypothesis ($r = -.056$, $p = .45$). This was the case for both male and female students in the sample, and held true regardless of whether students had taken an ethics course.

We also predicted students' scores on the three spheres of control measures would be associated with ethical awareness. The only support we found was that while sociopolitical control was unrelated to ethical awareness for male engineering students ($r = .12$, $p = .14$), it was negatively related to ethical awareness for female engineering students ($r = -.31$, $p = .02$) (Bairaktarova and Woodcock, 2014).

Our finding that taking an ethics class did not increase students' subsequent ethical awareness supported our view that the ethics curriculum and forms of presenting and discussing engineering ethics needs to be revised. We predict classes with more real-life scenarios and open-ended questions promoting discussion will provide students with a more complete exposure to engineering ethics, while at the same time raising the profile of engineering ethics. While questions on the FE are presented in multiple choice format, students must be prepared to face complex ethical dilemmas for which there is not always a single, straightforward answer. We felt including more realistic scenarios with the opportunity for students to respond in their own words, rather than via multiple choice, would provide a better test of ethical awareness and the role our individual difference measures.

We developed three ethical scenarios adapted from the experience of one of the author's engineering work practice. These scenarios involved ethical dilemmas that were more complex than the traditional vignettes used in the first study. We presented them in addition to three ethical vignettes adapted from the Lockheed Martin Ethics Challenge Game.

We conducted an exploratory study to assess whether students' ethical awareness would be influenced by the nature and complexity of the ethical dilemma, and the way they would respond to the scenarios. We developed three ethical scenarios adapted from the experience of one of the author's engineering work practice. These scenarios involved ethical dilemmas that were more complex than the traditional vignettes used in the first study. We presented them in addition to three ethical vignettes adapted from the Lockheed Martin Ethics Challenge Game.

We predicted students would display greater ethical awareness in response to the more traditional situations that are simulated and involve a clearer statement of potential breach of ethics, such as misconduct, conflict of interest, and intellectual property. We predicted person orientation would be positively associated with both types of scenarios where students provide an open-ended response, but more strongly for the more realistic scenarios, as they require more awareness of the motives of others. We predicted students' spheres of control would be more sensitive to the scenarios than the vignettes. Finally we were interested in the effect of having already taken an ethics class on students' ethical awareness in response to the vignettes and the scenarios.

The participants were 40 undergraduate engineering major students enrolled in a professional development course at a large Midwest university. During this course students are expected to develop an understanding of engineering ethics, teamwork, leadership, and professional responsibility. The students were predominantly male (88 percent), mainly from electrical, mechanical, or industrial and systems engineering majors (68 percent), and spoke English as their first language (85 percent). Only 29 percent reported that they had previously taken an ethics-related class. Participants responded to demographic questions, indicated whether they had ever taken an ethics class, read the three ethical vignettes and three scenarios, and were asked how they would respond to each situation by providing short typewritten answers. Once they responded to the cases, the students completed the person and thing orientations scale (Graziano, Habashi, and Woodcock, 2011), the spheres of control scale (Paulhus and Van Selst, 1990), and answered questions related to their general attitudes toward behaving ethically as an engineer.

Three raters independently evaluated the students' responses to each of the six ethical cases using the following scoring: 0 = no response, 1 = incorrect response, 2= correct response. One researcher completed the initial coding. In order to establish reliability, two other coders recoded the data using the same coding scheme. Reliability was calculated by comparing the number of agreements and disagreements of each researcher with the initial coding and calculating the average percentage of agreement. The inter-rater reliability was found to be 87 percent. The scores from the three Lockheed vignettes were totaled to create a vignettes (V) score, and the scores from the workplace scenarios were totaled to create a scenarios (S) score.

The V and S scales were not strongly correlated ($r = .23$, $p = 14$), which indicates they were qualitatively different. Contrary to our prediction, students scored significantly higher on the workplace scenarios ($M = 5.51$) than the traditional vignettes ($M = 4.22$) (Fig. 1). Eighty five percent of students gave a correct response to at least two of the three real-life scenarios compared with 44 percent for the vignettes. Also, contrary to our predictions, person orientation was related more strongly to the responses to the vignettes ($r = .36$,

p = .02) than the scenarios (r = .18, p = .27). Finally, consistent with expectations and our previous findings, neither the score on the vignettes nor the scenarios were influenced by having attended a previous ethics classes.

The three ethical vignettes were ambiguous and abstract, but they clearly presented situations where issues such as opportunity for misconduct, conflict of interest, and intellectual property could be easily recognized. However, our results show that students had quite a difficult time recognizing these issues and tended not to recommend appropriate actions. Students' responses on the scenarios showed a significant improvement in ethical awareness. Our finding that ethical awareness in response to both the vignettes and the scenarios was not influenced by having attended a previous ethics class is disappointing, yet not surprising. Considering the review on the status of current engineering ethics education, our finding is in alignment with the claim that engineering ethics education is not sufficient.

The awareness of a breech, or potential breech, in ethics is an important antecedent to selecting the most ethical course of action. Ethical dilemmas, both real and simulated, run the gamut from clear to ambiguous. A situation where behavior is ambiguous leaves the detection of a possible ethical breech in question. We hypothesized that individual differences in orientation to the social environment may help determine how individuals detect ethical breeches in ambiguous situations. We expected person orientation would be positively associated with detection of ethical awareness in response to both the vignettes and the scenarios, but more strongly for the scenarios. Our results indicated person orientation was related more strongly to ethical awareness in response to the vignettes rather than the scenarios. The ethical dilemmas in the scenarios were more complex than the vignettes and included both technical and ethical aspects. Engineers are extensively trained to design, test, and improve artifacts, systems, and processes through problem solving, so it is possible engineering students' solutions were driven by finding a solution for the technical aspect of the problem rather than the ethical dilemma the situation presented. This suggests that real-world ethical dilemmas likely involve finding a solution to problems with both technical and interpersonal aspects, and that student's may differentially assess ethical dilemmas with respect to these aspects. This finding is exemplified in engineering student's responses to one of our ethical scenarios where students were faced with choosing between two polymers for use in a biomedical device: one that is readily available, but comes with a low risk of damaging human skin, or one with no known risks but with a significantly higher price and a long delivery time.

"I think it depends on the ***application of the product****. If the product is in contact with human skin regularly then the newly developed polymer should be used. If the product rarely comes into contact with skin then the risks should be made clear to the handlers and the first one can be used."*

"Depending on how much ***research*** *has been done on the new polymer, I might choose to use the new polymer. However, if there has not been much* ***testing*** *on the new polymer, I would most likely use the polymer with potential risks associated, as long as there was a warning on the product to label the possible risks."*

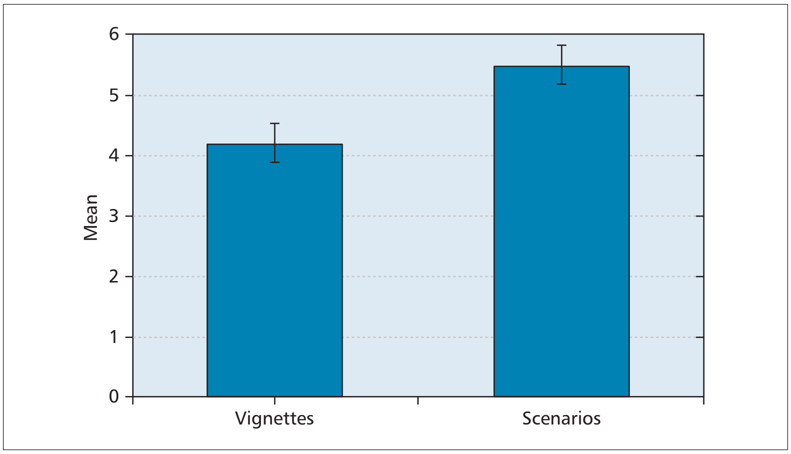


Figure 1. Ethical awareness in response to the vignettes and scenarios. Note: Error bars represent 95 percent confidence intervals.

Even though in both responses the students do display some regard for the safety and well being of the user, this concern does not come first. They are viewing the problem extensively by considering the product application or the process of testing. In the design of a new biomedical device, human safety and health must come first before any other concerns, profit, or deadlines.

The vignettes and the scenarios also differed with respect to which ethical decision was under the engineers' control. While the three vignettes required recognition of a potential breach of ethics, the decision to act ethically was entirely up to the engineer; but the three scenarios presented ethical dilemmas that varied with respect to the degree to which acting ethically or refusing to act unethically was within the engineers' control. For example, Vignette 1 asked the students to recognize an issue of misrepresentation of data; Vignette 2 required students to be able to recognize potential conflicts of interest; Vignette 3 provided students with a situation where they should recognize an issue of intellectual property. In Scenario 1, the decision to act ethically or unethically when choosing the polymer to use in the biomedical device was almost completely under the individual engineer's volition. In Scenario 2, students were presented with a situation that required the cooperation of others to behave in an ethical manner. In Scenario 3, the students were asked to recognize a potential breach of ethics that was embedded within a context of unethical behavior that runs through an entire project team, organization, or agency.

In addition to looking at students' ethical awareness by the recognition of a potential breach of ethics in the scenarios, students were also asked to recommend an ethical action. The three scenarios were real-life examples of situations engineers face, and are also qualitatively different with regard to an individual engineer's personal volitional control. We had anticipated the three spheres of control exemplified in the scenarios would be differentially related to students' spheres of control scores. This was only the case for the personal efficacy sphere, which was a significant predictor of ethical awareness in all three scenarios. One plausible explanation is that engineering students' sense of personal efficacy motivates them to report that they would act ethically in response to all types of ethical dilemmas. It is possible that as engineers-in-training become more seasoned, their awareness of the range of volitional control they have across different ethical dilemmas may increase.

Follow-up research is underway to test more scenarios and increase the reliability of such case studies for predicting engineering students' ethical awareness and behavior. The findings from this line of research could be specifically useful for engineering educators employing project-based pedagogies in courses with inclusion of ethical issues.

Below are some of the engineering student's responses to two of the ethical scenarios involving personal volitional control.

"Understanding the personal risk, either use a "whistleblower" hotline within the company if available, or (understanding the pressure on the electrical engineer) look to inform the CEO that this unethical practice will not be carried out by yourself and should not be carried out at all" [student response to one of Scenario 2, cooperation of others to behave in an ethical manner].

"I would write a white paper and call the agency's hotline or ask my manager. If neither worked out, I would continue up the food chain until I was either satisfied that there was no foul play or the necessary changes are made" [student response to one of Scenario 3, unethical behavior that runs through an entire project team, organization, or agency].

We present these results as preliminary, as the small sample size in this pilot limits the statistical power of the tests. Further research will need to build and test a bank of scenarios for each sphere of control to increase the reliability of them as a primer for ethical awareness.

Implications for Engineering Ethics Education

These findings suggest several implications for engineering ethics education. We first propose the use of more realistic scenarios with the opportunity for students to respond in their own words, rather than via multiple choice. We suggest when assessing students' outcomes of an engineering ethics class, instructors should use scenarios that look at ethical awareness and intention to act ethically rather than only vignettes with multiple choice questions to assess students moral judgment and recognition of ethical dilemmas. Presentations of more complex scenarios with the involvement of practicing engineers who experience ethical issues first hand, we argue, will promote discussion and alignment with practice. These scenarios should vary with respect to which students are required to consider the motives of others, and with respect to the amount of volitional control students have to enact ethical behavior. Lastly, we suggest the use of pedagogical methods that include graded assessment activities to increase students' motivation to take engineering ethics courses seriously.

In conclusion, this study proposes extending current practice from teaching moral reasoning only to considering students' ethical awareness and future behavior. We introduced motivational variables that engineering educators need to consider when designing engineering ethics curricula. The study findings suggest that these motivational factors may influence students' ethical awareness and predict their ethical behavior. Follow-up research is underway to test more scenarios and increase the reliability of such case studies for predicting engineering students' ethical awareness and behavior. The findings from this line of research could be specifically useful for engineering educators employing project-based pedagogies in courses with inclusion of ethical issues.

Additional Reading

[1] J. Colby and W. M. Sullivan, "Ethics Teaching in Undergraduate Engineering Education," *J. Engineering Education*, 2008, pp. 327–38.

[2] J. Rest, 1986a, *DIT Manual*, Minneapolis, MN: Center for the Study of Ethical Development, University of Minnesota; J. Rest, (1986b). *Moral Development: Advances in Theory and Practice*, New York, NY: Prager.

[3] L. Kohlberg, "The Psychology of Moral Development: The Nature and Validity of Moral *Stages," Essays on Moral Development*, vol. 2, 1984, Harper & Row.

[4] C. I. Malinowski and C. P. Smith, "Moral Reasoning and Moral Conduct: An Investigation Prompted by Kohlberg's Theory, *J. Personality and Social Psychology*, vol. 49, 1985; doi: 10.1037/0022-3514.49.4.1016, pp. 1016–27.

[5] L. J. Shuman, M. Besterfield-Sacre, and B. M. Olds, "Ethics Assessment Rubrics," L. A. *C. Mitcham, Encyclopedia of Science, Technology, and Ethics*, vol. 2, 2005, New York: Mac Millan Reference Books, pp. 693–95.

[6] T. S. Harding, D. D. Carpenter, and C. J. Finelli, "An Exploratory Investigation of the Ethical Behavior of Engineering Undergraduates," *J. Engineering Education*, vol. 101, no. 2, 2012, pp. 346–74.

[7] D. L. Paulhus, and M. Van Selst, "The Spheres of Control Scale: 10 Years of Research," *Pers. Indiv. Differ.*, vol. 11, no. 10, 1990; doi: 10.1016/0191-8869(90)90130-J, pp. 1029–36.

[8] W. G. Graziano *et al.*, "Orientations and Motivations: Are You a "People Person," a "Thing Person," or Both?" *Motivation and Emotion*, vol. 36, no. 4, 2012, 2012/12/01, doi: 10.1007/s11031-011-9273-2, pp. 465–77.

[9] R. Su and J. Rounds, "All STEM Fields are not Created Equal: People and Things Interests Explain Gender Disparities Across STEM Fields," *Frontiers in Psychology*, vol. 6, 2015, doi: 10.3389/fpsyg.2015.00189.

[10] A. Woodcock *et al.*, "Person and Thing Orientations: Psychological Correlates and Predictive Utility," *Social Psychological and Personality Science*, vol. 4, no. 1, 2013, pp. 117–24.

[11] D. Bairaktarova and A. Woodcock, "The Role of Personality Factors in Engineering' Students Ethical Decisions," *Proc. IEEE Conf. Engineering Ethics*, May 21–23, 2014, Chicago, Illinois.

[12] W. G. Graziano, M. M. Habashi, and A. Woodcock, "Exploring and Measuring Differences in Person-Thing Orientation," *Pers. Indiv. Differ.*, vol. 51, 2011, doi: 10.1016/j.paid.2011.03.004, pp 28–33.

Biographies

Diana Bairaktarova (dibairak@vt.edu) is an assistant professor in the Department of Engineering Education at Virginia Tech. She has more than 10 years of experience working as a mechanical design engineer. Through real-world engineering applications, Bairaktarova's research spans from engineering to psychology to learning sciences, as she uncovers how individual performance and professional decisions are influenced by aptitudes, personal interests, and direct manipulation of material objects.

Anna Woodcock is a social psychologist and a research faculty member at the California State University San Marcos. Her research interests lie in the broad areas of diversity, prejudice, and stereotyping. She conducts research on individual differences in motivations to pursue science, technology, engineering, and mathematics (STEM) careers, and contextual factors that promote and reinforce social disparities such as the underrepresentation of women and minorities in STEM.

IEICE Transactions on Communications

Call for Position Papers

Exploring Innovative Intelligence and Technologies in Communications

The IEICE Transactions on Communications announces that it starts accepting submissions of papers for a new category, POSITION PAPER, on October 1, 2015.

Features

POSITION PAPER is defined as an article with at least one of the following features, potentially having a significant influence to the theory and practice of communication technology:

- Introduction of novel viewpoints, frameworks, and/or paradigms
- Proposal of technologies or methods based on innovative ideas (not just extension of existing technologies)
- Novel ideas with the potential to bring about innovative technologies
- Prototyping or experimental results with noteworthy features, such as achievement of the world's best performance or the world's first realization

Only papers having the significant feature(s) above will be accepted, while comprehensive evaluations of the performance or the effectiveness will not be required.

Scope

Whole research areas covered by the IEICE Transactions on Communications, such as:

- Fundamental Theories for Communications
- Energy in Electronics Communications
- Transmission Systems and Transmission Equipment for Communications
- Optical Fiber for Communications
- Fiber-Optic Transmission for Communications
- Network System
- Network
- Internet
- Network Management/Operation
- Antennas and Propagation
- Electromagnetic Compatibility (EMC)
- Wireless Communication Technologies
- Terrestrial Wireless Communication/Broadcasting Technologies
- Satellite Communications
- Sensing
- Navigation, Guidance and Control Systems
- Space Utilization Systems for Communications
- Multimedia Systems for Communications

Further information

Please visit the journal web site at : **http://www.ieice.or.jp/cs/jpn/EB/**

*POSITION PAPER is available by **open access** through the website of the IEICE Transactions on Communications.

*POSITION PAPER submitted by September 30, 2017 will be exempted from the page charges up to the fee required for 50 reprints of four pages with electronic source data using the IEICE LaTeX style file.

The Institute of Electronics, Information and Communication Engineers

Integration of Ethical Training into Undergraduate Senior Design Projects on Wireless Communications

Wilmer Arellano, Ismail Guvenc, and Nezih Pala

The authors are with Florida International University.

ABSTRACT

Wireless communications engineers may face complex ethical dilemmas while designing products for consumers. At Florida International University's Electrical and Computer Engineering Department, we place great importance on training our students to address such ethical dilemmas, in alignment with ABET requirements. To this end, an ethical training framework is integrated into our two-semester senior design project course. In addition to the codes of ethics, our students use "The Theory Model" to make decisions based on ethical theories. For the solution of ethical dilemmas by means of ethical theories, they apply a modified version of the line drawing method. In this article we briefly explain the theories and methods that our students consider when facing ethical dilemmas. Then we present excerpts from four past senior design projects related to wireless communications. We also summarize the methodology the students use for identifying the best option to address ethical dilemmas.

INTRODUCTION

The importance of senior design courses and projects as a fundamental source for documenting the achievement of ABET's student outcomes [1] is widely recognized [2–4]. Some of ABET's indicated outcomes such as "(c) an ability to design a system, component, or process to meet desired needs within realistic constraints such as economic, environmental, social, political, *ethical*, health and safety, manufacturability, and sustainability", and "(f) an understanding of professional and *ethical responsibility*" [1], are often used as standards to allow the use of senior design courses as an assessment tool during accreditation. We can safely say that instructors entice their senior design students to illustrate their understanding and fulfillment of the needs expressed in outcomes (c) and (f) when writing the final project reports.

At Florida International University's Department of Electrical and Computer Engineering (FIU-ECE) we see senior design as a multifaceted experience that we often summarize as "project management with an example." The example is represented by the students' project designs in compliance with ABET's outcome (e): "an ability to identify, formulate, and solve engineering problems." On the other hand, project management is inspired by ABET's outcomes (c) and (f) described earlier, as well as outcomes (d) and (g), which relate to communication and team work [1]. Our senior design courses have been certified by the Writing Across the Curriculum (WAC) program and the Office of Global Learning Initiatives at FIU. With these, we aim to teach our students minimization of the barriers to trade and global success. In particular, we discuss topics related to the World Trade Organization (WTO), international standardization bodies, the International Electrotechnical Commission (IEC), the International Organization for Standardization (ISO), and the International Telecommunication Union (ITU). Our Senior Design Experience is divided into two semesters. During the first semester students learn topics directly associated with ABET's outcomes and write a comprehensive project proposal. During the second semester, students execute the project design and write the final report. There are several senior design projects every semester related to wireless communications and networks.

At FIU-ECE we place great importance on *ethical considerations* and *societal impact* in senior design projects, in alignment with ABET's outcomes (c) and (f). When dealing with ethics in senior design, recent works [5–7] use "The Code Model" for analyzing the code of ethics for engineering societies and corporations. The problem with this approach is that sometimes ethical codes do not address an ethical situation adequately, or do not address it at all. For this reason, in addition to the codes of ethics, we introduce the students to the process of making decisions based on ethical theories: "The Theory Model." This model provides guidance in making ethical decisions when the code model is not adequate. For the solution of

0163-6804/15/$25.00 © 2015 IEEE

ethical dilemmas by means of ethical theories, we use a modified version of the line drawing method presented in [8].

When preparing their project proposal and the final report, our students show their understanding of IEEE's code of ethics and further reflect on some of its canons that particularly relate to their project. In addition, our students must perform a broad search and identify an ethical issue that relates to the project, which in the event that it happens, would not be adequately analyzed by the code of ethics. Subsequently, the students are required to use ethical theories based on the *Theory Model*. This leads to interesting analysis of ethical issues by the students, where ethical analysis is sometimes highly hypothetical, but nevertheless it creates awareness for the students that sometimes engineers should step a little bit into philosophy.

The rest of this article is organized as follows. The following section shows the elements that our students consider when facing ethical dilemmas and societal impact while they are working on their senior design projects. Then we present excerpts from four past senior design projects related to wireless communications, as examples of the students' ethical considerations. The following projects are selected to demonstrate the results of our approach:

- uPark Assistant for vehicular ad-hoc networks
- GSM Relay Copter
- Copter Triangulation Locator
- Visible Light Communications

The final section concludes the article.

Ethical Considerations and Societal Impacts in Senior Design Projects

At FIU-ECE we consider that IEEE and its code of ethics represent important guidelines for both electrical and computer engineering majors. In their project proposals and final reports, our students must explicitly express that they are in compliance with IEEE's code of ethics, and highlight those canons when describing the main significance to their project. Students must find a complex dilemma related to their project, which cannot be adequately addressed by the code. Subsequently, this complex dilemma has to be resolved using the *Theory Model*.

Finding a complex ethical dilemma is not always an easy task. In order to help the students in this process, we suggest that the students consider these possibilities:

- Projects that pose an ethical dilemma. To illustrate this approach we analyze the case of "Robot workers versus human workers" [9]. In this example, the workers can be negatively affected by losing their job to a machine.
- Projects that help solve an ethical dilemma. We illustrate this approach with the case of "The Intelligent Wheel Chair" [10]. In this example, the handicapped individuals are positively impacted when an intelligent wheel chair helps them obtain a job.
- Projects in which flaws in the design create an ethical dilemma. We illustrate this approach with "Silicon Valley Programmer Indicted for Manslaughter" [11]. In this fictional scenario, the harm caused by a faulty design is shown.

From the many different ethical theories available, we made a selection of four theories that we believe have strong relation to engineering ethics. These theories are:

- Utilitarianism: The morally right action is the one that generates the greatest benefit (or least amount of harm) for the greatest number of individuals.
- Ethical Egoism: The morally right action is the one that safeguards and/or promotes your own or your organization's best interests.
- Kantian Ethics: The morally right action is the one that is based on rules, obtained by rationality, and will also be followed in similar situations.
- Rights Ethics: The morally right action is the one that respects society and the individual's rights.

The first two theories are called *Consequentialist*, as the goodness of actions is based on the consequence of those actions. The last two theories are referred to as *Non-Consequentialist*, and they represent two faces of the same coin. Kantian Ethics basically establishes that we have the duty to perform the morally correct actions, but at the same time, each duty is associated with other people's right. Our students use Kantian and Rights Ethics in their analysis, as looking from the opposite points of view of duty and right helps in the analysis.

In [8] the author proposes a "line drawing method" to solve ethical dilemmas. A line segment is drawn; the left end is associated with a morally incorrect extreme, while the right end is associated with a morally correct extreme. Several options are presented to solve the dilemma and placed on the line according to their level of goodness. An option that is better than the other should be closer to the right. The solution would be the option closest to the right end. The problem with this method would be that the options could be placed on the line subjectively. To help minimize this problem, we use a scoring system where we chose the option with the highest score. The score for each option is obtained by summing up how well it satisfies the different ethical theories. A score of 0 represents not satisfying the theory, while a score of 1 represents full satisfaction. Accepted score values are 0, 0.25, 0.50, 0.75, and 1.

We illustrate the method by presenting the options and scores of one of the student's examples that we present later. The team designing the uPark Assistant [12], after selecting their topic, found that privacy is an important issue when designing VANET applications. Ignoring the problem would be unethical while solving the problem would prevent them from completing the project on time. In Table 1 we show the options they presented, while in Table 2 we show the scores they determined. Options 2 and 3 were highly egoistic and received very low scores. The team chose Option 1 as their solution. This option produced the greatest good, respected

At FIU-ECE we consider that IEEE and its code of ethics represent important guidelines for both electrical and computer engineering majors. In their project proposals and final reports, our students must explicitly express that they are in compliance with IEEE code of ethics, and highlight those canons on main significance to their project.

Option #	Description
1	Tell users a new feature will be added to the design in order to prevent access to this information.
2	Ignore the chances that the location tracking information can be misused.
3	Let the users know this may happen, but do not offer any help.

Table 1. Possible solutions.

	Theories				
Option #	Utilitarianism	Egoism	Rights	Kantian	Score
1	1.00	0.00	1.00	0.50	2.50
2	0.00	1.00	0.00	0.00	1.00
3	0.00	1.00	0.25	0.00	1.25

Table 2. Possible solutions score.

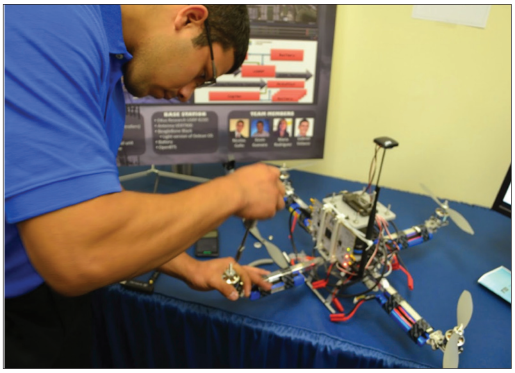


Figure 1. A senior design student at FIU, working on the GSM relay copter prototype [14].

people's rights, and the score to become a rule was 0.5. The score to become a rule was not perfect, because the solution would be available after the product sold and not at acquisition time.

Students are also required to write an essay-type sub-section on social impact. They need to write about how the project will contribute to local culture and global culture. In this section they include data obtained using two instruments: a survey they are required to create and run, and international interviews they are required to conduct. Both instruments gather acceptance questions along with technical questions from potential users.

Sample Projects and Ethical Components

In this section we will present four example senior design projects related to wireless communications and networks, which have been recently carried out at FIU-ECE. We will also discuss how the students have handled ethical aspects of the projects.

Park Assistant

uPark Assistant is a VANET application intended to help drivers find parking spots efficiently. In this system vehicles interact with a centralized server in a parking lot/garage. Upon connection and service acceptance, the system displays the nearest vacant parking location, and based on the GPS coordinates provided by the vehicle's onboard unit, directs the driver to the nearest vacant location. Once parked, the system keeps track of the vehicle's position and length of stay. Lastly, after the vehicle leaves the location, the system will store the parking lot usage data and update the database to make the spot vacant for the next vehicle.

After some research and analysis, the students became aware that privacy would be an issue, since the communication with the vehicle could provide information on the driver's location. The students analyzed the dilemma using the theory model, and considering the following options:

1. Inform users that a new feature will be added to the design in order to prevent access to this information.
2. Ignore the chances that the location tracking information can be misused;
3. Let the users know that this may happen, but do not offer any help.

The resolution of the ethical dilemma, after applying the method, was "1. Inform users that a new feature will be added to the design in order to prevent access to this information." The implementation was left for a hypothetical feature as this change would have exceeded the allotted time to complete the project.

GSM Relay Copter

Wireless communications and network infrastructure can be damaged during natural disasters and terrorist attacks. For example, the 2011 Tsunami in Japan severely damaged the cellular network infrastructure, which negatively affected search and rescue operations, emergency communications, and coordination among first responders [13]. In this senior design project, the students developed an unmanned aerial vehicle (UAV) system based on a software defined radio (SDR) platform (see Fig. 1), which can be rapidly deployed for use in emergency communication scenarios [14]. In particular, the UAVs can serve as flying GSM base stations and provide cellular network coverage to users within their vicinity. The universal software radio peripheral (USRP) equipment is used as the SDR equipment, and they are controlled via Beagle Bone boards that run OpenBTS software for implementing the GSM technology.

After some research, the students identified the following ethical and privacy related issues related to the use of the developed technology:

- The technology can be abused by those savvy enough to use the device to disrupt calls, push their own information onto a network, or drop into private conversations. The data that can be acquired by such means is desired by those who commit fraud.
- The physical device can be hazardous to those who are not cautious of the blades of

the propellers, which move with sufficient speed to injure an irresponsible operator.
- The batteries can combust if not properly maintained and might burn users who are not cautious while servicing the device.

The students came up with four options to deal with the ethical issues:
1. Add a warning label about rogue users.
2. Restrict access to users by using a user code or number.
3. Implement frequency hopping of data so other parties cannot easily sniff data.
4. Issue user or code number and frequency hopping.

After utilizing the scoring mechanism as in Table 3 for four different ethical theories, option 4 received the highest score, and students proceeded with this option.

	Theories				
Option #	Utilitarianism	Egoism	Rights	Kantian	Score
1	0	1	0	0	1
2	0.25	0.5	0.25	0.25	1.25
3	0.5	0.75	0.75	0.5	2.5
4	1	0.25	1	1	3.25

Table 3. Possible solutions score.

The students also investigated the social impacts of their project. For example, search and rescue operations were identified as a major application, in which their device can be used to create communications in places where a person is lost. As an example, they specified the Florida Everglades, which is a dead spot for communication. By creating a private GSM network in the Everglades, a stranded person can call for help. Another interesting use case is for crashed planes where a large area needs to be covered. The GSM Network Copter can comb the area and look for cell phones that are active. Once a cellphone is identified, the GSM Network Copter can notify the search and rescue team of the location.

Copter Triangulation Locator

This project aimed to develop a low-cost, accurate, efficient, and practical localization system using a quadcopter and software defined radio equipment such as USRPs. Such a product can be utilized to help locate missing people in large outdoor areas. By integrating different engineering concepts, i.e. a triangulation algorithm from software engineering, analog and digital signal processing, and wireless communication, the final product is envisioned to be an inexpensive deployment of triangulation techniques to be used in search and rescue operations [13]. The effectiveness of the system is validated using appropriately selected test cases and verification procedures. The design complements the existing localization methods by extending the operating scope into UAV based scenarios.

After some research, the students specified an ethical problem as follows: one of the quadcopters can go awry and decrease the accuracy of the whole system based on statistical data gathered. The students identified five ways to deal with this problem:
- Deny the existence of the problem.
- There is deviation in the system, the customer is informed of it, but no help is offered.
- A warning label says that the deviation should not be used for certain applications.
- Recall notices are sent out, and all deviated copters are replaced.
- Replacement copters are offered only if the customer notices the problem.

Using the scoring system of the theory model, as shown in Table 4, Option 4 received the highest score, and was chosen as the resolution mechanism.

	Theories				
Option #	Utilitarianism	Egoism	Rights	Kantian	Score
1	0	1	0	0	1
2	0	1	0.25	0	1.25
3	0.25	1	0.25	0.25	1.75
4	1	0	1	1	3
5	0.25	0.75	0.25	0.25	1.25

Table 4. Possible solutions score.

Visible Light Communications

Recently, a rapid increase in the number of mobile devices has pushed the radio frequency (RF)-based wireless technologies to their limits. This RF spectrum crunch has motivated the research community to look for solutions and alternative spectrum resources. A promising approach is to use the optical spectrum bands to complement the legacy RF technologies. The concept, known as optical wireless communication (OWC) or free-space-optical communication (FSO), loads the directional optical beams via non-negative modulation techniques and demodulates the light beam on a passive receiver, which is typically a photodetector (PD).

When the visible optical spectrum band is used, OWC/FSO translates into a particular form known as visible light communication (VLC) [15]. VLC is of interest particularly because its transmitters are light emitting diodes (LEDs), which are also the same devices used for solid-state lighting (SSL). The integration of VLC into SSL modules presents a great opportunity for a range of applications beyond just lighting and illumination. In this project the students aimed to build a multi-transceiver VLC access point in the form of a desk lamp using off-the shelf components with a price tag of $200. Along with all the technical, economic, environmental, and social aspects of the project, the students also discussed its ethical challenges.

Throughout their project the students followed the IEEE Code of Ethics, which states "to

	Theories				
Option #	Utilitarianism	Egoism	Rights	Kantian	Score
1	0	1.00	0	0	1.00
2	0	1.00	0.25	0	1.25
3	1.00	0.25	0.50	0.25	2.00
4	1.00	0	0.50	0.25	1.75
5	0.75	0.75	0.50	0.75	2.75

Table 5. Possible solutions score.

seek, accept and offer honest criticism of technical work, to acknowledge and correct errors, and to credit properly the contributions of others." They made changes in the product according to the feedback they received from various sources, such as the surveys they conducted. The project also requires compliance with the IEEE 802.15.7 standard, which provides a global standard for short-range optical wireless communication using the visible light spectrum. This standard provides access to several hundred THz of unlicensed spectrum, immunity to electromagnetic interference and noninterference with RF systems, and additional security by allowing the user to see the communication channel. It also allows for communication augmenting and complementing existing services from visible-light infrastructures. The IEEE 802.15.7 standard adheres to applicable eye safety regulations.

The ethical dilemma selected by the students was the action against a hardware problem. To use the Theory Model for assessing the ethical choices, they first identified the possible actions. The first option is to do nothing and deny that the problem even exists. The second option would be to tell consumers that there is a possibility of existence of a problem. The third option they came up with is to refund money to consumers limited to a set time period after the discovery of the problem. The fourth option is to send a technical support employee to resolve the problem. The last option they came up with is to give consumers some sort of incentive such as a coupon or gift card for future purchases of their products.

If option 1 were chosen the consumers would be enraged that nothing was being done, and as a result they would lose business. Option 2 would forewarn consumers that there may be a problem, so they would know ahead of the time. However, consumers may still become enraged and frustrated once the problem arises. Option 3 would show consumers that the manufacturer cares about them and acknowledges that there was a mistake and will make up for it. At the same time the manufacturer would save money by only refunding for a set time period. Option 4 would cost the manufacturer significant profit by sending technical support to all our consumers, and the consumers would most likely still be angry that they have to wait for help to arrive. Option 5 might make the consumers happier; even though the manufacturer would not fix the problem, they gave consumers discounts on more products. This shows the customers that the manufacturer did sympathize with them, and it saves the company money at the same time.

As shown in Table 5, the students scored the options according to the four ethical theories on the ethical line diagram. They concluded that the fifth option was the best due to the higher score between the alternatives available through all of the different ethical theories proposed. The third option was also a very good option, coming in second place, and it would most likely make the consumers a lot happier if they went with that option.

Concluding Remarks

In this article we discussed how we train our undergraduate students at FIU-ECE to address ethical dilemmas, via senior design projects. We discussed four representative senior design project examples in which students came up with a hypothetical ethical dilemma, identified multiple options to address the dilemma, and used a methodological approach to choose the best option. With emerging 5G wireless technologies, wireless systems will be integrated into trillions of connected Internet of Things (IoT) devices, immersed into our daily lives. We will continue evolving our ethical training framework at FIU-ECE through senior design projects to address ethical issues that may arise as a result of emerging applications and technologies.

Acknowledgment

The authors would like to thank the senior design students at FIU-ECE, who have worked on the individual projects, namely: D. Guerra, V. Ramos, M. Forest, C. Boza, and E. Fiallo (uPark Assistant project); K. Guevara, M. Rodriguez, N. Gallo, and G. Vellasco (GSM Relay Copter project); R. Calderon, L. Perez, L. Gu, and H. Liang (Copter Triangulation Locator project); S. Asratyan, S. Coriolan, R. Gonzalez, R. Luabeya Mafu, and M. Salazar (Visible Light Communications project). This work was supported in part by NSF CNS-1453678, NSF AST-1443999, and NSF CNS-1422062.

References

[1] "Criteria for Accrediting Engineering Programs, 2015–2016|ABET;" available: http://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2015-2016/#objectives, accessed: 02-Jul-2015.

[2] J. R. Goldberg, "Senior Design Capstone Courses and ABET Outcomes," *IEEE Eng. Med. Biol. Mag.*, vol. 25, no. 4, Jul. 2006, pp. 84–86.

[3] V. Wilczynski and A. C. Foley, "Designing a Capstone Design Course to Achieve Student Outcomes," *Proc. ASME Mechanical Engineering Congress and Exposition*, 2014, pp. V005T05A026–V005T05A026.

[4] S. J. Stagg *et al.*, "Incorporating Design into Undergraduate Biomedical Engineering Curriculum," *Proc. ASEE Gulf-Southwest Annual Conf.*, San Antonio, TX, 2015.

[5] C. L. Dym, P. Little, and E. J. Orwin, *Engineering Design: A Project-Based Introduction, 4th ed.*, New York: Wiley, 2014.

[6] G. E. Dieter and L. C. Schmidt, *Engineering Design, 5th ed*. New York: McGraw-Hill, 2013.

[7] H. Jack, *Engineering Design, Planning, and Management*, Amsterdam; Boston: Elsevier/AP (Academic Press is an imprint of Elsevier), 2013.

[8] C. B. Fleddermann, *Engineering Ethics, 4th ed.*, Upper Saddle River: Prentice Hall, 2012.

[9] "Ethical Issues Concerning Robots and Android Humanoids Technology, Links999;" available: http://www.links999.net/robotics/robots/robots_ethical.html; accessed: 09-Jul-2015.

[10] "Choice: Ethical and Legal Rehabilitation Challenges;" available: http://www.worksupport.com/documents/proed9.html.htm, accessed: 09-Jul-2015.

[11] "Killer Robot," available: https://ethics.csc.ncsu.edu/risks/safety/killer_robot/, [Accessed: 10-Jul-2015].

[12] D. Guerra *et al.*, "uPark Assistant," Senior Design Project, Florida International University, Miami, Fl., 2013.

[13] A. Merwaday and I. Guvenc, "UAV Assisted Heterogeneous Networks for Public Safety Communications," *Proc. IEEE Wireless Communications and Networking Conf. (WCNC) Workshops (WCNCW)*, Mar. 2015, New Orleans, LA, pp. 329–34.

[14] K. Guevara *et al.*, "UAV-Based GSM Network for Public Safety Communications," *Proc. IEEE SoutheastCon*, Apr. 2015, Fort Lauderdale, FL, pp. 1–2.

[15] A. Sahin *et al*., "Hybrid 3D Localization for Visible Light Communication Systems," submitted to *IEEE J. Lightwave Tech*., May 2015.

Biographies

Wilmer Arellano (arellano@fiu.edu) is a full time senior instructor in the Department of Electrical and Computer Engineering, Florida International University, Miami, Florida. His research areas of interest include vehicular ad hoc networks (VANETs) and swarm intelligence. He has authored several publications on VANETs and a book chapter on swarm intelligence. His M.S. and B.S. degrees are in electronic engineering from Universidad Simon Bolivar, in Caracas, Venezuela. He is a member of the IEEE and Tau Beta Pi.

Ismail Guvenc (iguvenc@fiu.edu) received his Ph.D. degree in electrical engineering from the University of South Florida in 2006. He was a research engineer at DOCOMO Innovations from 2006 to 2012. Since August 2012 he has been an assistant professor at Florida International University. His recent research interests include heterogeneous wireless networks and 5G wireless systems. He is a recipient of the 2014 Ralph E. Powe Junior Faculty Enhancement Award and 2015 NSF CAREER Award.

Nezih Pala (npala@fiu.edu) is an associate professor in the Electrical and Computer Engineering Department at Florida International University, Miami, FL. He received his Ph.D. degree in electrical engineering from Rensselaer Polytechnic Institute, Troy, NY in 2002. His research interests include the design, fabrication and characterization of nanoscale systems, electronic and optoelectronic devices for biological and chemical sensing, energy harvesting and storage, plasmonics, THz applications, free space optical communication, and visible light communication.

The Effect of a Stand-Alone Ethics Course in Chilean Engineering Students' Attitudes

Ruth I. Murrugarra and William A. Wallace

Ruth I. Murrugarra is with Universidad Adolfo Ibañez.

William A. Wallace is with Rensselaer Polytechnic Institute.

ABSTRACT

Engineering ethics education is taking on increasing importance worldwide, but in Chile the percentage of universities that have a mandatory course concerning ethics is still small. Traditionally, Chilean universities with existing ethics courses teach them using a philosophical or theological perspective, limited to occidental theories, and usually from a Christian point of view. This article studies the impact of a new methodology and technique to teach ethics in Chile: case-based, non-normative, and with a critical-descriptive approach. An empirical study is conducted to assess the relative impact of an ethics class on students individual and inherent moral values and attitudes, and understand the factors that contribute to this impact. Results indicate that even though the importance of religion in Chile is decreasing, it is still a major source of students' ethical principles and moral values. In addition, results suggest that a change in moral values develops when discussions among groups with different points of view occur.

INTRODUCTION

Chile has 180 institutions that provide higher education, 59 of which are colleges or universities. The first 30 universities according to the Chilean ranking, which include the first 10 universities in Industrial Engineering, enroll almost 60 percent of the students seeking higher education. Among those 30 universities, only 36 percent have a mandatory course concerning ethics. "Professional Ethics" is the ethics course most often taught, and it is usually taught from a Christian point of view.

The most frequent topics included in the current Chilean ethics programs are: ethics and society, morals and ethics, ethics principles, occidental ethical theories such as utilitarianism, Kantianism, and virtue ethics (Aristotle), and the code of ethics of an engineer.

A small percentage of Chilean universities have a mandatory course concerning ethics, and the universities that have an existing ethics course in their program usually follow a philosophical or theological perspective from a Christian point of view, not a case-based approach. Hence, it is not surprising that very few ethical cases with a Chilean context have been developed, and that the few courses that use cases rely on American cases.

For this study, a course designed initially for U.S. students and adapted to the Chilean culture was used and taught as an elective course in the Department of Science and Engineering at Universidad Adolfo Ibañez in Chile, with the title "Ethics and Engineering" [1].

The ethics course was designed as a result of a National Science Foundation (NSF) award for "Educational Simulation for Computing and Information Ethics." This course introduces students to a wide range of ethical theories (eastern, western, and contemporary) and addresses a diverse range of ethical challenges facing us in the information age from a non-normative but a critical-descriptive approach [2].

Throughout the course, students have small-group assignments to enhance learning and develop social skills, such as cooperation, and self-confidence [3]. Lectures in the first half of the semester are related to moral values and a variety of ethical theories, along with discussions of ethical cases. For the second half of the course, a learning tool is used to complement the material. This tool is a case-based educational computer simulation, also developed as part of the "Educational Simulation for Computing and Information Ethics" NSF project, for this course specifically and initially for U.S. students. During this part of the semester, each week a new life-like case is discussed. Students grouped in teams use the case-based educational computer simulation to discuss and confront different ethical issues related to the Information Age using the theories learned throughout the course.

In addition, each team needs to model and simulate an ethical issue using agent simulation techniques. This project involves the modelling aspects of a target phenomenon, the utilization of this model to simulate activities for this target phenomenon, and the utilization of ethical principles to assess motivations, processes, or the related consequences of these activities.

0163-6804/15/$25.00 © 2015 IEEE

Sample

Data were collected during three consecutive years (Fall 2012, Fall 2013, and Fall 2014) from senior industrial engineering students enrolled in a 16-week course. In addition, data from two controlled groups (Fall 2013 and Fall 2014) were collected to verify the impact from externalities.

The sample consists of 130 students in total; 94 have taken the course, while 36 belong to the control group who had not taken the ethics class. The sample includes female and male Industrial Engineering students, all of them in their senior year, all of them born in Chile with Spanish as their first language. Additional information such as gender, religion, and Chilean region of origin were collected. A summary of the students' demographic data is presented in Table 1. The control groups were from senior year Industrial Engineering elective courses where students work on independent projects in teams.

It is relevant to mention that Catholicism was the official religion in Chile until 1925 and is still the main religion in the country, although its predominance has being declining. According to a survey performed in 2014 by a well known survey company in Chile (Latinobarómetro), the percentage of Catholics has decreased from 74 percent in 1995 to 57 percent in 2013, while Atheists, agnostics, and persons with no religion rose from 8 percent in 1995 to 25 percent in 2013. In the sample group, among all Christians only one student considered himself as Protestant, while the remaining considered themselves as Catholics.

Data Collected

Students individually completed the Schwartz Value Survey (SVS), which identifies ten basic personal values, in order to build a moral value profile for each student [4]. Each survey was completed two times during the semester: once in the beginning of the semester before any exposure to course material and once at the end of the semester, after all course material had been presented and all group-assignments had been completed.

The 10 basic personal values are organized on four dimensions or underlying motivations (openness to change, self-transcendence, conservation, and self-enhancement). Adjacent personal values are conceptually close to one another, hence their underlying motivations are similar. On the other hand, the further apart the individual values are the more antagonistic are their underlying motivations (Fig. 1).

Methodology

First, a multivariate analysis of variance, or MANOVA, is used to analyze differences in means among groups when having multiple independent variables and multiple dependent variables; the former are nominal and the latter are quantitative variables. The null hypothesis is that all groups have the same mean for each dependent variable. A significant difference is found when the p-value is less than the confidence level of 0.05 [5].

Sample group	Sex		
	Male	Female	N
Ethics 2012	69 %	31%	30
Ethics 2013	71%	29%	29
Ethics 2014	68%	32%	22
Control group 2013	62%	38%	18
Control group 2014	60%	40%	18
	Religion		
	Christian	None/agnostic/ atheist	
Ethics 2012	45%	55%	
Ethics 2013	55%	45%	
Ethics 2014	52%	48%	
Control group 2013	57%	43%	
Control group 2014	55%	45 %	
	Region of origin		
	Capital	Other	
Ethics 2012	95%	5%	
Ethics 2013	98%	2%	
Ethics 2014	90%	10%	
Control group 2013	92%	8%	
Control group 2014	97 %	3%	

Table 1. Summary of students' demographics.

The first comparison is done to the initial Schwartz Value survey data, to detect differences in the value profiles based on the demographic characteristics of each student. The dependent variables are the 10 basic personal values; the independent variables are gender, region of origin, religion, and sample group.

The second comparison uses data from the second SVS, to detect differences in the final value profiles of each student based on their demographic characteristics and the characteristics of the team they conformed. Each team was characterized based on the demographic characteristics of its members, such as gender composition and religious composition.

Then, to verify the MANOVA results, a Kruskal-Wallis test is applied. This test does not

This project involves the modelling aspects of a target phenomenon, the utilization of this model to simulate activities for this target phenomenon, and the utilization of ethical principles to assess motivations, processes, or the related consequences of these activities.

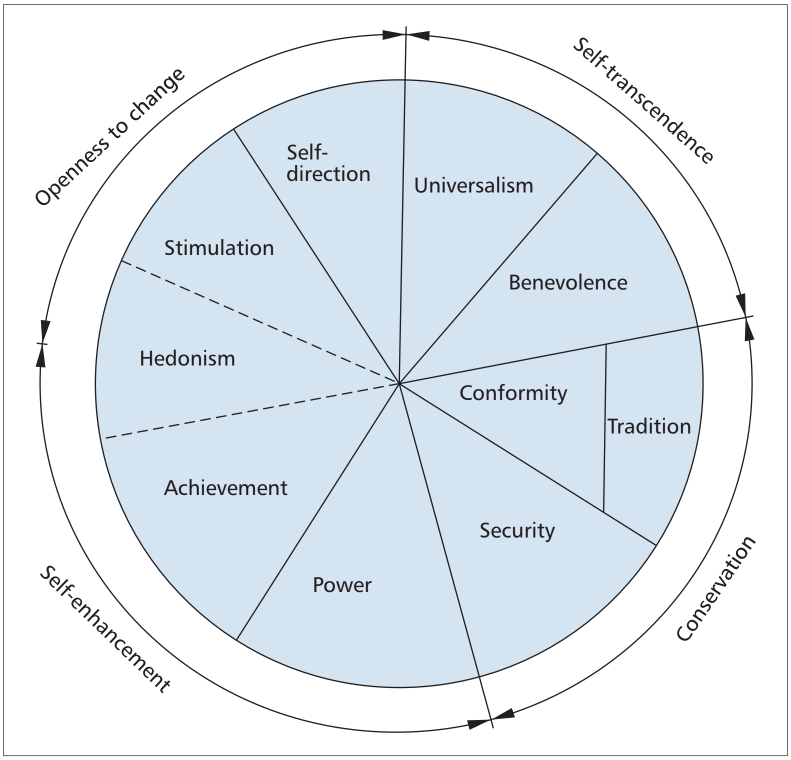


Figure 1. Schwartz Value Survey basic personal values and their underlying motivations.

assume normality, and verifies if data from different groups come from the same population. It is a non-parametric test whose null hypothesis is that all data have the same distribution [5].

Finally, to measure the changes in individual moral values from the beginning to the end of the course, the Wilcoxon signed-ranked test with a confidence level of 0.05 is used. It is a non-parametric test that compares two paired samples. It does not assume that the data comes from a normally distributed population, works with ordinal data, and since it is a rank-based test, it is robust to outliers [5]. The null hypothesis is that the medians from two dependent samples do not differ significantly. To further identify which group is different, a Dunn's post-test is performed. This test compares the difference in the median for each pair of groups [5].

Results

The multivariate analysis performed on data from the first SVS survey (taken at the beginning of the academic semester) indicates that neither gender nor region of origin has an effect on any of the 10 basic personal values. However, there are statistically significant differences based on sample groups.

Almost all sample groups have no significant differences in the mean scores of their basic personal values, except for sample group Ethics 2012, which has significantly higher mean scores on the personal values of Hedonism and Power. These two individual values lie under the same motivational value, Self-Enhancement (Fig. 2).

The MANOVA test also reveals that religion does significantly influence the value of Tradition, where Christians have higher mean scores than the group conformed for Agnostics, Atheists, or students with no religion.

It is interesting to notice that although no significant differences were found in the value Tradition among the different sample groups, the sample group Ethics 2012, which is the only one that has Christians as a minority, has the lowest mean score.

The Kruskal-Wallis test for independent samples confirms that the values of Power and Hedonism are significantly different when comparing all sample groups, with p-values <0.005. In particular, Ethics 2012 reports higher values than all other sample groups on these basic personal values.

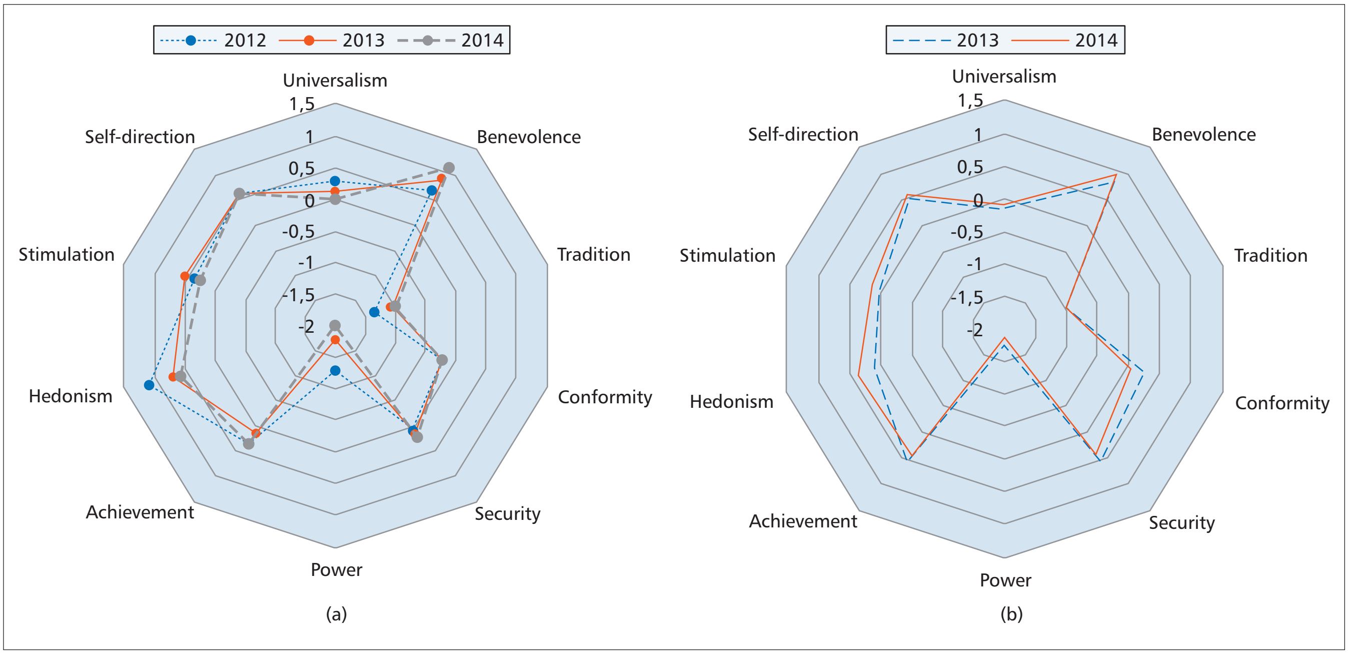


Figure 2. Results of first Schwartz Value Surveys: a) ethics and engineering courses; b) control groups.

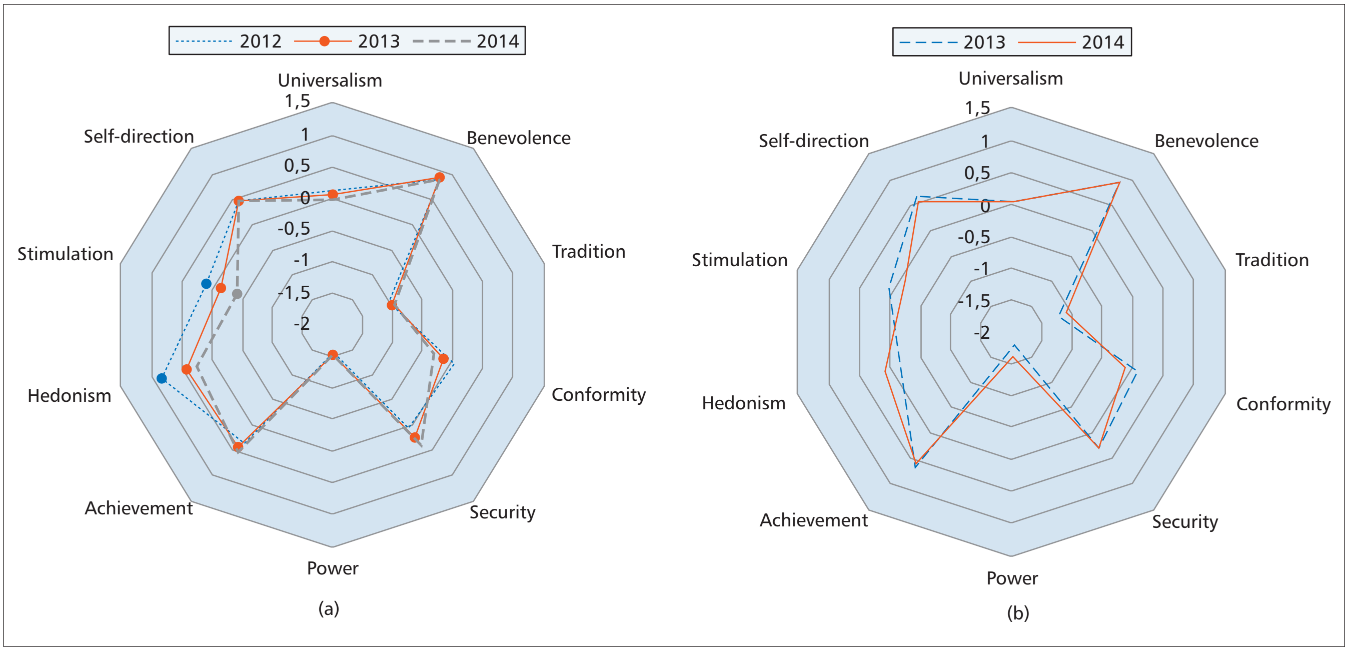


Figure 3. Results of second Schwartz Value Surveys: a) ethics and engineering courses; b) control groups.

The Wilcoxon-Ranked test for dependent samples indicates that no sample group except for Ethics 2012 has significant changes in some of their personal values mean score. In that particular sample group, the values of Power and Stimulation significantly decreased after taking the ethics class. Self-Enhancement and Openness to Change are the underlying motivations of Power and Stimulation, respectively, both having a Personal focus. No changes were significant in the personal values under the motivations of Conservation or Self-Transcendence, which have a social focus. The p-value results are shown in Table 2.

A second MANOVA test applied to the second SVS survey (taken at the end of the academic semester) indicates basically the same as the multivariate ANOVA applied to the first SVS. Neither gender, region of origin, nor the team interaction has an effect on any of the 10 basic personal values, but sample group and religion are factors that affect the mean scores of some of them.

Once again, all sample groups except Ethics 2012 have no significant differences in the mean scores of their basic personal values. However, after taking the ethics class, the sample group Ethics 2012 only had higher scores on the personal values of Hedonism (see Fig. 3). Religion remained influencing the value of Tradition.

It is worth mentioning that although after taking the ethics class the only significant differ-

Personal values	Ethics 2012	Ethics 2013	Ethics 2014	Control group 2013	Control group 2014
Power	0.021*	0.318	0.535	0.498	0.533
Achievement	0.100	0.759	0.642	0.702	0.142
Hedonism	0.535	0.610	0.910	0.444	0.769
Stimulation	0.010 *	0.300	0.326	0.709	0.651
Self-direction	0.691	0.226	0.569	0.434	0.914
Universalism	0.496	0.876	0.501	0.779	0.922
Benevolence	0.201	0.401	0.569	0.230	0.217
Tradition	0.148	0.551	0.776	0.305	0.256
Conformity	0.670	0.931	0.816	0.614	0.624
Security	0.306	0.718	0.326	0.768	0.938

Table 2. P-Value results for Wilcoxon test for dependent samples.

Even though the importance of religion in Chile is being decreasing, it is still a major source of students ethical principles and moral values. Different results on the sample groups can be attributed to the only difference that this study finds among groups, Christian minority/majority.

ence occurs in the sample group Ethics 2012 and in particular in the value of Power, all sample groups are more homogeneous in terms of their mean scores.

Conclusions

This research finds that students who participated in this study, those taking the ethics class as well as those from the control group, have similar initial personal values scores independent of their gender or region of origin. Their value profiles rate higher values related to Openness to Change, such as Self-Direction, Stimulation, and Hedonism, and the value of Benevolence, which is related to Self-Transcendence.

In most of the sample groups the majority of students are Christians except sample group Ethics 2012, which had a majority of Agnostics, Atheists, or those with no religion. Statistically higher initial scores on personal values Power and Hedonism (Self-Enhancement) were found in this particular sample group. In addition, although not statistically different, a lower score on Tradition was found as well.

The statistical analysis also reveals that religion does significantly influence the value of Tradition, where Christians have higher mean scores than the group conformed for Agnostics, Atheists, or students with no religion.

None of the sample groups, including the control groups, with Christians as a majority significantly changed their personal values after taking the ethics class. The only sample group that presented statistically significant changes in some of the personal values' mean scores after taking the ethics class was sample group Ethics 2012. In this particular sample group, the values of Power and Stimulation significantly decreased after taking the ethics class, both related to personal interests or characteristics. And, although not statistically significant, the value of Tradition increased to similar levels as the other sample groups.

Even though the importance of religion in Chile is decreasing, it is still a major source of students' ethical principles and moral values. Different results in the sample groups can be attributed to the only difference that this study finds among groups, Christian minority/majority. Since statistically significant changes occur in the most heterogeneous sample group, in terms of religious beliefs, results suggest that a change in moral values develops when discussions among groups with different points of view occur.

Results are in line with previous research done by the authors [6]. The same experiment was done in a U.S. university and the study concluded that the team interaction with students from different cultures (different points of view) does have an impact on the students' values.

References

[1] R. I. Murrugarra and W. A. Wallace, "A Cross Cultural Comparison of Engineering Ethics Education: Chile and United States," C. Murphy *et al.*, (Eds.), *Engineering Ethics for a Globalized World*, Springer, pp. 189-211, 2015.

[2] K. R. Fleischmann, R. W. Robbins, and W. A. Wallace, "Information Ethics Education in a Multicultural World," *J. Information Systems Education*, vol. 22, no. 3, 2011, pp. 191–202.

[3] K. T. Henson., *Curriculum Planning: Integrating Multiculturalism, Constructivism, and Education Reform, Fifth Edition*, Waveland Press, Long Grove, IL, 2015.

[4] S. H. Schwartz *et al.*, "Refining the Theory of Basic Individual Values," *J. Personality and Social Psychology*, vol. 103, no. 4, 2012, pp. 663–88.

[5] R. C. Martella *et al*., *Understanding and Interpreting Educational Research*, New York, NY., Guilford Press, 2013.

[6] R. I. Murrugarra and W. A. Wallace. "Cross-Cultural and Cross-National Impact of Ethics Education on Engineering Students," *Proc. 2014 IEEE Int'l. Symposium on Ethics in Science, Technology and Engineering*, 23–24 May 2014, pp. 1–6.

Biography

Ruth I. Murrugarra (ruth.murrugarra@uai.cl) is an assistant professor at Universidad Adolfo Ibañez.

William A. Wallace is Yamada Corporation Professor at Rensselaer Polytechnic Institute.

Telecommunication Engineering Education (TEE): Making the Case for a New Multidisciplinary Undergraduate Field of Study

Tarek S. El-Bawab

ABSTRACT

Six years of sustained efforts to recognize telecommunication engineering (TE) as a distinct education discipline came to a successful conclusion on November 1, 2014, with ABET's approval of its 2015-2016 Criteria for Electrical, Computer, Communications, Telecommunication(s) and Similarly Named Engineering Programs. This paper narrates the history of the Telecommunication Engineering Education (TEE) initiative and movement (2008–2014) which resulted in recognition of this field as a distinct engineering education discipline in the US and 28 other countries. We describe how the case for this recognition was made successfully, based on advances in network science and engineering. We discuss some aspects of the new ABET criteria, and how they relate to ongoing and anticipated changes in the arena of engineering education.

INTRODUCTION

Modern telecommunication(s) (Telecom) has changed the way we live our lives, and has become the indispensable foundation of our social and economic activities. Telecom has impacted virtually every aspect of our living: health care, welfare, education, research, social life, transportation, business, commerce, banking, entertainment, tourism, defense, security, and many others. The Telecom industry has evolved to embrace a new mix of traditional and emerging industries and partners, including equipment vendors (from the telephony, data, and TV worlds), service providers (telcos, cable multi-service operators, Internet service providers, utility companies, and even municipalities), and a spectrum of domestic, business, government, and civil/military end users.

Telecommunication is a field of science, engineering, and technology that is concerned with designing, building, operating, and managing the networks and systems used to transport voice, data, image, and video signals. The field also encompasses studies in regulation, legislation, standardization, business, society, and politics. Telecommunication engineering (TE) has undergone paradigm shifts over the last two decades. The technical foundations of the field have changed drastically. Legacy networks optimized for circuit-switched voice evolved to transport packet-based multimedia. Networks are built today based on new principles, theories, and technologies for innovative applications and services. While signs of these trends were looming in the 1990s, and despite early calls for education programs focusing on networks [1, 2], engineering curricular and accreditation criteria lagged behind. Today there are very few specialized telecommunication engineering programs in the US. The nature of our engineering education system, coupled with the history of our Telecom industry, made it difficult for academia to fully appreciate and recognize telecommunication engineering as a distinct field of undergraduate study in the U.S. [3, 4].

The Accreditation Board for Engineering and Technology, Inc. (ABET) is the organization that accredits university programs in applied science, computing, engineering, and engineering technology in the U.S. and in many other countries. ABET is a federation of societies that set policy, develop strategy, and conduct accreditation activities on behalf of their professions. The Institute of Electrical and Electronics Engineers, Inc. (IEEE) is the lead member society in ABET concerning Electrical Engineering (EE) and Computer Engineering (CpE). Before 1982, roughly, there were no ABET accreditation criteria for any engineering discipline. Each constituent society had guidelines applied to its programs. The IEEE had separate guidelines for EE and CpE starting from late 1970s. The concept of a program (discipline based) criteria was adopted by ABET in the early 1980s. As a result, the IEEE developed separate program criteria for EE and CpE. In the 1994-1995 timeframe, the two criteria were merged into combined (ECE) criteria, which remained in use until 2014–2015. In the year 2000 timeframe, ABET

Tarek S. El-Bawab is with the Department of Electrical and Computer Engineering, Jackson State University, USA.

0163-6804/15/$25.00 © 2015 IEEE

also started to place special emphasis on student learning outcomes [5]. The 2014–2015 ECE criteria had no specialized provisions for telecommunication engineering. Telecommunication was one of those areas where ABET accreditation criteria existed for engineering technology programs, but not for engineering programs.

It was quite a surprise, however, to discover that this program, based upon which I was supposed to teach and perform research, was fairly irrelevant as far as modern telecommunication engineering is concerned. My attempts to shed light on such an unfortunate reality were not appreciated.

ABET's process to approve criteria changes is robust and well thought of. Therefore, a strong case must be made for a change to occur. The process involves two main steps. The first step is at the commission level. ABET has four commissions, and the Engineering Accreditation Commission (EAC) is the one responsible for engineering programs. The second step is at the ABET Board of Directors' level. Change-in/new accreditation criteria in a given field of engineering may be proposed by the member society concerned with that field. They are submitted to EAC to be examined. If EAC approves the proposal, with or without comments, it proceeds to the Board of Directors for further approval. A proposal takes this route twice before becoming fully approved and effective. In the first cycle/year, the proposal is examined for approval by the commission and then by the Board. The approval of the Board in this cycle releases the proposal for Public Comments. Upon conclusion of the public comments period, the proposal is sent back to the commission for a second reading and approval, taking public comments into consideration. Then the proposal is passed over to the Board of Directors for final approval.[1]

The Roots and Proceedings of the TEE Movement

Telecom professionals in the U.S. lived in an era of tremendous growth in the 1990s. However, the Telecom industry plunged into a sharp downturn in the first decade of the 21st century. Many professionals and engineers left the industry during that time, including the author of this paper. I was interested in a transition to academia and accepted an offer from one of the very few U.S. universities with a telecommunication engineering program. It was quite a surprise, however, to discover that this program, based upon which I was supposed to teach and perform research, was fairly irrelevant as far as modern telecommunication engineering is concerned. My attempts to shed light on such an unfortunate reality were not appreciated. It turned out that most of the progress and paradigm shifts the telecommunication field experienced throughout recent decades were largely a blur in the eyes of many in our education system.

In 2008 I started to publicly campaign for modern undergraduate telecommunication engineering education programs in the U.S. I wrote an article in this regard, which was ultimately published in the *IEEE Communications Magazine* in January 2010 [3]. The article was well received and captured the attention of many colleagues. The IEEE Communications Society's (ComSoc) Education Board (EB) (now Education and Training Board, ETB) Director at that time was Stefano Bregni. He invited me to form a taskforce (later called a workgroup) to promote Telecommunication Engineering Education (TEE) in the U.S. and to take that cause up with ABET.

The decision to form the TEE Taskforce/Workgroup became effective at the Education Board meeting at Globecom 2010 in Miami (December 7, 2010), but it took time to put the group together. First, a mission had to be defined and outlined properly. I completed the first draft of the TEE position paper in June 2011. This draft was the basis for recruiting the TEE group members, and for my communications with IEEE Educational Activities in 2012. The TEE workgroup became fully composed near the end of March 2012. The final version of the position paper, incorporating remarks from Mehran Esfendiari, George Rouskas, Anura Jayasumana, and Michael Kincaid, was published in March 2013 [4]. TEE evolved from a personal initiative to a ComSoc movement. The affiliation with IEEE ComSoc gave this effort considerable momentum.

IEEE Educational Activities is the entity in IEEE that works with ABET. During the discussion of the TEE effort at the Globecom 2011 Education Board meeting (Houston, December 6, 2011), Tariq Durrani, who was VP, IEEE Educational Activities Board (EAB), discussed the protocol involved in proposing and making ABET criteria changes. Later, he introduced me to the IEEE officers whom I would work with. Several audio meetings were organized, including a conference on April 9, 2012 with Burt Dicht (IEEE Director of University Programs) and Ken Cooper (Chair of the IEEE Committee on Engineering Accreditation Activities (CEAA)) to discuss how accreditation changes can be proposed, and how the work toward that goal may be carried out. On May 27, 2012 Ken informed me that the IEEE had officially asked ABET to change its program name (Electrical and Computer Engineering) to include Communications Engineering. He expected that this request would be approved, a task force would be formed to write a draft of the criteria changes to be made, and that CEAA would ask me to participate in this task.

In July 2012 ABET's EAC voted on a first reading to include Communications in the Electrical and Computer Engineering accreditation criteria. On July 27, 2012 Claire L. McCullough (Chair of CEAA's Subcommittee on Criteria) asked me to join the virtual team she was forming to develop the proposed criteria changes. The team convened on September 7, 2012. It was composed of the CEAA subcommittee on criteria, a representative from ComSoc (the author of this article), and other volunteers. The mission was to recommend communications engineering specific criteria provisions to add to the existing ABET Electrical and Computer Engineering (ECE) criteria, and to produce a draft in this regard to put before the next CEAA meeting in early 2013. Then two members from the TEE workgroup volunteered to participate in these discussions upon my request: Frank Effenberger and Michael Kincaid.

The virtual team discussions took five months, and there were extensive deliberations. By the end of the discussions, two drafts for the new criteria, slightly different from each other, were produced for CEAA to consider. In February 2013 the committee approved a refined version of these drafts,

[1] ABET adopted revisions to its Constitution and By-Laws in October 2012, and reorganized lately. While policy and procedure are under revision, the process of making criteria changes remains the same at the time of this writing.

which was mostly in line with the TEE group recommendations. EAC approved the proposed draft on first reading during the commission's meeting in July 2013. The ABET Board did the same on October 25, 2013, and the proposal of the modified criteria was published for public comments.

The public comments period was open through June 15, 2014. Several colleagues extended their support to the TEE movement and to our cause during that period, including writing to ABET. ComSoc backed up the movement with support and publicity over the society's media outlets.[2] Michele Zorzi, who has been the ETB Director since January 1, 2014, is one of the biggest supporters of the TEE movement. On June 12, 2014 the President of the IEEE Communications Society, Sergio Benedetto, wrote to ABET expressing ComSoc's support of the new criteria, which will pave the way for new education avenues and new research in several emerging areas of engineering knowledge. The proposal of the new criteria was approved by EAC in July 2014. The ABET Board of Directors approved the same on November 1, 2014. The new criteria were published shortly thereafter and are now used for ABET visits worldwide, as of fall 2015.

The new (2015–2016) criteria add specialized provisions for communications/telecommunication(s) engineering, thereby replacing the old *Program Criteria for Electrical, Computer, and Similarly Named Engineering Programs* by the new *Program Criteria for Electrical, Computer, Communications, Telecommunication(s) and Similarly Named Engineering Programs*. These criteria constitute an outstanding milestone for the telecommunication engineering field and profession. The magnitude of this accomplishment may not be obvious to some for a couple of reasons. First, the vast majority of professionals in our field have electrical engineering (EE), computer engineering (CpE), computer science (CS), or other education backgrounds. This has been the case for several generations, and it was sufficient until few years ago. Second, like CpE, telecommunication engineering (TE) is already recognized as an EE spinoff in some countries. Hence, for many in these countries it is hard to be impressed by the norm. However, the fact of the matter is that TE was not as distinctly recognized in the U.S. and in most countries. That status quo was not sustainable in the long term, and this situation has now changed.

This is the first, and largest, change of its type in these particular ECE criteria in at least 35 years. Hence, it is nothing short of history making in terms of engineering education accreditation in the U.S. and in 28 other countries (where ABET accredits programs). It is also an accomplishment for ComSoc as a society and for the global Telecom community. The news was celebrated as such by ComSoc,[2] its Board of Governors, and IEEE [6]. In effect, we developed a definition of what academic telecommunication engineering programs should be about for the first time. We also introduced a fundamental change in the philosophy of criteria writing, since the only difference between EE and CpE in the old criteria was the type of math involved. This has changed when it came to our field, and this can impact future IEEE/ABET criteria.

Matter of Terminology

The new (2015–2016) ABET Program Criteria for Electrical, Computer, Communications, Telecommunication(s), and Similarly Named Engineering Programs make a distinction between communication(s) engineering and telecommunication(s) engineering as follows:

"The curriculum for programs containing the modifier "communication(s)" or "telecommunication(s)" in the title must include topics in communication theory and systems.

The curriculum for programs containing the modifier "telecommunication(s)" must include design and operation of telecommunication networks for services such as voice, data, image, and video transport."

This distinction may pose a terminology issue as the words communication(s) and telecommunication(s) are used interchangeably by many in the field. For the record, the TEE movement did not make this distinction. However, one can see that it makes sense for IEEE, ABET, and existing ECE programs to have this distinction in place, at least at this point in time. The prefix "tele," which designates distance between communicating ends, has been inherent to the words telephony and telegraphy, which defined telecommunication(s) and the industry based thereupon for more than 100 years. Since recent advances in this field have further conquered physical distances in unprecedented ways, it can be argued that this prefix is now even more appropriate to emphasize. Others may see that this prefix is not necessary because communications can now happen at virtually any distance. This is why the two words are interchangeable.

While English language dictionaries provide several meanings of the word *communication(s)*, ranging in their application from human expression and art to science and engineering, these dictionaries are very specific when it comes to the word *telecommunication(s)*, where it restricts the meaning to the field of science, engineering, and technology discussed herein. This contrast is echoed in technical dictionaries, such as the IEEE standard dictionary of electrical and electronics terms [7], where the word telecommunication(s) is described in more detail than communication(s). The new ABET criteria are in line with these dictionary definitions of the two words, since the one with unique relevance

The new ABET criteria are in line with these dictionary definitions of the two words, since the one with unique relevance to the field (i.e. telecommunication) is associated with more specific criteria requirements. The new criteria also make sense in recognizing so many EE programs with emphasis on communication(s).

[2] See, for instance, two ComSoc blogs/releases: Telecommunication Recognized as Distinct Engineering Education Discipline,
www.comsoc.org/blog/telecommunication-recognized-distinct-engineering-education-discipline
May 22, 2014, and Breaking news! Telecommunication Engineering Now Has Official Accreditation Criteria,
www.comsoc.org/blog/breaking-news-telecommunication-engineering-now-official-accreditation-criteria
November 14, 2014.

These goals, which constitute a worthy cause in their own right, have all been accomplished. Much can be done now to capitalize on this achievement, but this is no longer the mission of a single movement of a society or a group of professionals. Further steps are for the global telecom community and industry to take.

to the field (i.e. telecommunication) is associated with more specific criteria requirements. The new criteria also make sense in recognizing so many EE programs with emphasis on communication(s). It is natural that these programs coexist with specialized telecommunication programs. The latter, however, must focus on network science and engineering. Telecommunication engineering curricula cannot be limited to traditional/classical EE curricula and courses in the 21st century.

Finally, some engineering programs have emerged in recent years under the title of network engineering, derivatives of this title, or with a focus on certain aspect(s) of network science and engineering. These programs fall under the telecommunication engineering discipline per the philosophy now adopted by IEEE and ABET. Program/curricular designs may adopt various flavors within the broad multidisciplinary arena of telecommunication(s). The appearance of these new programs, along with other changes happening in engineering education [8, 9], also suggests that the terminologies we use today may further evolve in the future.

New Horizons in Engineering Education

The realization of more programs in new and traditional multidisciplinary areas of studies, such as telecommunication/network engineering, mechatronics engineering, systems engineering, ecological engineering, biomedical engineering, biological engineering, pharmaceutical engineering, nanoengineering, and engineering management, is one of a number of ongoing and anticipated changes in engineering education [8, 9]. Today, engineering is evolving in several directions. In terms of technical scope, it is moving toward cross-disciplinary frameworks as opposed to the classical framework, mostly organized around a handful of primary engineering disciplines. A mix of traditional programs (electrical, computer, civil, mechanical, etc.) and a new generation of multidisciplinary programs can exist in the future, and this would benefit our society and economy. In terms of context, engineering is also evolving to a global scope that embraces human, social, and environmental issues. Tomorrow's engineers will need to acquire a broader perspective of their social and environmental responsibilities. Soft skills are becoming more critical for a new generation of engineers who will practice their professions in the global village of tomorrow.

According to a National Science Board study [8], there are concerns in the U.S. about the American public's perception of engineering education as difficult and as not concerned with humanity and/or society. There are related concerns about declining interest of U.S. citizens in engineering study, about the ability of our schools to retain engineering students, and about the future of the U.S. workforce's capabilities in engineering. Engineering education must evolve to face these challenges. According to this study, the general public perceives engineering as focusing on "things" rather than people. The fact that this is the perception, although our engineering curricula have a considerable component of human/social sciences, suggests a need to revisit general education in our engineering programs. For example, maybe integrating social/human skills into engineering courses instead of enrolling students in courses designed for social/human majors is an approach worth trying, at least in part. This approach can also free some credit hours for new cross-disciplinary engineering coursework. Innovative multidisciplinary programs, such as telecommunication engineering, can also help attract students to engineering study and improve their retention. Some youth who are fascinated with how the Internet works, for example, are potential telecommunication engineering students. Today many of them are probably pursuing electrical and computer engineering where they do not receive sufficient exposure to topics they are most interested in.

What Lies Ahead

The primary goals of the TEE movement were to define modern telecommunication engineering (TE) from an education perspective, explore what an academic TE program must be about, and introduce specialized accreditation provisions for TE programs into ABET's criteria. These goals, which constitute a worthy cause in their own right, have all been accomplished. Much can be done now to capitalize on this achievement, but this is no longer the mission of a single movement of a society or a group of professionals. Further steps are for the global Telecom community and industry to take. This will also depend on how the field and the profession will evolve while facing future challenges.

While the telecommunication/networking literature is rich in scholarly papers, theses, studies, and research books, it is remarkably poor in textbooks that are suitable for telecommunication engineering course and curriculum development. The field may have exploded in content in such a way that is not easily traceable by formal university-type education literature. Today many students/professionals can only keep up with progress through short-lifetime tutorials and white papers, which are not sufficient for rigorous academic training. As an extension of the TEE effort, and capitalization on its outcomes, a new series of textbooks in telecommunication engineering was launched earlier this year [10]. The series aims at exploring new areas in modern telecom and filling a gap thereto in today's literature, for both undergraduate and graduate studies. Now curricular development efforts should also cover graduate programs and modernize them.

Since standardization and regulation are critical in Telecom, the U.S. Department of Commerce's National Institute for Standards and Technology (NIST) has recently decided to fund a project to develop a course in telecommunication standards and standardization at Jackson State University, as part of its 2015 Standards Services Curricula Development (SSCD) program [11]. The course will be offered to senior undergrads and to first-year graduate students.

In recent years I had the honor of receiving communications from colleagues in different countries expressing support and interest in pursuing similar efforts in their regions and communities. My colleagues on the ComSoc ETB have started an effort to explore ways for TEE to make further inroads into additional international territories.

More of these efforts are desirable. Development of telecommunication engineering curricular models is important. Industry should provide support for curricular design so that telecommunication engineering programs are no longer shaped by processes not sufficiently informed about the current state of the art. Industry can also help facilitate laboratory equipment, student internships, and start-up/adjunct faculty. Federal support and leadership, by NSF, NIST, and NTIA,[3] for instance, is important.

Conclusion

Telecommunication engineering (TE) has undergone fundamental paradigm shifts and tremendous progress in recent decades. The field has expanded in theories, enabling technologies, design methods, and applications. The resulting wealth of new engineering and scientific knowledge prompted the Telecommunication Engineering Education (TEE) initiative and movement (2008–2014). Over a period of six years, the movement accomplished its goals to define modern telecommunication engineering from an academic perspective, explore what a TE program must be about, and introduce specialized provisions for TE programs into the accreditation criteria of the Accreditation Board for Engineering and Technology (ABET). We have discussed the roots and proceedings of this movement and explored future directions the global Telecom community may take to further capitalize on its accomplishments. Several colleagues contributed to the success of this movement, which underlines the importance of community effort. Collaborative efforts, involving all stakeholders, are necessary for advancing and modernizing engineering education and for social and economic progress.

Several colleagues contributed to the success of this movement, which underlines the importance of community effort. Collaborative efforts, involving all stakeholders, are necessary for advancing and modernizing engineering education and for social and economic progress.

Acknowledgment

I would like to acknowledge IEEE, CEAA, and ABET for making the criteria change possible. Ken Cooper and David L. Soldan were the CEAA Chairs I worked with throughout this endeavor. David was also the source of information concerning criteria history, and I appreciate his help in this regard. Claire McCullough chaired the virtual team that deliberated and developed the new criteria proposal. Edwin Jones helped the team converge. The TEE movement progressed under three ComSoc administrations. Len Cimini (VP, Technical Activities, 2012–2013) helped us cross administrative boundaries. I must thank Sergio Benedetto (ComSoc President, 2014–2015) for his support. I am grateful for the three EB/ETB Directors who backed my effort generously: Stefano Bregni (2010–2011), Dave Michelson (2012–2013), and Michele Zorzi (2014–2015). I am indebted to Steve Gorshe for putting the TEE position paper under the limelight. John Pape was behind the ComSoc visibility effort in 2014. I am thankful to all the TEE Workgroup Members for their trust in me and support: P. Baniewicz, F. Effenberger, M. Esfandiari, V. Frost, A. Jayasumana, L. Kazovsky, M. Kincaid, M. Medard, and G. Rouskas.

References

[1] Research Priorities in Networking and Communications, workshop report, NSF Division of Networking and Communications Research and Infrastructure, 1992.

[2] W. Yurcik and D. Doss, “The Beginning of a New Discipline: Undergraduate Telecommunications Programs in the USA,” *Proc. ISECON 2001*, v 18 (Cincinnati): §04b.

[3] T. S. El-Bawab, “Is it Time for Specialized Telecommunication Engineering Education in the United States,” *IEEE Global Communications Newsletter*, *IEEE Commun. Mag.*, vol. 48, no. 1, Jan. 2010, pp. 20.

[4] T. S. El-Bawab *et al.*, “Towards Specialized Undergraduate Telecommunication Engineering Education in the U.S.,” *IEEE Commun. Mag.*, vol. 50, no. 9, pp. 14–16, Sept. 2012; and errata in vol. 51, no. 3, Mar. 2013, p. 168.

[5] L. R. Lattuca *et al.*, “The Changing Face of Engineering Education,” *The Bridge* (National Academy of Engineering), vol. 36, no. 2, Summer 2006, pp. 5–13.

[6] K. Pretz, Telecommunications Engineering Is Now a Distinct Education Discipline, *The Institute*, *IEEE,* 21 Nov. 2014.

[7] *IEEE Standard Dictionary of Electrical and Electronics Terms* (IEEE Std 100-1992, 5th Ed., last edition published).

[8] The National Science Board, “Moving forward to Improve Engineering Education,” NSB-07-122, The National Science Foundation (NSF), Nov. 19, 2007.

[9] *The Bridge* (National Academy of Engineering), vol. 43, no. 2, Summer 2013.

[10] T. S. El-Bawab (Series Editor), *Textbooks in Telecommunication Engineering*, Springer, 2015

[11] T. S. El-Bawab (PI), “Advancing and Integration Knowledge of Telecommunication Standards into STEM Education,” project funded by the National Institute for Standards and Technology’s SSCD Program, Federal Award Number 70NANB15H342.

Biographies

Tarek S. El-Bawab [SM ’01] (telbawab@ieee.org) initiated and led the Telecommunication Engineering Education (TEE) movement, which resulted in recognition of telecommunication engineering as a distinct ABET-accreditable education discipline (2008–2014). He is the first (2015) recipient of the IEEE ComSoc Education Award, due to this work. He is a member of the IEEE ComSoc Board of Governors and Director of Conference Operations. He is an active/founding member of several ComSoc technical committees, and has served as Chair of the Transmission, Access, and Optical Systems (TAOS) Technical Committee for two terms. He has served as an organizer and/or Symposium Chair for several ICC and Globecom Conferences, and chaired the ICC/Globecom International Workshop on Optical Networking Technologies (IWONT) for 10 years. He is Board member of the ComSoc Education and Training Board. He is also a member of the IEEE Computer, Electron Devices, and Photonics Societies. His research interests include telecommunication(s), network architectures, optical networks, performance analysis, and enabling electronic/photonic technologies, devices, and systems. Currently, he is a professor with the Department of Electrical and Computer Engineering, Jackson State University, USA. Before this he was with Alcatel-Lucent as a project manager with the Network Strategy Group (CTO organization), USA. Earlier, he was involved in networking research with a number of organizations, including Alcatel-Lucent; the Department of Electrical Engineering, Colorado State University (USA); and the Department of Electronic Systems Engineering, University of Essex (UK). Before this he led large-scale international telecommunication engineering projects for 10 years. He has more than 70 scholarly journal/conference papers and patents. His book *Optical Switching* is one of the most comprehensive references in its subject. He is the Editor of Springer’s new series, Textbooks in Telecommunication Engineering. He has a B.Sc. in electrical engineering, and a B.A. in history, both from Ain Shams University, Cairo, Egypt. He holds an M.Sc. in solid state science from the American University in Cairo, and an M.Sc. in telecommunications and information systems from the University of Essex, UK. He obtained his Ph.D. in electrical engineering from Colorado State University.

[3] National Telecommunications and Information Administration.

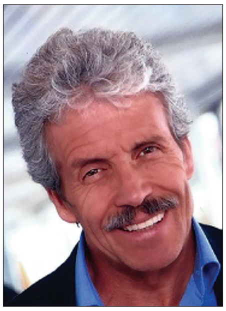
Guest Editorial

Underwater Wireless Communications and Networks: Theory and Application: Part 1

Xi Zhang

Jun-Hong Cui

Santanu Das


Mario Gerla

Mandar Chitre

The Earth is a water planet, two-thirds of which is covered by water. With the rapid developments in technology, underwater communications has become a fast growing field, with broad applications in commercial and military water based systems. The need for underwater wireless communications exists in applications such as remote control in the off-shore oil industry, pollution monitoring in environmental systems, collection of scientific data from ocean-bottom stations, disaster detection and early warning, national security and defense (intrusion detection and underwater surveillance), as well as new resource discovery. Thus, the research of new underwater wireless communication techniques has played the most important role in the exploration of oceans and other aquatic environments. In contrast with terrestrial wireless radio communications, the communication channels in underwater wireless networks can be seriously affected by the marine environment, by noise, and by limited bandwidth and power resources, and by the harsh underwater ambient conditions. Hence, the underwater communication channel often exhibits severe attenuation, multipath effect, frequency dispersion, and constrained bandwidth and power resources, etc., which turn the underwater communication channel into one of the most complex and harsh wireless channels in nature. When facing these unique conditions in diverse underwater applications, many new challenges, which were not encountered in terrestrial wireless communications, are emerging in underwater acoustic, optical, and RF communications for future underwater wireless networks. Of these challenges, acoustic and optical are the most compelling, and somewhat complementary, owing to the potential for longer range and high bandwidth networked communications in size- and power-constrained modems and unmanned systems.

Inspired by the attractive and unique features and potential benefits of advanced underwater communications, the topic of underwater wireless networks has attracted increasing attention from researchers not only in academia, but also in the military and industrial sectors. While a great deal of research efforts have been made in recent years to underwater wireless networks, the aforementioned challenges posed by underwater acoustic as well as optical wireless channel exploitation in future underwater wireless system developments still remain an open problem. As we are launching the first Feature Topic of *IEEE Communications Magazine* focusing on underwater wireless communications and networking, we aim at addressing the urgent needs in both theory and application aspects by industry, military, and the research community in order to better understand the recent progress, explore the future potential research directions, and define new research paradigms in underwater wireless communications and networks. The response to our Call for Papers on this Feature Topic was overwhelming, with a total of 52 articles submitted from all around the world. Going through the rigorous two-round review process, Part 1 of this Feature Topic contains eight excellent articles focusing on the key issues and emerging concepts of contemporaneous underwater wireless networks and techniques.

The first article, "Realizing Underwater Communication through Magnetic Induction," introduces the magnetic induction as a new alternative communication paradigm tackling the high propagation-delay, low data-rates, and highly environment-dependent underwater wireless communications and networks. The second article, "Undersea Laser Communication with Narrow Beams," demonstrates the two main advantages of narrow-beam optical communication: increased power throughput and decreased temporal spread using Monte Carlo analysis under the undersea scattering environment. The third article, "Security and Privacy in Localization for Underwater Sensor Networks," addresses the security and privacy issues in underwater localization by proposing their schemes against the attacks and investigating the techniques for privacy preservation during the localization process. The fourth article, "Software-defined Underwater Acoustic Networks: Towards a High-rate Real-time Reconfigurable Modem," investigates adopting the software-defined radio principles in underwater acoustic networks and proposes and analyzes the software-defined acoustic modem prototype. The fifth article, "Routing Protocols for Underwater Wireless Sensor Networks," overviews the existing routing protocols in under-water-sensor-networks through classifying them into two categories based on a route decision maker, and investigates and compares their various performance issues. The sixth article, "Turbo Equalization for Single-Carrier Underwater Acoustic Communications," investigates the time- and frequency-domain Turbo equalization schemes with low-complexity for MIMO Single-Carrier Modulation systems in underwater acoustic communications infrastructure. The seventh article, "Structured Sparse Methods for Active Ocean Observation Systems with Communication Constraints," presents an ocean monitoring infrastructure with multiple mobile/static acoustic sensors for field reconstruction, target tracking, and exploration-exploitation. The eighth article, "Underwater Sensor Networks: A New Challenge for Opportunistic Routing Protocols," proposes the promising opportunistic-routing protocols for underwater wireless sensor networks by employing dynamic-relay schemes to overcome bandwidth, reliability, and propagation-delay constraints.

We would like to thank all the authors for their excellent contributions and all the reviewers for their valuable reviewing comments. We also appreciate strong supports from Dr. Sean Moore, the former Editor-in-Chief, and Dr. Osman Gebizlioglu, the current Editor-in-Chief of *IEEE Communications Magazine*, and the IEEE Publications team. Finally, we hope that the readers find this Feature Topic interesting and stay tuned for new developments in this research area and Part II of this feature topic in February 2016.

BIOGRAPHIES

XI ZHANG received his Ph.D. degree from The University of Michigan. He is a Full Professor at Texas A&M University, has published over 300 research papers, received U.S. NSF CAREER Award in 2004, is IEEE Distinguished Lecturer, and received four IEEE Best Paper Awards. He is author of an IEEE BEST READINGS journal paper. He has been Editors for numerous IEEE Transactions/Journals, TPC Chair for IEEE GLOBECOM 2011, and TPC Vice-Chair for IEEE INFOCOM 2010.

JUN-HONG CUI received her Ph.D. degree from UCLA in 2003. She is now a Full Professor at the University of Connecticut. Her recent research mainly focuses on underwater sensor networks, autonomous underwater vehicle networks, cyber-aquatic systems, smart ocean technology and ocean computing. Jun-Hong co-founded ACM WUWNet (International Conference on Underwater Networks and Systems), and is now serving as its steering committee chair. She was the recipient of NSF CAREER Award and ONR Young Investigator Award.

SANTANU DAS is the Program Manager of Communications and Networking program within the C4ISR Dept. of Office of Naval Research, where he has broad responsibility for planning, executing and providing leadership for integrated Science & technology projects to develop new capabilities for Naval communication networks. He received a Ph.D. in Electrical Engineering from University of Alberta, Edmonton, Canada and conducted research at AT&T Bell Labs, Whippany, NJ in areas of 3G-wireless and fiber-optic communications.

MARIO GERLA [F'92] received the Ph.D. degree from UCLA. He was part of the team that developed the early ARPANET protocols under the guidance of Prof. Leonard Kleinrock. He joined the UCLA Computer Science Dept. in 1976. He is leading several advanced wireless network projects under Industry and Government funding. His team is developing a Vehicular Testbed for safe navigation, content distribution, urban sensing and intelligent transport.

MANDAR CHITRE received holds a Ph.D. degree in electrical engineering via research in underwater acoustic communications. He currently holds a joint appointment with the Department of Electrical and Computer Engineering at the National University of Singapore as an Assistant Professor and with the Tropical Marine Science Institute as the Head of the Acoustic Research Laboratory. His current research interests include underwater communications, autonomous underwater vehicles, and acoustic signal processing.

Realizing Underwater Communication through Magnetic Induction

Ian F. Akyildiz, Pu Wang, and Zhi Sun

Ian F. Akyildiz is with the Georgia Institute of Technology.

Pu Wang is with Wichita State University.

Zhi Sun is with the State University of New York at Buffalo.

ABSTRACT

The majority of the work on underwater communication has mainly been based on acoustic communication. Acoustic communication faces many known problems, such as high propagation delays, very low data rates, and highly environment-dependent channel behavior. In this article, to address these shortcomings, magnetic induction is introduced as a possible communication paradigm for underwater applications. Accordingly, all research challenges in this regard are explained. Fundamentally different from the conventional underwater communication paradigm, which relies on EM, acoustic, or optical waves, the underwater MI communications rely on the time varying magnetic field to covey information between the transmitting and receiving parties. MI-based underwater communications exhibit several unique and promising features such as negligible signal propagation delay, predictable and constant channel behavior, sufficiently long communication range with high bandwidth, as well as silent and stealth underwater operations. To fully utilize the promising features of underwater MI-based communications, this article introduces the fundamentals of underwater MI communications, including the MI channel models, MI networking protocols design, and MI-based underwater localization.

INTRODUCTION

Underwater communication networks have drawn the attention of the research community in the last decade and a half, driven by a wealth of theoretical and practical challenges. This growing interest can largely be attributed to new applications enabled by large-scale networks of underwater devices (e.g., underwater static sensors, unmanned autonomous vehicles, and autonomous robots), which are capable of harvesting information from the aquatic and marine environment, performing simple processing on the extracted data, and transmitting it to remote locations. Significant results in this area over the last few years have ushered in a surge of civil and military applications. However, the underwater environment imposes great challenges on reliable and real-time communications primarily based on acoustic waves as well as electromagnetic (EM) and optical waves.

EM waves experience high attenuation under water, which severely limits the achievable communication range. To increase the communication range, large antennas are required for low-frequency EM communication, which is not practical for small underwater vehicles and robots. For example, the antenna size of an EM transmitter is a couple of meters for a 50 MHz operating frequency. Optical waves experience multiple scatterings of light, which results in intersymbol interference and short transmission range [1]. To prolong the transmission range, the transmission of optical signals requires high precision in pointing narrow laser beams at the receiver, which is difficult for highly mobile underwater vehicles and robots. The common and most used acoustic waves, while promising long communication ranges under water, exhibit high propagation delay along with unreliable and unpredictable channel behavior and low data rate, which is caused by complex multi-path fading, prevalent Doppler effects, and significant variation of these properties due to temperature, salinity, or pressure [2].

In the last decade our research on magnetic induction (MI)-based communications in challenged and harsh environments [3, 4] such as soil, oil reservoirs, and water pipelines has demonstrated many promising features of MI-based communication. Adopting similar communication principles as in [3, 4], recent research provides mathematical analysis of the MI-based communication channel in underwater applications [5–7]. Compared to commonly used acoustic, optical, and EM communication, MI-based communication has the following advantages:

Negligible signal propagation delay: Different from acoustic waves that propagate at a speed of 1500 m/s under water, MI waves propagate at a speed of 3.33×10^7 m/s under water. This extremely high propagation speed of MI waves can significantly improve the delay performance of underwater communications, while facilitating the design and implementation of the underwater networking protocols, such as medium access control (MAC) and routing, and the underwater networking services (e.g., localization). Moreover, physical layer synchronization among wireless devices becomes simple and reliable due to the negligible delay and stable channel.

Predictable and constant channel response: Since the radiation resistance of a coil is much

0163-6804/15/$25.00 © 2015 IEEE

Communication paradigm	Propagation speed	Data rates	Communication ranges	Channel dependency	Stealth operation
MI	3.33×10^7 m/s	~ Mb/s	10–100 m	Conductivity	Yes
EM	3.33×10^7 m/s	~ Mb/s	≤ 10 m	Conductivity, multipath	Yes
Acoustic	1500 m/s	~ kb/s	~ km	Multipath, Doppler, temperature, pressure, salinity, environmental sound noise	Audible
Optical	3.33×10^7 m/s	~ Mb/s	10–100 m	Light scattering, line of sight communication, ambient light noise	Visible

Table 1. Comparison of underwater MI, EM, acoustic, and optical communications.

smaller than that of an electric dipole, only a very small portion of energy is radiated to the far field by the coil. Hence, compared to acoustic communication, multi-path fading is not an issue for MI-based underwater communication. Moreover, because of the high propagation speed of MI waves, the frequency offsets caused by Doppler effect can be greatly mitigated. Without suffering from multi-path fading and Doppler effect, the MI channel conditions (e.g., data rate and packet loss rate over a given transmission range) are highly constant and predictable. Moreover, without suffering from light scattering as in optical communications, the transmission range and channel quality of MI communications are independent of water quality factors such as water turbidity. In addition, both acoustic and optical communications have to deal with a high level of acoustic and ambient light noises. The EM noise experienced by MI channels is limited under water because the high-frequency noise is absorbed by the water medium.

Sufficiently large communication range with high data rate: In MI-based communications, the transmission and reception are accomplished through the use of a pair of small-size wire coils, that is, coil antennas. Different from the dipole antenna used in most EM wave-based communications, there is no minimum frequency below which the antenna cannot work. On the one hand, the time varying magnetic field can be generated no matter how small the coil is at the MI transmitter. On the other hand, as long as there is magnetic flux going through the coil, the MI receiver can capture the signal even if the frequency is as low as the Megahertz band. This property means that each small coil antenna can be utilized to emit low-frequency MI signals, which allow small underwater robots and vehicles to communicate over sufficiently long distances. Moreover, the operating frequency of MI coils can reach Megahertz bands while maintaining predictable and constant channel quality, which leads to much higher data rates than in the acoustic communications.

Stealth underwater operations: While acoustic and optical communications depend on the generation, propagation, and reception of audible sounds or visible lights, respectively, underwater MI communications utilize non-audible and non-visible MI waves, suitable for a wide range of civilian and military applications that require stealth underwater operations.

The comparison of MI, EM, acoustic, and optical communications is summarized in Table 1. It is also worth noting that the cost of MI coils is very low, generally less than US$1.00 per coil for mass production. Therefore, it is very suitable for large-scale deployment of MI-based underwater nodes. The promising features of underwater MI communications can enable many new and emerging underwater applications:

Collaborative sensing and tracking with underwater swarming robots: Fish behavior shows an astonishing ability to efficiently find food sources and favorable habitat regions through schooling behavior, that is, rapid orienting and synchronized moving of a group of fish with respect to environmental gradients, such as local variations in chemical stimuli such as odorant plumes or other environmental properties such as phytoplankton density. Underwater MI communications can enable a swarm of underwater robots (e.g., agile robotic fish) to mimic this collective and synchronized intelligence of fish by exchanging control and environmental gradient information with guaranteed delay bounds. In such a way, swarming robots can collaboratively track sources of pollution, toxicity, and biohazard with high convergence speed and accuracy.

Stealth and real-time underwater surveillance and patrol: The high bandwidth along with the constant and reliable channel conditions achieved by underwater MI communications can enable real-time underwater surveillance, which demands high-speed delivery of a large volume of multimedia contents (e.g., audio, video, and scalar data). In addition, the stealth and silent features of underwater MI communications allow underwater surveillance to be carried out in stealth mode.

Disaster assessment, search, and rescue in a cluttered underwater environment: Underwater structures, such as leaking submarine cabins, sunken ships, and completely submerged buildings, can create a cluttered underwater environment, which is constituted by confined, hard-to-reach, and complex underwater spaces. Such environments prevent the deployment of bulky underwater vehicles and the application of conventional acoustic communications, which inevitably experience severe multi-path fading with significantly degraded channel quality. In this case, a small and agile underwater robot equipped with small coil antenna can maneuver in such clustered environments much more easily, while allowing reliable and real-time MI

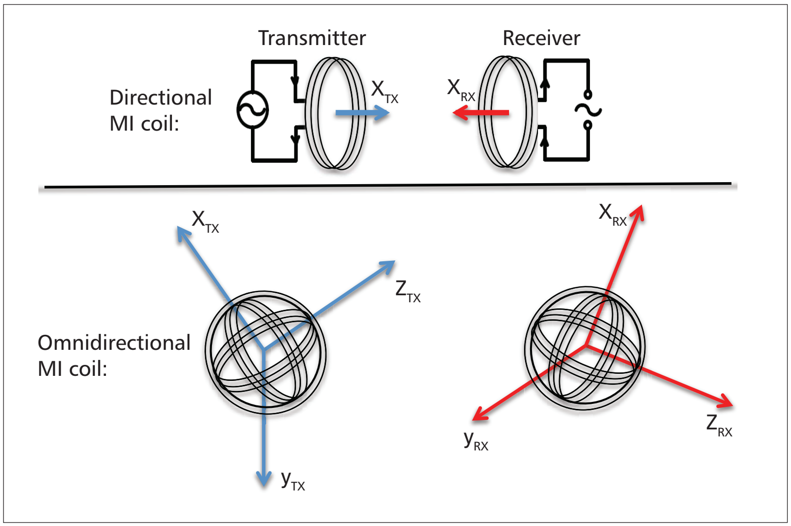


Figure 1. Illustration of MI-based communications with a tridirectional coil antenna.

underwater communications for comprehensive in situ disaster assessment as well as timely survivor search and rescue.

In the remainder of this article, we first focus on the MI channel modeling and physical layer techniques. Then we introduce the design guidelines for underwater MI networking protocols. Finally, we highlight the underlying principles of the MI-based underwater localization method.

Underwater MI Channel Modeling

MI Channel Path Loss

With MI communication, the data information is carried by a time varying magnetic field. Such a magnetic field is generated by the modulated sinusoid current along an MI coil antenna at the transmitter. The receiver retrieves the information by demodulating the induced current along the receiving coil antenna, as shown in Fig. 1. In MI communications, the transmission distance is shorter than a wavelength. As a result, the communication channel experiences less absorption due to the lossy underwater medium. Moreover, the multipath fading is negligible in underwater MI systems. As previously mentioned, the MI transceivers work at the Megahertz band that has a wavelength of tens of meters. Since the communication range in underwater MI systems is within one wavelength, even if there are multiple paths between the transceivers, the phase shifting of multiple paths is so small that the coherence bandwidth is much larger than the system bandwidth. Hence, the fading and channel distortion are negligible.

To rigorously characterize the above unique underwater MI communication channel, an analytical channel model is of great importance. In the MI channel models [3], the EM fields around the transmitter and receiver coils are first expressed using Maxwell's equations. Based on the field analysis, the coupling between the transceiver coils is modeled using the equivalent circuit method. Finally, the MI path loss can be derived as a function of the operating frequency, transmission distance, coil antenna size, number of turns, the relative angle between the two coils, and the underwater environmental conditions, especially the water conductivity.

While the MI channel models in [3, 5, 6] consider the directional coil antennas, Fig. 2 shows the numerical path loss of the underwater MI communication channel with small-size omnidirectional coil antennas [7]. We consider the MI transceivers to be equipped with coil antennas with 10 cm radius and 20 turns of AWG26 wire. The operating frequency is 10 MHz. Three types of underwater environments are considered: sea water with conductivity 4 S/m, lake water with conductivity 0.005 S/m, and drinking water with conductivity 0.0005 S/m. According to Fig. 2a, using a pocket-sized wireless device can achieve 20 m communication range in the drinking water, more than 10 m range in the lake water, but less than 1 m range in the sea water. (Note that here the maximum communication range is considered to be the distance where the path loss reaches 100 dB.) Such big differences in communication ranges are due to the orders of magnitudes differences in the medium conductivities in the sea water, lake water, and drinking water. Highly conductive sea water induces significant Eddy current that incurs very high path loss, which is the problem for both the EM waves and MI techniques in the seawater applications.

The transmission range of underwater MI communications can be increased by using the optimal operating frequency and larger coil antennas. Figures 2b and 2c show the influence of the operating frequency and coil antenna size on the MI channel path loss in the lake water. As the operating frequency increases from 100 kHz to 15 MHz, the MI path loss decreases at first and then keeps increasing after a turning point. The minimum value in Fig. 2b indicates that there is an optimal operating frequency for MI communications in the lake water. On the one hand, the higher frequency can enhance the MI coupling between the transmitter and the receiver. On the other hand, due to the non-zero connectivity in the lake water, the higher frequency also encounters higher loss due to Eddy current. Hence, there is an optimal frequency that achieves the minimum path loss and maximizes the communication range. Moreover, the coil size has a monotonically positive influence on the MI coupling, which means that less path loss and longer communication distance can be realized by increasing the coil antenna size, as shown in Fig. 2c. In addition, the path loss can be further reduced by increasing the wire turns of the coil antenna. For example, based on our channel model, the path loss can be reduced by 6 dB if the number of turns is doubled. It should be noted that the performance shown in Fig. 2 is based on the assumption that the transmitter coil and receiver coil are coaxially positioned. However, the orientation of the coil antenna can be highly random, since underwater robots and vehicles can move and rotate freely in the target underwater environment. Hence, we introduce underwater MI-based omnidirectional communication.

Omnidirectional Communication

Despite the promising properties of MI underwater communications, according to the derived channel models, if the underwater robots and vehicles use the single coil antenna, the path loss performance of the MI channel varies according to the uncontrollable coil orientations. To solve the problem, we propose utilizing a tridirectional coil antenna at the receiver and transmitter, which consists of three coils that are perpendicular to each other, as shown in Fig. 1. Here we show that underwater robots and a vehicle equipped with a tridirectional coil antenna constitute an omnidirectional receiver, and the coil direction has almost no effect on communication performance. It should be noted that, for illustrative purposes, we only let one of the three coils work at the transmitter in Fig. 1. Due to the field distribution pattern of the coil, the orthogonal coils on the same wireless device do not interfere with each other since the magnetic flux generated by one coil becomes zero at the other two orthogonal coils.

To prove the advantages of the tridirectional coil antenna, we numerically compare the performance of the systems using either a unidirectional or tridirectional coil in the MI communication channel. When the receiver is a unidirectional coil, the path losses with different antenna orientations are shown in Fig. 3a. In contrast, when the receiver is a tridirectional coil, the path losses with different antenna orientations are given in Fig. 3b. According to the results, the tridirectional coil antenna has much lower path loss and more reliable performance than the unidirectional coil when the antenna deviates from its optimal orientation. Moreover, since the path loss of the tridirectional coil antenna does not obviously vary when the orientation of the antenna changes, the tridirectional receiver can be regarded as omnidirectional.

Open Research Issues

3D MI channel modeling: Although tri-coil MI communications have demonstrated promising omnidirectional features, an analytical 3D model of the tri-coil MI channel still needs to be developed. Based on such an analytical channel model, the optimal omnidirectional performance can be achieved by properly combining the received signals at the three coil antenna elements by maximizing transmission gain through optimal power allocation, and by adopting spatial-temporal coding mechanisms.

Transmission range extension through underwater MI waveguide: The transmission range of MI communications can be extended by adopting the MI waveguide technique. The channel model of the MI waveguide was developed in [3] delicately for underground wireless communications. The MI waveguide consists of a series of relay coils between two transceivers. An MI relay is just a simple coil without any power sources or processing devices. Since the MI transceivers and relays are coupled, the relays will get the induced currents one by one up to the receiver, so the signal strength at the receiver side becomes large enough over long distance. Since the original channel model of MI-waveguide is developed for 2D directional MI-coil antennas [3], the omnidirectional MI-waveguide channel model still needs to be developed in the 3D aquatic space.

Underwater MI coil antenna design: The parameters of the coil antenna, including the size, number of turns, as well as the lumped and the distributed impedance, have significant influence on the underwater MI communications. Therefore, to maximize the underwater MI communication distance as well as the bandwidth, the optimal coil antenna parameters need to be designed by jointly considering the trade-off between the induced current and the absorption due to skin depth, the trade-off between the MI path loss and the antenna size, as well as the

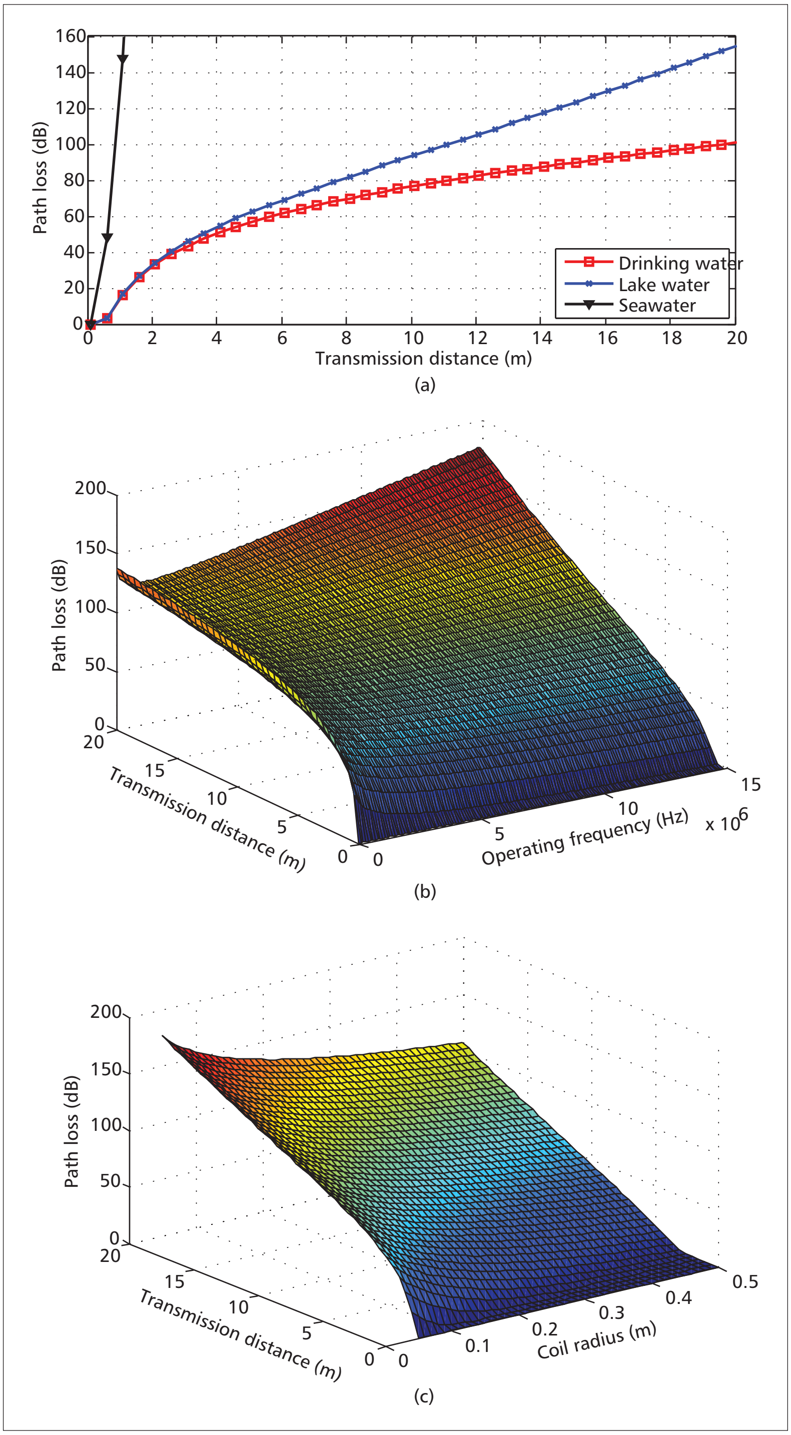


Figure 2. Path loss of the underwater MI communication channel: a) path loss of MI underwater communications; b) influence of operating frequency (lake water); c) influence of coil antenna size (lake water).

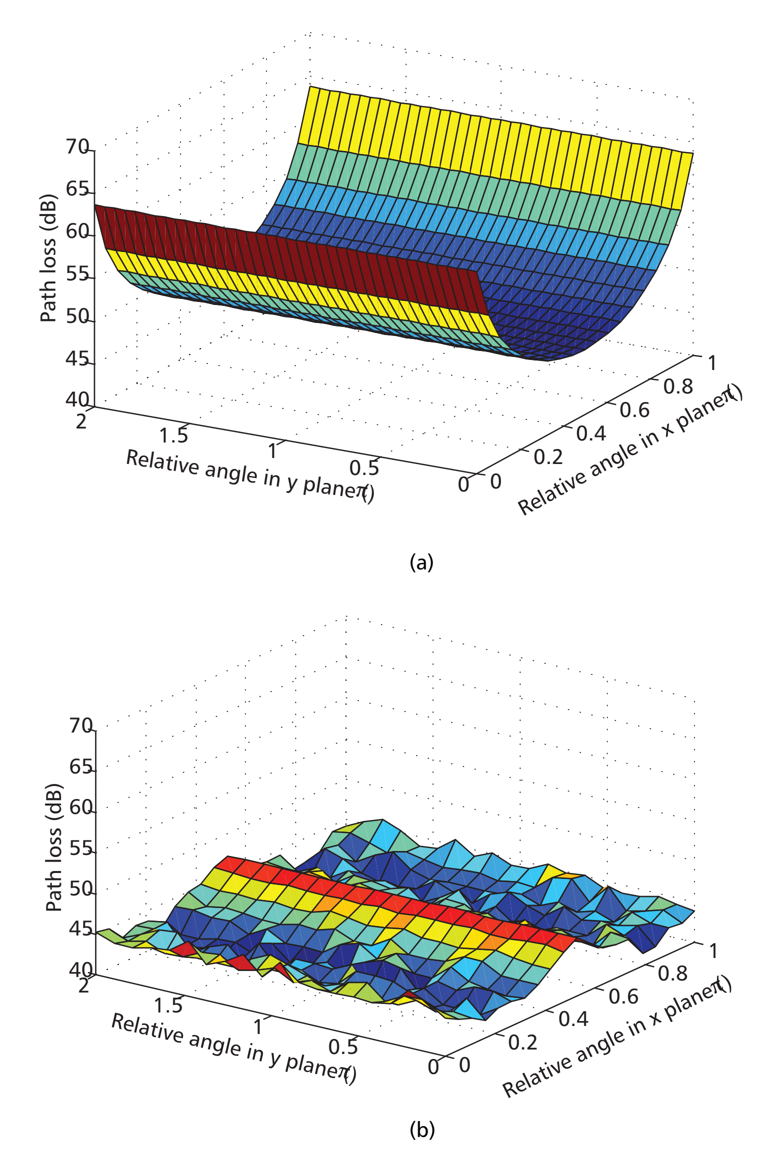


Figure 3. Comparison between the path loss in the MI system with unidirectional and tridirectional coils (as a function of antenna orientations): a) path loss of a unidirectional receiver; and b) path loss of a tridirectional receiver.

trade-off between the quality factor of the coil circuit and the system bandwidth.

Underwater MI transceiver design: To realize the underwater MI communication scheme, the MI transceiver needs to be designed and implemented, which consists of the RF front-end, analog-to-digital/digital-to-analog converter (ADC/DAC), modulator, and equalizer. In particular, since the MI channel is insensitive to multipath fading, Doppler effect, and underwater acoustic noise, an extremely high-order modulation scheme and corresponding channel equalization scheme can be jointly designed based on the underwater MI channel models, which can provide high data rate MI communication links under water.

Energy modeling: Since the underwater robots and vehicles are usually powered by batteries and are expected to operate for a long time period, energy-aware communication, and networking mechanisms are of great importance, which rely on an accurate energy consumption model of the whole system. Therefore, a precise energy model needs to be developed based on the unique underwater MI channel model as well as the optimal antenna and transceiver parameters.

Underwater MI-Based Cross-Layer Design

Overview of MI-Based Cross-Layer Protocol

The performance of underwater MI-communications can be further enhanced through the cross-layer design approach. For example, we have proposed a cross-layer communication framework for the MI-based wireless sensor networks [4]. This cross-layer module provides a unified solution for underground, soil, oil reservoir, and underwater environments because it can adaptively adjust the communication parameters according to the channel characteristics such as bandwidth and path loss under different environments.

More specifically, the proposed cross-layer communication framework is a distributed routing/MAC/PHY solution and allows each node to jointly optimize the next-hop selection, transmitted power, modulation scheme, and forward error correction (FEC) coding rate, with the objective of simultaneously minimizing energy consumption and maximizing packet transmitted rate, while satisfying the application-specific QoS requirements such as delay and packet loss rate. The proposed solution relies on a highly scalable geographical routing paradigm and adopts the direct sequence code-division multiple access (DS-CDMA) as the MAC-layer protocol, the performance of which is further enhanced through distributed game-theoretic power control to combat the near-far problem [4]. Our preliminary results show that the proposed cross-layer protocol outperforms the layered protocol solutions with 50 percent energy savings and 6 dB throughput gain.

Open Research Issues

Underwater MAC protocol design: The existing MAC design in underwater mainly aims to combat the adverse impact of acoustic channels [8, 9]. The underwater MAC protocols via MI-based communications are still unexploited. Carrier sense multiple access (CSMA) schemes are easily implemented in a distributed and low-complexity manner, and are robust to time-varying network topology caused by node mobility. However, conventional underwater acoustic communications face fundamental limitations in implementing CSMA-like schemes because of the high and variable propagation delay of underwater acoustic waves [1]. On the contrary, there are no such barriers to implement CSMA-like schemes for underwater MI communications because the propagation delay MI waves under water are negligible. In addition, the relatively high operating frequency of MI waves paves the way for the design and implementation of frequency-division multiple access (FDMA) protocols under water, which is not feasible for underwater acoustic communication because of its low operating frequency. Similar to MI-based communications, EM-based communications also exhibit low propagation delay and can operate at high frequency. However, the existing MAC protocols designed for EM-based communications in the terrestrial case cannot work effectively and efficiently for underwater MI-based communications because they fail to capture the unique features of the MI channel. For example, because MI-based communications depend on the time-varying near-field magnetic strength, the signal strength of MI-based communications will drop dramatically in the far field. This means that instead of considering the complicated physical model, that is, the signal-to-interference-plus-noise ratio (SINR) model, the MAC protocol design for underwater MI-based ad hoc networks can safely adopt the simple protocol model, where the impact of interference from a transmitting node is only determined by whether or not a receiver resides within the communication range of this transmitting node. Moreover, because of the constant and predicable MI channel, each node can accurately estimate its instantaneous transmission rate only based on its relative location to the receiving party. This can greatly enhance the performance of the celebrated maximum weight scheduling protocols, the throughput optimality of which depends on the channel estimation accuracy.

Routing protocol design: Because of the long propagation delay and high bit error rate of acoustic channels, it is very difficult to provide reliable, timely, and energy-efficient data transfer in large-scale underwater networks over multihop communications. To address the challenges of acoustic channels, sophisticated underwater routing protocols [10, 11] have been developed. However, the promising features of MI channels necessitate revisiting the underwater routing protocol design, taking into account the 3D feature of the underwater environment and the robust underwater MI channel. Moreover, different from the routing operations for EM, acoustic, or optical communications, the nodes in underwater MI-based communications can forward the packets to the next-hop node by acting as either a passive relay working in the MI waveguide mode or a conventional active relay by first receiving the packets and then transmitting them to the next-hop node. A passive relay does not consume any energy, while the forwarding distance is relatively small. An active relay can achieve longer forwarding distance by inducing the transmitting and receiving energy consumption. Therefore, by adaptively selecting different relay modes, the optimal routing protocols can be designed to yield the best trade-off between network lifetime and network latency.

Error control techniques: Since the underwater MI channel exhibits totally different characteristics from conventional acoustic channels, it is necessary to investigate the most suitable error control technique. For example, an MI channel experiences much lower variations, and such a stable channel may only require low-complexity FEC code with high coding rate to achieve high reliability. Our recent research [4] has shown that the low-complexity Bose, Ray-Chaudhuri, Hocquenghem (BCH) code, which is more energy-efficient than convolutional codes, is suitable for MI-based communications in challenging underground environments. Therefore, it is necessary to investigate the optimal FEC coding schemes that yield the best trade-off among the complexity, energy efficiency, and error correction capability of underwater MI-based communications.

Underwater MI-Based Localization

Localization is an important functionality for underwater communication networks. In particular, coordination and maneuvering of underwater robots and vehicles requires each robot to be aware of the relative 3D positions of its immediate neighbors as well as the absolute coordinates in the 3D underwater space. However, the absence of GPS underwater imposes great challenges in the realization of effective relative and absolute localization under water.

The existing underwater localization methods can be mainly categorized into acoustic-based and dead-reckoning ones, both of which have fundamental limitations. Acoustic-based methods, such as long baseline (LBL) and short baseline (SBL) systems, determine the relative or absolute positions of underwater vehicles by measuring the travel time of acoustic waves between vehicles or between the vehicle and anchor nodes, based on the measured or assumed sound speed, and/or the known locations of the anchor nodes. The acoustic-based methods require large acoustic antennas for emitting low-frequency sound, and are thus not suitable for small-scale underwater robots, and suffer unavoidable localization errors caused by the inherent multi-path propagation of acoustic waves especially in cluttered, confined, or shallow underwater environments. On the other hand, dead-reckoning methods estimate the current position of an underwater vehicle based on its last known position along with its speed and heading measured by the onboard sensors, such as gyroscopes, Doppler velocity sonar (DVS), and accelerometers. Utilizing the dead-reckoning methods, the errors in the position estimate grow without bounds as the underwater vehicle travels continuously.

To conquer the above problems, we propose the novel received magnetic field strength (RMFS)-based localization method, which uses the by-product of MI-based communication with tridirectional antenna: the RMFS between the 3 × 3 transmitting and receiving coil antennas. The novelty of the proposed solution relies on the unique multi-path fading free propagation properties of MI-based signals and the inherent orthogonality of tri-coil MI antennas, which guarantee accurate, simple, and convenient relative localization strategy. Moreover, such a method does not induce additional cost, occupy extra space, or increase the weight of the resource constrained underwater nodes. Specifically, each time underwater nodes communicate with each other, the RMFS is measured. Then the mutual distance between underwater nodes can be estimated using the MI channel model and maximum likelihood estimation (MLE) method. Because of the multi-path fading free feature, the RMFS-based localization method

This cross-layer module provides a unified solution for underground, soil, oil reservoir, and underwater environments because it can adaptively adjust the communication parameters according to the channel characteristics such as bandwidth and path loss under different environments.

To fully utilize the promising features of underwater MI-based communications, this article introduces the fundamentals of underwater MI communications, including the MI channel models, MI networking protocol design, and MI-based underwater localization.

will yield much more accurate estimation than the conventional acoustic-based solutions. Relying on real-time RMFS measurement, it does not induce accumulated estimation error as the dead-reckoning methods do. Moreover, since the MI-based technique is sensitive to direction, by using three coils in orthogonal planes, we can estimate the distance in each of the three directions separately, and calculate the distance and angular coordinates of the other underwater nodes relative to a fixed known reference [12]. The open research issues include the development of robust localization under inter-node MI interference for swarming underwater nodes.

Conclusions

In this article, we introduce a new underwater communication paradigm, namely underwater magnetic-induction (MI) communication. Fundamentally different from the conventional underwater communication paradigm, which relies on EM, acoustic, or optical waves, underwater MI communications rely on the time-varying magnetic field to covey information between the transmitting and receiving parties. MI-based underwater communications exhibit several unique and promising features such as negligible signal propagation delay, predictable and constant channel response, sufficiently large communication range with high bandwidth, and silent and covert underwater operations. To fully utilize the promising features of underwater MI-based communications, this article introduces the fundamentals of underwater MI communications, including the MI channel models, MI networking protocol design, and MI-based underwater localization.

Acknowledgment

This work was supported by the US National Science Foundation (NSF) under Grant No. CNS-1446557 and CNS-1446484. We thank Tommaso Melodia, Dario Pompili, Ozgur Akan, Mari Carmen Domingo, and Eylem Ekici for their valuable comments.

References

[1] I. Akyildiz, D. Pompili, and T. Melodia, "Underwater Acoustic Sensor Networks: Research Challenges," Elsevier *Ad Hoc Networks J.*, vol. 3, no. 3, Feb. 2005, pp. 257–79.

[2] L. Lanbo, Z. Shengli, and C. Jun-Hong, "Prospects and Pproblems of Wireless Communication for Underwater Sensor Networks," *Wireless Commun. and Mobile Computing*, vol. 8, no. 8, 2008, pp. 977–94.

[3] Z. Sun and I. F. Akyildiz, "Magnetic Induction Communications for Wireless Underground Sensor Networks," *IEEE Trans. Antenna and Propagation*, vol. 58, no. 7, 2010, pp. 2426–35.

[4] S. C. Lin *et al.*, "Distributed Cross-Layer Protocol Design for Magnetic Induction Communication in Wireless Underground Sensor Networks," IEEE Trans. Wireless Commun., vol. 14, no. 7, 2015, pp. 4006–19.

[5] M. C. Domingo, "Magnetic Induction for Underwater Wireless Communication Networks," IEEE Trans. Antennas and Propagation, vol. 60, no. 6, 2012, pp. 2929–39.

[6] B. Gulbahar and O. B. Akan, "A Communication Theoretical Modeling and Analysis of Underwater Magneto-Inductive Wireless Channels," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, 2012, pp. 3326–34.

[7] H. Guo, Z. Sun, and P. Wang, "Channel Modeling of MI Underwater Communication Using Tri-Directional Coil Antenna," *Proc. IEEE GLOBECOM '15*, San Diego, CA, 2015.

[8] S. Shahabudeen, M. Motani, and M. Chitre, "Analysis of a High Performance MAC Protocol for Underwater Acoustic Networks," *IEEE J. Oceanic Eng.*, vol. 39, no. 1, 2014, pp. 74–89.

[9] Y. B. Zhu *et al.*, "Toward Practical MAC Design for Underwater Acoustic Networks," *IEEE INFOCOM*, 2013, pp. 683–91.

[10] Z. Zhou and J. H. Cui, "Energy Efficient Multi-Path Communication for Time-Critical Applications in Underwater Sensor Networks," *ACM MohiHoc*, 2008.

[11] A. Jain and X. Zhang, "Aqua-Route: Reliable and Energy-Efficient Routing Protocol in Underwater Wireless Sensor Networks," *ACM WUWNET*, 2012.

[12] X. Tian, Z. Sun, and P. Wang, "On Localization for Magnetic Induction-Based Wireless Sensor Networks in Pipeline Environments," *IEEE ICC*, 2015.

Biographies

Ian F. Akyildiz [F'96] (ian@ece.gatech.edu) received his B.S., M.S., and Ph.D. degrees in computer engineering from the University of Erlangen-Nuremberg, Germany, in 1978, 1981 and 1984, respectively. Currently, he is the Ken Byers Chair Professor in Telecommunications with the School of Electrical and Computer Engineering, Georgia Institute of Technology (Georgia Tech), Atlanta, director of the Broadband Wireless Networking (BWN) Laboratory, and chair of the Telecommunication Group at Georgia Tech. Since 2013, he has been a FiDiPro Professor (Finland Distinguished Professor Program supported by the Academy of Science) with the Department of Electronics and Communications Engineering, Tampere University of Technology, Finland, and the founding director of the Nano Communications Center. Since 2008, he has also been an honorary professor with the School of Electrical Engineering at Universitat Politecnica de Catalunya, Barcelona, Spain, and the founding director of the NaNoNetworking Center in Catalunya. Since 2011, he has been a consulting chair professor with the Department of Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. He is Editor-in-Chief of Elsevier's *Computer Networks Journal* and founding Editor-in-Chief of Elsevier's *Ad Hoc Networks Journal*, Elsevier's *Physical Communication Journal*, and Elsevier's . He is an ACM Fellow (1997). He has received numerous awards from IEEE and ACM. His current research interests are in wireless sensor networks in challenged environments such as underwater and underground, nanonetworks, terahertz band, 5G cellular systems, molecular communication, and software defined networking.

Pu Wang [M] (pu.wang@wichita.edu) received his B.E. degree in electrical engineering from Beijing Institute of Technology, China, in 2003, and his M.E. degree in electrical and computer engineering from Memorial University of Newfoundland, Canada, in 2008. He received his Ph.D. degree in electrical and computer engineering from Georgia Tech under the guidance of Prof. Ian F. Akyildiz in August 2013. Currently, he is an assistant professor with the Department of Electrical Engineering and Computer Science at Wichita State University, Kansas. He received the BWN Lab Researcher of the Year Award from Georgia Tech in 2012. He received the TPC top ranked paper award at IEEE DySPAN 2011. He was also named Fa ellow of the School of Graduate Studies, Memorial University of Newfoundland in 2008. His current research interests are wireless sensor networks, cognitive radio networks, cloud radio access networks, software defined networking, cyber-aquatic systems, nanonetworks, and wireless communications in challenged environments.

Zhi Sun [M] (zhisun@buffalo.edu) received his B.S. degree in telecommunication engineering from Beijing University of Posts and Telecommunications and his M.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2004 and 2007, respectively. He received his Ph.D. degree in electrical and computer engineering from Georgia Tech under the guidance of Prof. Ian F. Akyildiz in 2011. Currently, he is an assistant professor in the Electrical Engineering Department at the State University of New York at Buffalo. Prior to that, he was a postdoctoral fellow at Georgia Tech. He won the Best Paper Award at IEEE GLOBECOM 2010. He was given the outstanding graduate award at Tsinghua University in 2007. He received the BWN researcher of the year award at Georgia Tech in 2009. His expertise and research interests lie in wireless communications, wireless sensor networks, and cyber physical systems in challenged environments.

Undersea Laser Communication with Narrow Beams

Andrew S. Fletcher, Scott A. Hamilton, and John D. Moores

The authors are with MIT Lincoln Laboratory.

ABSTRACT

Laser sources enable highly efficient optical communications links due to their ability to be focused into very directive beam profiles. Recent atmospheric and space optical links have demonstrated robust laser communications links at high rate with techniques that are applicable to the undersea environment. These techniques contrast to the broad-angle beams utilized in most reported demonstrations of undersea optical communications, which have employed LED-based transmitters. While the scattering in natural waters will cause the beam to broaden, a narrowly directive transmitter can still significantly increase the optical power delivered to a remote undersea terminal. Using Monte Carlo analysis of the undersea scattering environment, we show the two main advantages of narrow-beam optical communication: increased power throughput and decreased temporal spread. Based on information theoretic arguments, gigabit-per-second class links can be achieved at 20 extinction lengths by utilizing pulse position modulation, single-photon-sensitive receivers, and modern forward error correction techniques.

INTRODUCTION

Undersea wireless communications is a significant challenge due to the highly attenuating nature of seawater for most electromagnetic frequencies. While acoustic communications has been demonstrated over long propagation distances (e.g., 1–10 km), it is limited to sub-megabit-per-second data rates, can suffer severe multi-path, and introduces orders of magnitude greater latency than optical signaling. By contrast, with blue-green optical communications, gigabit-per-second rates can be achieved, over distances potentially up to hundreds of meters in the clearest waters, enabling a host of new applications. Laser light can be collimated into extremely narrow beams, with sub-milliradian-class diffraction-limited divergence angles. Even in seawater, despite significant scattering, narrow transmitted beams yield an advantage by maximizing the power delivered to a remote terminal, provided the two terminals can point to each other with sufficient accuracy. In this article, we explore the potential benefits as well as the challenges for undersea wireless communication with narrow optical beams.

The undersea systems analysis herein is inspired by lessons learned in the last few decades developing high-performance laser communications (lasercom) for atmospheric and space applications. The recent Lunar Laser Communications Demonstration (LLCD) [1] achieved error-free communications from a moon-orbiting satellite to the Earth's surface at rates up to 622 Mb/s. Both the space terminal (10 cm aperture transmitting 0.5 W of optical power) and ground terminal (array of four 40 cm receive apertures) were modest in terms of aperture and optical power. Another highly successful atmospheric lasercom demonstration was the Free-Space Optical Communication Airborne Link (FOCAL), achieving error-free 100 GB file transfers over 25+ km from an aircraft to a ground terminal at a rate of 2500 Mb/s, with robust tracking out to 60 km [2]. Again, the optical power (0.5 W) and aperture sizes (2.5 cm on the airplane, four 1.2 cm apertures on the ground) were very modest. Neither system required the complexity of adaptive optics.

LLCD and FOCAL provide lessons in optical communications applicable to undersea. Both systems used diffraction-limited beams to maximize power delivery. Each terminal tracked light transmitted from the remote terminal through a cooperative means for accomplishing pointing, acquisition, and tracking (PAT). The PAT systems were robust amid platform vibrations and through the turbulent atmosphere. Due to the extremely long range, LLCD approached lasercom from the perspective of fundamental performance bounds. Two vital ingredients were careful channel characterization and the information capacity analysis of the modulator/receiver pairs. In addition, LLCD demonstrated the operational utility of single-photon-sensitive detectors.

In this work, we explore the similarities and differences of the atmospheric and undersea channels, the technologies available to undersea,

0163-6804/15/$25.00 © 2015 IEEE

	Absorption coefficient **a (m^{-1})**	**Scattering coefficient** **b (m^{-1})**	**Extinction length** **$(a + b)^{-1}$ (m)**
Clear ocean	0.114	0.037	6.6
Coastal ocean	0.179	0.219	2.5
Turbid harbor	0.366	1.824	0.46

Table 1. Representative absorption and scattering coefficients at wavelength of 514 nm, taken from Petzold (1972) [3].

and the applicability of nuanced atmospheric PAT techniques. The article is organized as follows. We begin with a description of undersea channel modeling, follow with a discussion of the benefits of narrow-beam optical systems, and conclude with implementation considerations.

Undersea Channel Characterization

Successful undersea lasercom will require a system design informed by robust and accurate characterization of the propagation channel. This will especially be true with narrow-beam communications that seeks to push performance to the frontiers of what is physically realizable. Fortunately, ocean engineers have extensively worked to characterize the propagation of light through various seawater conditions. We rely on these characterization efforts and interpret them in the context of narrow-beam optical communication.

Signal attenuation due to absorption and scattering is by far the dominant loss term in any undersea optical communication link. While the scale of attenuation varies dramatically depending on the water characteristics, all undersea propagation is characterized by a loss exponential with propagation distance. The standard method of describing this loss is in terms of an absorption coefficient (typically given as a) and a scattering coefficient (given as b), both in units of m^{-1}. A beam attenuation length, or extinction length, of $(a + b)^{-1}$ m refers to the propagation distance that results in a power reduced by a factor of $e^{-1} \approx 0.37$ due to absorption and scattering. Alternatively, a scattering length of b^{-1} m refers to a reduction of e^{-1} due to scattering alone. Thus, small values of a and b denote clear water, and allow light to propagate for longer distances. Typically referenced values are listed in Table 1. Higher scattering coefficients correspond to waters with higher concentrations of biomaterial, such as phytoplankton or chromophoric dissolved organic matter (CDOM), or in some cases suspended sediment. We can see a variation in extinction greater than a factor of 10; we also see that even in clear ocean conditions the exponential extinction is significant.

Atmospheric and free-space link losses are typically dominated by a beam spreading term proportional to R^{-2}, where R is the propagation range. For collimated beams of light undersea, the extinction loss $e^{-(a+b)R}$ dominates, and the R^{-2} loss becomes nearly negligible in comparison. Note that this is only true for nearly collimated light; if the light source has a broad initial divergence angle, as with typical LED-based sources, the R^{-2} loss plays a much more important role in the link budget calculations. In direct contrast to atmospheric lasercom, diffraction is almost irrelevant in terms of calculating link losses for undersea lasercom.

The most straightforward approximation of scattering effects, referenced above, is to treat all scattered photons as a link loss term. This "scattering as loss" approximation is highly appropriate for atmospheric lasercom links. However, in seawater, light is strongly forward-scattered, and some non-negligible fraction of the scattered light will in fact be collected by the receiver; a "scattering as loss" interpretation of the undersea channel is especially pessimistic for scenarios such as the turbid harbor, where the scattering is substantially stronger than absorption. Inclusion of scattered light is vital to a correct comparison of wide-beam vs. narrow-beam optical systems.

Scattering also has a significant temporal effect on the optical waveform. Photons that scatter one or more times but still reach the receive terminal have traveled a longer distance than "ballistic" photons that propagate without being scattered, resulting in a temporal spreading of the received waveform. Temporal spreading is a function of several system parameters, including the scattering coefficient, the propagation distance, the transmit beam size, the receive aperture size, and the receiver field of view (FOV). Intuition regarding the receiver FOV is particularly important for our narrow beam analysis: a narrow FOV receiver will limit the received photons to scattered light with very small scattering angles, while a wide FOV will accept light that scatters significantly off axis. The former will have propagation distances very close to the ballistic photons, thus minimizing the temporal spread. Wide angle photons will have traveled longer distances, and their inclusion makes the temporal spread more severe.

Modeling the channel response due to scattering is best performed by means of ray-tracing Monte Carlo simulation to compute the random paths of an ensemble of photons. Each photon is subject to a series of independent scattering events, with the frequency of such events characterized by the scattering coefficient b, and the resulting angle randomly drawn from a scattering phase function with a strong forward scattering emphasis. Given an initial distribution of photons constituting the lasercom beam exiting the transmit aperture, we compute the independent random walk for each photon and approximate the distribution of photon density, arrival angle, and time of arrival at the receive aperture, using the invertible analytic volume scattering function (VSF) in [4, Appendix B]. An example photon density distribution is given in Fig. 1. Despite the approximation's dismissal of coherence effects, such Monte Carlo ray tracing simulations are widely valued computational methods for modeling the undersea channel.

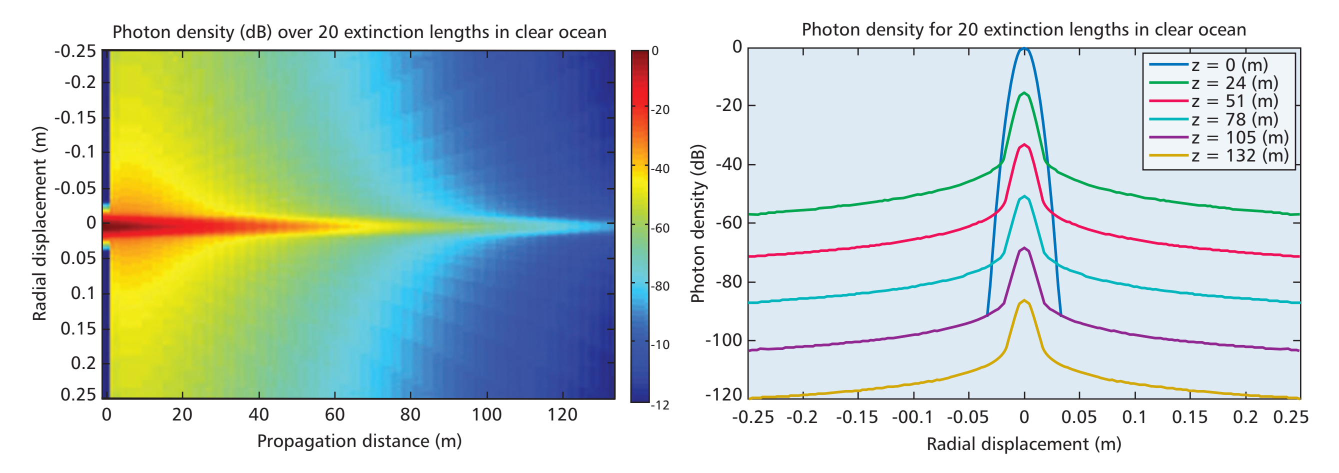


Figure 1. Simulated beam profile in clear ocean conditions. The transmit beam is Gaussian with a 1 cm radius beam-waist. The profile is given over distances up to 20 extinction lengths.

Benefits of Narrow-Beam Lasercom

Narrow-beam undersea lasercom has three significant advantages over wide beam optical communication: an increase in the light transmitted across the channel, a reduction of the temporal spread of the signal, and enhanced spatial and spectral filtering options to reduce background light. To illustrate these impacts on communication performance, we begin with a discussion of modulation and information capacity in photon-starved channels. We follow with example scenarios in clear and turbid waters.

Modulation and Information Capacity for the Photon-starved Channel

The recent optical communication demonstration from the moon's orbit to an Earth ground station (the aforementioned LLCD) demonstrated high-rate optical communication in what is sometimes referred to as the "photon-starved channel." Such a classification is given to a system where the total signal flux relative to the data rate is very limited. Deep space links (due to their astronomical distances) and undersea links (due to sea water's exponential extinction) can both exhibit photon-starved channels. While large apertures and high optical powers can partially compensate, practical size and power limitations encourage maximizing photon efficiency. In the case of LLCD, photon efficiency was increased by utilizing optical bandwidth (a plentiful resource) and detectors sensitive to single photons.

For the high-rate LLCD downlink, multiple bits of information were communicated for every received photon. LLCD used a high-bandwidth (5 GHz slot rate) signaling scheme utilizing 16-ary pulse position modulation (PPM) and half-rate forward error correction (FEC). For each 16-slot symbol, exactly one contained an optical pulse; the temporal location of the pulse-containing slot indicated which of 16 symbols was transmitted. The receiver deployed an array of single-photon detectors with precise time of arrival resolution. By detecting the arrival of even a single photon per symbol, multiple bits of information were transmitted.

PPM signaling and single-photon receivers are directly applicable to the undersea environment. Information theory allows us to compute the best achievable efficiency with PPM and single-photon-sensitive receivers. Modeling the photon arrivals as Poisson distributed (characteristic of laser light), we calculate the channel capacity for PPM signaling in the ideal case of no background light. (For a detailed derivation of channel capacity for optical receivers, see [5].) Figure 2a plots the achievable sensitivity vs. the bandwidth expansion. A 16-ary PPM system with 1/2-rate FEC has a bandwidth expansion of 8 and can achieve –4.6 dB photons/bit, or 2.9 b/photon. Figure 2b shows the capacity of 16-ary PPM when N_b photons of background light, detector dark counts, and temporally spread signal photons are included. In a low-noise case such as a deep-water or night scenario, the sensitivity is close to the noiseless result. Higher background levels (e.g., from upward-looking, near-surface, daytime scenarios) impact the achievable sensitivity. Spatially and spectrally narrow filters increase the information capacity by reducing stray light to the detector.

Clear Ocean Scenario: Increase Photon Delivery

Consider a clear ocean scenario with absorption and scattering coefficients of $a = 0.114\ \text{m}^{-1}$ and $b = 0.037\ \text{m}^{-1}$, yielding an extinction length of 6.6 m. A low-size low-power narrow-beam lasercom system can close the link over 20 extinction lengths (132 m). Consider transmit and receive terminals with a 2 cm diameter and a collimated transmit beam at a wavelength of 515 nm. Even with such small terminals, 132 m still represents near-field transmission, so in vacuum an aligned system would couple all of the light to the receiver. (We assume that the terminals are properly pointed; discussion of methods for pointing and tracking follows in a later section.) The channel loss due to absorption and scattering is 87 dB, where we assume all scattered light is lost. For a

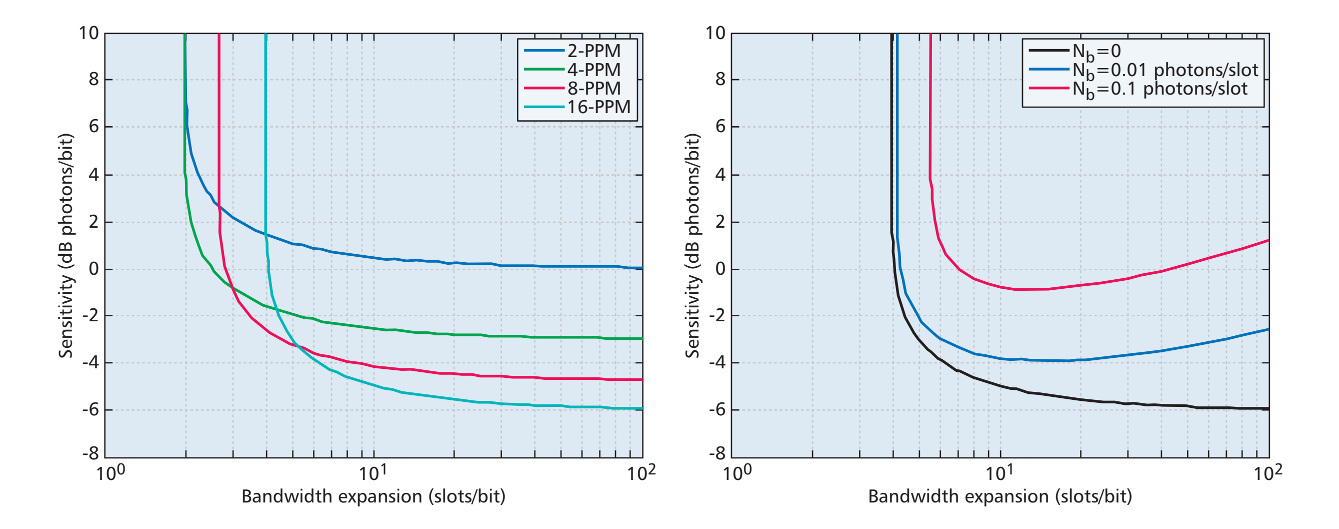


Figure 2. PPM sensitivity at capacity with an ideal single-photon detecting receiver. In a), PPM is compared for orders 2, 4, 8, and 16 with no noise photons included. In b), 16-PPM sensitivity at capacity for noise levels of N_b = 0.01 and 0.1 average photons per slot are compared to the noiseless case.

1 Gb/s link using 16-ary PPM and 1/2-rate FEC, we saw above that the noiseless sensitivity at capacity is 2.9 b/photon. Such a system requires –109 dBW of received power. Even allowing 12 dB for imperfections (background light penalty, optics losses, FEC losses, pointing losses, non-unity detection efficiency, etc.), the narrow-beam system could close the 1 Gb/s link with 100 mW of transmit power.

For comparison purposes, we consider a wide-beam optical communication system (with the same aperture sizes) in which the transmitter illuminates an entire hemisphere (2π steradians), and the receiver similarly has a hemispherical field of view. Coupling into a 2 cm aperture incurs substantial loss, partially mitigated by the fact that scattered light also couples to the aperture. 132 m of propagation leads to a loss of 153 dB. By implementing a narrow-beam system, we can increase the coupled light by over six orders of magnitude. With the same 100 mW transmit power, the maximum data rate of a wide beam system at that distance is 3.5 kb/s.

Turbid Harbor Scenario: Reduce Temporal Spread

In the second case study, we consider a turbid harbor scenario with absorption and scattering coefficients of a = 0.366 m^{-1} and b = 1.824 m^{-1}. We assume the same 2 cm terminals. In these waters, 20 extinction lengths are a distance of only 9.1 m. In contrast to the clear ocean, scattering contributes significantly more to extinction than absorption. Both the narrow-beam and wide-beam cases must account for the arrival of scattered photons. In Monte Carlo simulation, the narrow-beam system observes a loss of 67 dB, while the wide-beam system sees 80 dB of loss. The narrow-beam advantage in throughput, while present, is not nearly as dramatic as for the clear ocean case.

The advantage of the narrow-beam system in the harbor case is more evident in the temporal spread of the photon arrival. As seen in Fig. 3, the wide-beam system has a significant spread in the time of arrival of the received photons. This can be understood intuitively by noting that most arriving photons have scattered multiple times, causing them to follow random trajectories. More than 10 percent of the received photons arrive 10 ns after the earliest arrivals; it requires 200 ns before 99 percent of the photons arrive. The temporal spread causes inter-symbol interference (ISI) unless the signaling rate is reduced. (For PPM, ISI has an effect equivalent to background photons.) In contrast, the delay spread is significantly reduced in the narrow-beam case. Here only 1 percent of the photons arrive 10 ns after the earliest photons. Furthermore, because this is a tracked system, the receiver can spatially filter the received photons by their angle of arrival. Even a modest filter, such as 0.1 rad, dramatically reduces the delay spread so that 99 percent of the photons arrive within 1 ns, as shown in the figure. Of course, such a filter reduces the number of arriving photons as well, by 4 dB in this example. Narrow spatial filtering enables high throughput signaling in the presence of severe scattering.

Narrow-Beam Lasercom Implementation Considerations

Communication System Dynamic Range

Tracked narrow-beam optical systems are intended to achieve high data throughput at long distances. To accomplish their objective, the terminals would be designed for some specified low power level representing a desired number of extinction lengths. Water clarity, however, is variable over time. In addition, a robust system should be able to handle distances shorter than the maximum range. By either shortening the propagation distance by a few extinction lengths or clarification of water, the communication link could find itself with orders of magnitude more optical power available. A robust system design would gracefully adjust to such power variations and be effective in a wide variety of undersea environments.

In the scenarios above, we cited 20 extinction length scenarios with throughputs ranging from –87 dB up to –67 dB for narrow-beam systems in clear ocean and turbid harbor seawater, respectively. We have asserted the possibility for designing gigabit-per-second class communication systems under these circumstances. Such a receive terminal would certainly require sensitive optical detectors, which typically have dynamic ranges between 10 and 20 dB. A range reduction of 2–5 extinction lengths could move such a receiver into saturation, all other factors being equal.

The simplest mitigation method reduces the power coupled to the receiver. This could be done by transmit power reduction, transmit beam widening, or a receiver optical iris or other variable receiver attenuator. On the other hand, increased received signal strength is typically a boon to a communication system. The single-photon-sensitive receiver described above is applicable to the longest propagation range, but a higher-bandwidth receiver uses the increased signal strength to increase the data rate. A notional undersea lasercom terminal would have dual electro-optic front-ends for the low signal and high signal cases.

Pointing, Acquisition, and Tracking

Initializing and maintaining a link between two lasercom terminals is often the most challenging aspect of a lasercom system due to their highly directive beams. Pointing, acquisition, and tracking are the three basic functions for a narrow-beam lasercom link. To begin, a transmitter terminal must know its position, its attitude, and the location of the receive terminal with sufficient accuracy to place light on the receiver. For terrestrial systems, location and attitude may be derived from GPS and inertial navigation systems. The transmitter typically sends a broader beam than will eventually be utilized and may scan across the predicted receiver location (known as the uncertainty region). Acquisition by the receive terminal is achieved when it detects light from the transmit terminal. Detectors used for acquisition typically cover a relatively wide FOV (milliradian-class). Tracking requires the receive terminal to adjust its own pointing solution such that it is aligned on-axis with the incoming beam from the transmit terminal. Lasercom systems typically employ a high-bandwidth control loop to null out disturbances on the receive platform and maintain the receive terminal on-axis pointing. In most lasercom systems, acquisition and tracking are employed at both the transmit and receive terminals in a bidirectional manner.

The PAT methodology described here can also be used to enable narrow-beam undersea lasercom systems. A primary driver on PAT architecture for undersea systems is likely to be imprecise location information for the transmit and receive terminals. GPS connectivity will be intermittent (or unavailable), and navigation systems lose accuracy over time without a reference. The net effect will be that the uncertainty region the transmit terminal is required to scan may be significantly larger than for typical atmospheric lasercom systems. Fortunately, the relatively slow (compared to aircraft or spacecraft) undersea platform velocities provide opportunities to accommodate increased acquisition times. An effective undersea lasercom PAT system design will trade transmit (or beacon) beamwidth, acquisition standoff range, and required acquisition time against anticipated open-loop pointing uncertainty to enable a practical PAT architecture for undersea lasercom systems.

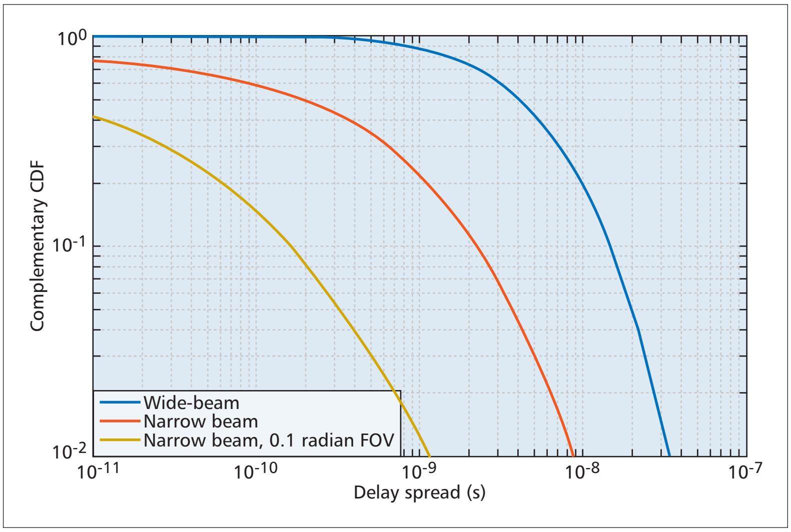


Figure 3. Temporal spread of photon arrivals after 20 extinction length propagation in the turbid harbor case (from Monte Carlo simulation). The vertical axis is the complementary cumulative distribution function; thus, 90 percent of photons arrive within 10 ns in the wide-beam case, 2 ns in the narrow-beam case, and 90 ps in the narrow-beam case with a 0.1 mrad FOV.

Narrow-beam transmission under water may be subject to brief, intermittent signal outages. These may occur due to tracking glitches (although such can be engineered to be rare), or marine organisms or detritus blocking the beam. A robust system should accommodate such outages by robust reacquisition and disruption-tolerant communication protocols (e.g., packet acknowledgment and retransmission).

Transmitter Technology

Much of the past work in undersea-undersea optical communications has exploited LED single device and multi-element array-based transmitters. These are commercially available in the low-attenuation undersea wavelength bands, are relatively low cost, and are capable of multi-Watt output. It is, however, difficult to drive LEDs with high extinction ratios at rates above ~10 MHz, and the low spatial coherence of LED transmitters make them hard to focus into narrow beams. Lasers, in contrast to LEDs, are highly coherent, can have very small divergence angles, and can have narrow spectral linewidths.

Many commercially available semiconductor lasers in blue and green generate ~100 mW power levels and can be directly modulated at ~100 MHz rates. Some development may be required in blue-green for faster signaling, by either direct laser modulation or external modulators. If additional power is required, a master-

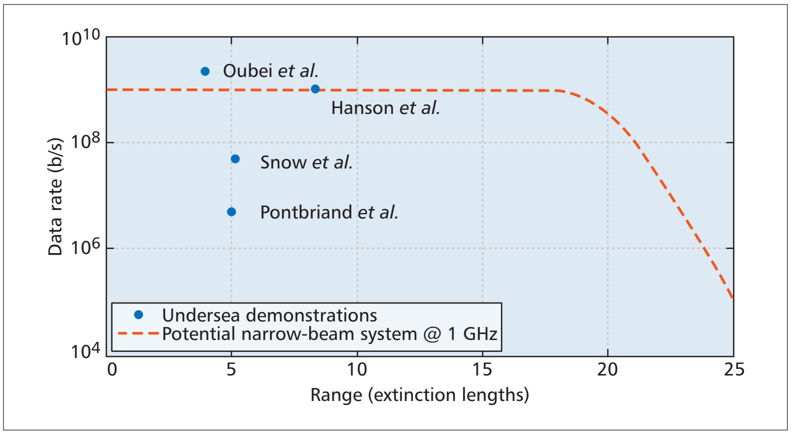


Figure 4. Fundamental channel capacity analysis for an undersea lasercom system employing a narrow-beam transmitter and a photon-counting receiver predicts significant achievable performance gains compared to prior demonstrations. The assumed system operates at 1 GHz, with 2 cm transmit and receive apertures and 100 mW of transmit power. Range is plotted in beam attenuation lengths (extinction lengths).

oscillator power amplifier (MOPA) architecture can be employed. While optical amplification is straightforward in the telecommunications wavelength bands and near 1 μm, there are few amplification options in blue-green. However, amplification in combination with frequency conversion is feasible. For example, a 1060 nm Yb laser can be followed by an external modulator, an optical amplifier, and finally a second-harmonic generation (SHG) crystal.

Optical amplification provides a valuable benefit for communications links with widely varying dynamic range. The output of optical amplifiers is average power limited. As the pulse repetition frequency is lowered, the energy per pulse increases, improving the receiver signal-to-noise ratio (SNR). Thus, the transmitter can be dynamically adapted to the varying loss of a particular link.

Receiver Filter Technology

Optical filtering enables the receiver to spectrally distinguish the signal of interest from out-of-band ambient light, a critical function for undersea receivers in shallow waters and daytime operations. An advantage of using narrow beams for communication between undersea terminals is that commercial off-the-shelf (COTS) interference filters can be used effectively, despite their incidence-angle-dependent passbands. This enables the designer to avoid the use of the specialized filters (Lyot, atomic line filter) required for wide FOV narrowband filtering.

Detection Technology

High-sensitivity photon-counting detectors are highly desired in undersea optical communications systems. These detectors include photomultiplier tubes (PMTs), microchannel plate (MCP), detectors, and avalanche photodiodes (APDs).

PMTs have been the mainstay of undersea optical communications systems. They can be operated at reasonably high rates, enabling tens to hundreds of megabits per second data rates. PMTs also tend to have large active areas, facilitating the collection of received light. They can be very sensitive, even to the point of counting single photons. They have quantum efficiency in blue-green of up to a few tens percentage. PMTs can have some disadvantages as well, including noisiness, limited dynamic range, and limited lifetime, particularly when operated at high count rates.

MCPs also utilize electron multiplication, with a large number of parallel photocathodes and channels (tubes or slots), each ~10 μm in diameter. MCPs can offer very good response times, and, being arrayed, can provide spatial resolution, unlike a single-pixel PMT. A disadvantage of most MCPs is that they cannot thermally tolerate sustained current and therefore only support low count rates. As with PMTs, MCPs can be noisy and have lifetime limitations.

APDs are solid-state semiconductor devices that do not suffer the secondary electron emission damage common to PMTs and MCPs. APDs can be operated in linear mode, typically for high-speed applications, or in Geiger mode (with extremely high gain) for single-photon sensitivity [6]. Silicon APDs can be designed for efficient operation in blue-green and also offer relatively low multiplication noise. They can operate at rates of tens to hundreds of megahertz in Geiger mode, and gigahertz rates in linear mode. APDs typically have much smaller area than PMTs, but can be arrayed, enabling spatial resolution. With proper readout circuit design, the arraying of Geiger mode APDs can be exploited to improve temporal performance as well [6]. Each APD pixel has a finite reset time, but if the readout circuit is designed such that each pixel can be re-armed after firing (rather than periodically re-arming the entire array), the effective bandwidth of the array will scale with the number of pixels in the array. We have architected asynchronous readout circuits for this purpose, and with sufficient investment, the architecture should be scalable to gigacount-per-second rates [7].

Conclusion

To illustrate the potential capacity gains achievable using narrow beams, in Fig. 4 we plot predicted achievable capacity (in red) vs. performance reported by published demonstrations ([8–11]). The predicted performance invokes PAT, sensitive detectors, photon-efficient modulation, and FEC as described in the article. The terminal assumptions are modest, with 100 mW of transmit power and 2 cm apertures as discussed in the scenario examples. Theoretically, ~60 dB of sensitivity increase is possible at gigbit-per-second data rates (in the analysis, we arbitrarily imposed a 1 GHz signaling rate). Implementing a near-capacity system that approaches this previously unrealized undersea communications performance gain should be feasible using the systematic design methodologies described herein. Of course, cost-effective realization for widespread deployment would require further maturation and productization of some devices and subsystems.

Many previous undersea lasercom systems have been designed to operate with over-powered wide-beam optical power transmitters that fall short of optimal performance. We have outlined a path to achieving significant performance improvement for undersea lasercom links using narrow-beam actively pointed transmitters and photon-counting receivers designed to operate with high-sensitivity waveforms and powerful error correction coding. This class of narrow-beam undersea lasercom systems can also be extended to accommodate received photon flux that varies over an extremely large dynamic range for different sea water types or link distances. We believe that undersea lasercom systems designed using the methodology described in this article will enable longer-range higher-rate links using practical optical transmit powers for compatibility with a broad range of undersea platforms and application requirements.

We believe that undersea lasercom systems designed using the methodology described in this article will enable longer-range, higher-rate links using practical optical transmit powers for compatibility with a broad range of undersea platforms and application requirements.

Acknowledgment

This work is sponsored by the Office of Naval Research under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the author, and are not necessarily endorsed by the United States Government.

References

[1] D. M. Boroson *et al.*, "Overview and Results of the Lunar Laser Communication Demonstration," *Proc. SPIE*, Vols. 8971, 2014, pp. 89710S-1–11.

[2] F. G. Walther *et al.*, "Air-to-Ground Lasercom System Demonstration Design Overview and Results Summary," *Proc. SPIE*, vols. 7814, 2010, pp. 78140Y-1–9.

[3] C. D. Mobley, *Light and Water: Radiative Transfer in Natural Waters*, Academic Press, 1994.

[4] J. W. McLean, J. D. Freeman, and R. E. Walker, "Beam Spread Function with Time Dispersion," *Applied Optics*, vol. 37, 1998, pp. 4701–11.

[5] B. I. Erkmen, B. E. Moision, and K. M. Birnbaum, "A Review of the Information Capacity of Single-Mode Free-Space Optical Communication," *Proc. SPIE*, vol. 7587, 2010.

[6] J. A. Mendenhall *et al.*, "Design of an Optical Photon Counting Array Receiver System for Deep-Space Communications," *Proc. IEEE*, vol. 95, 2007, pp. 2059–69.

[7] J. P. Frechette *et al.*, "Readout Circuitry for Continuous High-Rate Photon Detection with Arrays of InP Geiger-Mode Avalanche Photodiodes," *Proc. SPIE*, vol. 8375, 2012, pp. 83750W-1–9.

[8] F. Hanson and S. Radic, "High Bandwidth Underwater Optical Communication," *Applied Optics*, vol. 47, 2008, pp. 277–83.

[9] J. B. Snow, J. P. Flatley, and D. E. Freeman, "Underwater Propagation of High-Data-Rate Laser Communications Pulses," *Proc. SPIE*, vol. 1750, 1992, pp. 419–27.

[10] C. Pontbriand *et al.*, "Diffuse High-Bandwidth Optical Communications," *OCEANS*, 2008.

[11] H. M. Oubei *et al.*, "2.3 Gbit/s Underwater Wireless Optical Communications Using Directly Modulated 520 nm Laser Diode," *Optics Express*, vol. 23, no. 16, 2015, pp. 20,743–48.

Biographies

Andrew S. Fletcher (fletcher@ll.mit.edu) is a technical staff member in the Optical Communications Technology Group at MIT Lincoln Laboratory. He received B.S. and M.S. degrees in electrical engineering from Brigham Young University in 2001, and received a Ph.D. in electrical engineering from the Massachusetts Institute of Technology in 2007. He specializes in classical and quantum communications theory, forward error correction, and laser communication system architectures.

Scott A. Hamilton [SM] is with MIT Lincoln Laboratory, where he leads the Optical Communication Technology group. In this role, he is responsible for multiple R&D programs to develop free-space laser and quantum communication system architectures and associated technologies. He studied electrical engineering at the University of California, Davis, and received B.S., M.S. and Ph.D. degrees in 1993, 1996, and 1999, respectively. He is a member of the OSA.

John D. Moores [SM] is assistant group leader in the Advanced Lasercom Systems and Operations group at MIT Lincoln Laboratory. He received a B.S. degree in electrical, computer, and systems engineering from Rensselaer Polytechnic Institute, and S.M. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology. He is a member of the Oceanographic Society.

Security and Privacy in Localization for Underwater Sensor Networks

Hong Li, Yunhua He, Xiuzhen Cheng, Hongsong Zhu, and Limin Sun

ABSTRACT

Underwater sensor networks are envisioned to enable a wide range of underwater applications such as pollution monitoring, offshore exploration, and oil/gas spill monitoring. Such applications require precise location information as otherwise the sensed data might be meaningless. On the other hand, security and privacy are critical issues as underwater sensor networks are typically deployed in harsh environments. Nevertheless, most underwater localization schemes are vulnerable to many attacks and suffer from potential privacy violations as they are designed for benign environments. However, a localization scheme that does not consider security and privacy could lead to serious consequences, especially in critical applications such as military monitoring. In this article, we discuss the security and privacy issues in underwater localization, and investigate the techniques that can provide security and preserve node privacy in underwater sensor networks.

Hong Li is with Beijing Key Laboratory of IOT Information Security Technology, Institute of Information Engineering, CAS and George Washington University.

Yunhua He is with Xidian University and George Washington University.

Xiuzhen Cheng is with George Washington University.

Limin Sun and Hongsong Zhu are with Beijing Key Laboratory of IOT Information Security Technology, Institute of Information Engineering, CAS.

Hongsong Zhu is the corresponding author.

INTRODUCTION

Underwater sensor networks consist of a variable number of underwater sensors and vehicles (unmanned underwater vehicles, autonomous underwater vehicles, etc.) to perform collaborative monitoring tasks over a given underwater area. Such a networking technology is envisioned to enable many underwater applications such as pollution monitoring, offshore exploration, and oil/gas spill monitoring.

To support these applications, underwater sensors and vehicles should first configure their locations for various tasks such as data tagging, node tracking, and target detection. Location information is also needed to improve the performance of medium access and network protocols. A large number of underwater localization schemes have been proposed in past years. These schemes typically involve two phases, the *location-related information collection* phase and the *position estimation* phase. In the first phase, nodes measure location-related information such as distance, angle, and hop count to each other or to anchors. In the second phase, their locations are estimated by a centralized node or calculated by themselves locally.

As most of the proposed localization schemes were designed without taking security into consideration, the two phases are both vulnerable to many security threats such as replay attacks, Sybil attacks, and wormhole attacks. Attackers can exploit these security loopholes to interfere with the localization process, or make estimated location imprecise, which could lead to serious consequences in many critical applications such as military monitoring. Furthermore, most of the localization schemes suffer from potential privacy leakage (e.g., location privacy), since a node must reveal a lot of information in order to be localized. Privacy leakage may make the nodes easily captured by enemies. It may also lead to many other security issues such as location spoofing attacks. In this article, we first discuss the security and privacy issues in localization for underwater sensor networks. Then we survey a few secure and privacy-preserving localization schemes and discuss their suitability for underwater sensor networks.

The remainder of this article is structured as follows. First, we present an overview of the localization schemes proposed for underwater sensor networks. Then the security and privacy issues of underwater localization are investigated, respectively. Finally, open research issues and challenges are discussed, and the article is concluded.

THE LOCALIZATION SCHEMES IN UNDERWATER SENSOR NETWORKS

Generally speaking, there are two kinds of nodes in an underwater sensor network: *unknown* or *to-be-localized* nodes with locations that need to be determined, and *anchor*, *reference*, or *beacon* nodes the locations of which are known a priori. Anchor nodes define the coordinate system and provide beacon signals to assist in localizing the unknown nodes. Many localization algorithms have been proposed for underwater sensor networks. These schemes typically consist of a location-related information collection phase and a position estimation phase.

0163-6804/15/$25.00 © 2015 IEEE

Location-Related Information Collection Phase

In the location-related information collection phase, a node estimates its distances, angles, or hop counts to other nodes or to anchor nodes. Such information will be fed to the next phase for position estimation. The methods of information collection can be classified as either range-based or range-free.

Range-Based Approaches: In range-based approaches, the distances between nodes are measured by time of arrival (ToA), time difference of arrival (TDoA), angle of arrival (AoA), or received signal strength indicator (RSSI). For example, the Underwater Positioning Scheme (UPS) [1] estimates the range differences between an unknown node and four anchor nodes based on TDoA. Given the locations of anchor nodes A, B, C, and D (as shown in Fig. 1), UPS calculates the range differences in two steps. In the first step, master node A, which is responsible for initiating a localization process, broadcasts a beacon signal. B replies to A with a beacon signal containing the time difference between receiving A's beacon and sending its own beacon. C replies to A with a beacon containing the time difference between receiving A's beacon and sending its own beacon after receiving beacon signals from both A and B. After receiving the beacons from A, B, and C, D performs the same process as B and C. Sensor S measures the arrival times of the beacon signals from anchor nodes A, B, C, and D locally. In the second step, these time differences are transformed into range differences from the unknown sensor to the anchor nodes, which are used in the trilateration equations to estimate the location of the sensor node.

Range-Free Approaches: In range-free approaches, nodes do not estimate distances or angles; instead, they measure hop count or network connectivity. For example, the Area Location Scheme (ALS) [2] estimates the area where an unknown node resides. In ALS, anchor nodes send out acoustic beacon signals at varying power levels. Each beacon packet contains the ID of the anchor node and the power level at which the signal is emitted. An unknown node passively listens to the beacon packets, keeps a list of IDs and their corresponding power levels, and sends this information to a sink node. The sink node processes the received information to estimate the area in which the unknown node is located based on the anchor nodes' locations and the signal propagation model. This process is illustrated in Fig. 2. Obviously, ALS can only provide a coarse location estimation for the unknown nodes within a certain area.

Position Estimation Phase

In this phase, the positions of unknown nodes are computed based on the information collected in the first phase. Position estimation can be either centralized or distributed.

Centralized Position Estimation: In centralized position estimation, the locations of the anchor nodes and the information collected in the first phase are sent to a centralized node. The centralized node then estimates locations for all the unknown nodes using techniques such as trilateration, multilateration, and triangulation. An example mechanism, the Hyperbola-Based Localization Scheme (HLS), was proposed in [3], which uses a hyperbola-based approach to localize unknown nodes. As shown in Fig. 3, an unknown node *D* sends a message to anchor nodes after detecting an event. Anchor nodes A, B, and C receive the message at times t_1, t_2, and t_3, respectively. Then these times and the locations of the anchor nodes are sent to a centralized node. As the difference between AD and BD is a constant that can be estimated by multiplying the speed of acoustic signals and the difference between t_1 and t_2, the unknown node is located on the hyperbola H_{AB}. Similarly, the unknown node is also located on the hyperbola H_{BC}. Then the centralized node can estimate the unknown node's location by computing the intersection of the two hyperbolas.

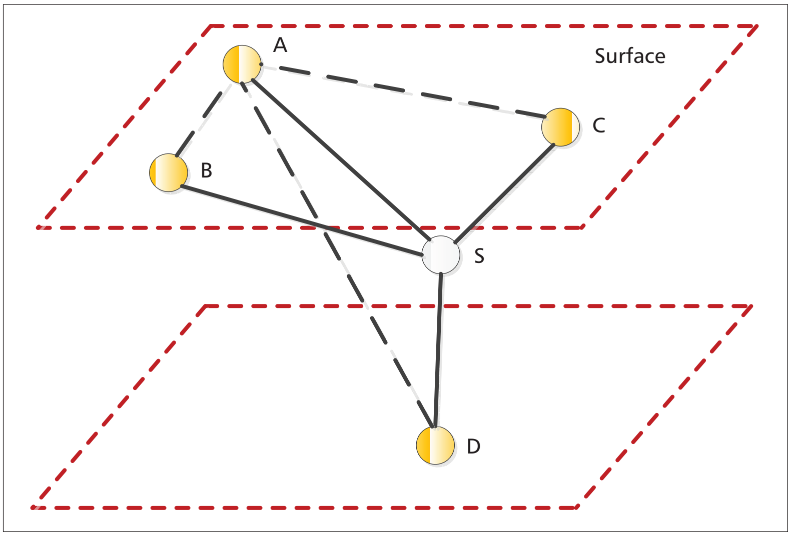


Figure 1. The Underwater Positioning Scheme (UPS).

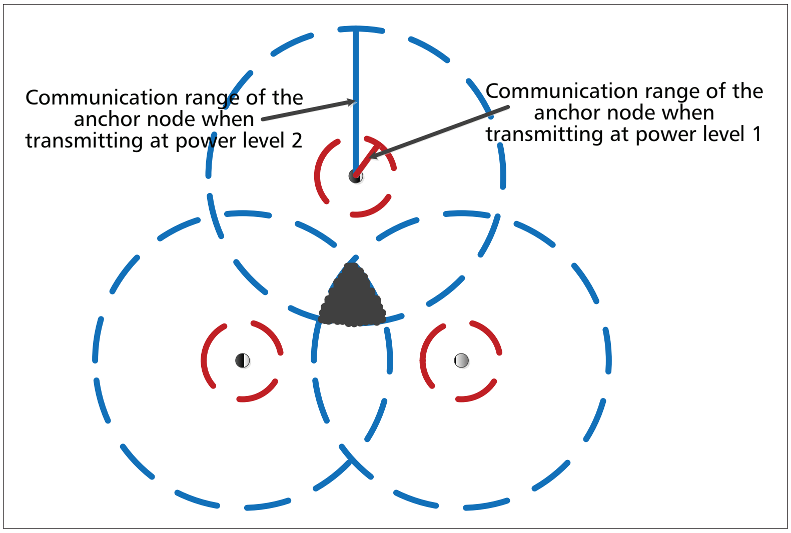

Communication range of the anchor node when transmitting at power level 2
Communication range of the anchor node when transmitting at power level 1

Figure 2. Three anchor nodes send out beacon messages at power levels 1 and 2. If an unknown node receives beacon messages from all three anchor nodes transmitting at power level 2, it resides in the shaded region

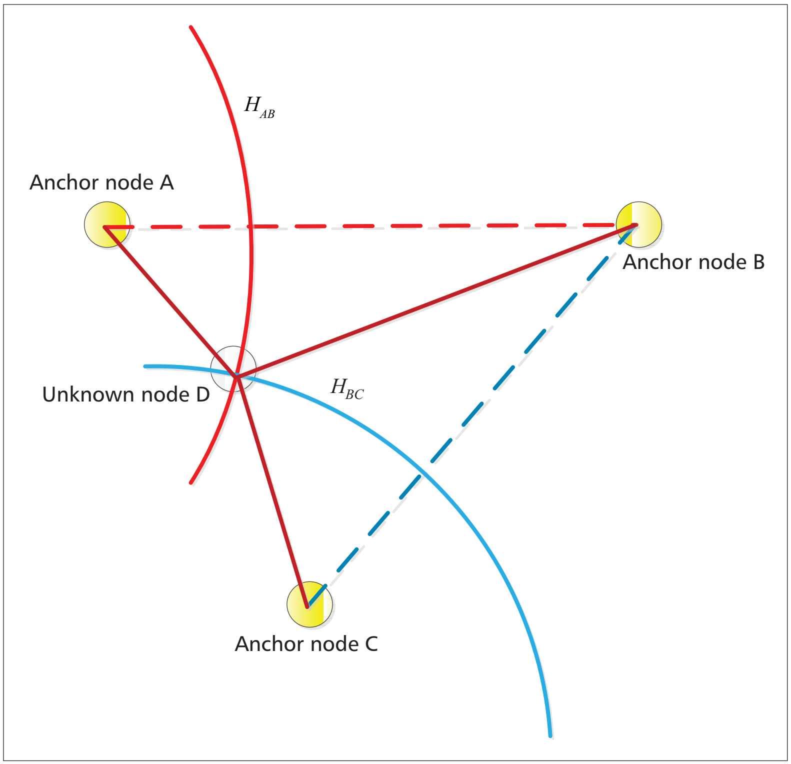


Figure 3. The Hyperbola-Based Localization Scheme.

Distributed Position Estimation: In distributed position estimation, each unknown node runs a localization algorithm locally after collecting location-related information. An example scheme was presented in [4], in which mobile anchor nodes first learn their coordinates via GPS before sinking and then broadcast beacons containing their positions as they are diving. Unknown nodes passively listen to the broadcast messages and use the ToAs of these messages to measure the ranges to the anchors. After hearing from several beacons, unknown nodes locally estimate their positions using triangulation.

Security Attacks on Underwater Localization and Countermeasures

The localization schemes described above perform well in secure environments. However, underwater sensor networks are usually deployed in harsh environments and operate unattended, making them extremely vulnerable to many security attacks. For instance, an attacker can disable the localization system or cause unknown nodes to have imprecise locations, which could lead to severe consequences in many critical applications. In this section, we investigate the security issues of underwater localization, and then discuss the techniques to secure underwater localization.

Attacks on Underwater Localization

DoS Attacks: Underwater localization can be affected by denial of service (DoS) attacks of the following kinds:

- Jamming attack: This is a common attack in wireless networks and has been well studied, especially for terrestrial sensor networks. Underwater links are mainly based on acoustic channels with narrow frequency bands; thus, an attacker can easily interfere with a physical channel to disable the localization process through narrowband jamming.
- Attacking critical nodes: In underwater localization systems, the failure of critical nodes such as anchors could directly affect the localization process. For example, if the master node in UPS [1] is compromised or destroyed by an attacker, the localization process may not be started.

Attacks on Range-Based Measurement: These attacks mainly target range-based localization schemes that first measure the distances between nodes. An attacker can make a node appear closer to or farther away from another node. Such attacks can be launched with many different ways:

- Replay attack: In a replay attack, an attacker first intercepts the message while jamming the legitimate communication channel, and then replays the same message. As with a jamming attack, a replay attack is not specifically designed for underwater localization, but it can also make unknown nodes get imprecise locations. When replay attacks are launched during the location-related information collection phase, an unknown node could get an imprecise propagation time and signal strength, causing imprecise distance estimation based on ToA/TDoA or signal strength.
- Delay/advanced response: In ToA/TDoA-based localization schemes, a node is supposed to reply immediately after receiving other nodes' beacon messages during the location-related information collection phase. A compromised node can delay the reply to appear farther away from the sender, or send the response before receiving other nodes' beacon messages to appear closer.
- Changing transmission power: In some range-based localization schemes, the distance between two nodes is estimated by the signal strength. A malicious node can change its transmission power to make it appear closer to or farther away from other nodes.

Attacks on Range-Free Measurement: In range-free localization, a malicious node can simply adjust its transmission power to change the network topology, which could result in imprecise estimation of hop count and proximity. This phase can also be affected by wormhole attacks. An attacker connects two or more nodes through low-latency direct links (called wormhole links) that can be established by a variety of means such as a cable in underwater. Then one node records a packet at a location in the network and forwards the message to its colluding partners in other parts of the network by wormhole links. After receiving the message, the colluding partners replay it. In this way the wormhole attack changes the network topology and deteriorates the positioning accuracy of range-free localization by either enlarging the neighborhood to affect proximity measurement (e.g., ALS [2]) or shortening the shortest routing path between two nodes to impact the hop count measurement (e.g., DV-Hop [5]).

Advertising False Information: Since most localization algorithms take locations of the anchor nodes as input, a malicious node can attack the location estimation process by providing false information:

- Providing false locations of the anchor nodes: A compromised anchor can send a false position to unknown nodes or centralized nodes. An advertised false position of an anchor node can lead to erroneous position computation even when the distances are precisely estimated. In [6], the authors showed that a receiver can easily be spoofed to an arbitrary location in GPS localization, which is also possible in underwater localization.
- Providing false range-based/range-free information: In centralized localization, all information collected in the first phase should be sent to a centralized node. Thus, an attacker can forge many identities and advertise erroneous information by a Sybil attack. The erroneous information can largely decrease the accuracy of localization even though other legitimate nodes measure the range-based/range-free information precisely in the first phase.

Non-Cooperation: Most of the localization algorithms require a minimal number of anchor nodes for location estimation. For example, HLS [3] needs at least three anchor nodes. If some of the anchor nodes are compromised or destroyed by an attacker, and the number of anchors falls below a threshold, location estimation can fail. In many distributed localization schemes, unknown nodes should cooperate to estimate their locations, as demonstrated in CLS [7] and UPS [1]. If an attacker compromises some of the nodes and launches DoS attacks, the localization estimation process may not function correctly. In centralized localization, an attacker can disrupt the location estimation process by compromising the centralized node that estimates the locations for the unknown nodes.

Techniques for Securing Underwater Localization

Compared to the traditional security attacks in terrestrial sensor networks, the attacks against underwater localization are more difficult to defend due to the unique characteristics of acoustic channels characterized by high bit error rate, large and variable propagation delay, and low bandwidth. Encryption is a straightforward way to address security attacks, but it consumes a lot of energy, and it also cannot defend against most of the attacks described above. Securing underwater localization is still an unexplored research area. Several secure localization schemes have been proposed in the last few years to provide secure positioning of the nodes in terrestrial sensor networks. Generally speaking, secure localization schemes can be classified into the following categories.

Misbehavior Detection: Such schemes were proposed to detect compromised nodes or location anomalies. In [8], the authors detected malicious anchor nodes by comparing the distance measured from the beacon signal of an anchor node and the distance calculated using the location information provided by the anchor node. If the difference between them is larger than the maximum distance error, the detecting node can infer that the received beacon signal is malicious, as depicted in Fig. 4. In [9], the authors identified location anomalies by verifying whether the derived locations were consistent with the deployment knowledge. In underwater sensor networks, nodes may drift with water current and oceanographic animals, rendering the proposed schemes mentioned above inapplicable underwater.

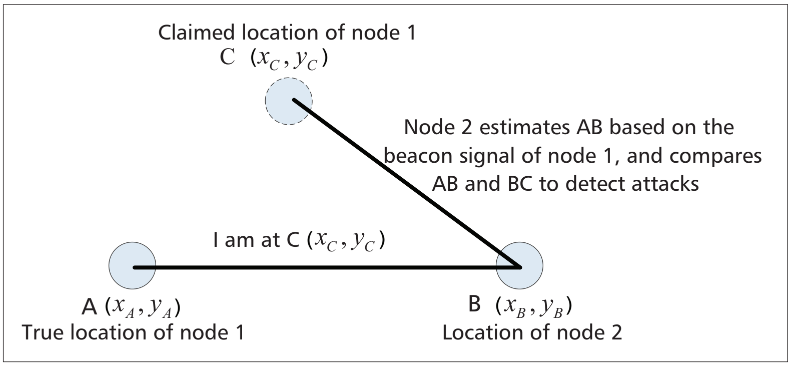

Claimed location of node 1
C (x_C, y_C)
Node 2 estimates AB based on the beacon signal of node 1, and compares AB and BC to detect attacks
I am at C (x_C, y_C)
A (x_A, y_A)
True location of node 1
B (x_B, y_B)
Location of node 2

Figure 4. If malicious node 1 advertises a false location C, the difference between AB and BC should be larger than the measurement error.

Robust Location Computation: Robust location computation aims to precisely estimate unknown nodes' locations in an untrusted environment. In [10], the authors developed two attack-resistant location estimation techniques to tolerate the malicious attacks against range-based location discovery in wireless sensor networks. The proposed scheme first identifies malicious location references by examining the inconsistency among them (indicated by the mean square error of the estimation), and defeats malicious attacks by removing malicious data. Then each anchor node votes on the cell in which the node may reside. In [10], the authors proposed a range-independent localization scheme called SeRLoc, which is robust against wormhole attacks, Sybil attacks, and the compromising of network entities. SeRLoc first detects attacks based on properties such as sector uniqueness and communication range violation using directional antennas and then filters out the attacked locators. Finally, unknown nodes determine their locations based on the beacon information transmitted by the locators, which are equipped with omnidirectional antennas.

Location Verification: Location verification schemes try to validate the reliability of location computation. In [12], the authors proposed a distance bounding protocol that can be used to verify the proximity of two devices connected by a wired link. In [13], the authors presented an Echo protocol to verify whether a node is inside a particular region. In Echo, a verifier sends a packet containing a nonce to the prover using RF; then the prover immediately echoes the packet back to the verifier using ultrasound; finally, the verifier checks whether the prover node is in the claimed region by estimating the round-trip time. This process is depicted in Fig. 5.

In many critical applications, the positions of the sensor nodes are very sensitive. Location privacy leakage could bring many problems. For example, enemies can easily destroy the anchor nodes to disable the whole network if they harvest the locations of the anchors in a military reconnaissance application.

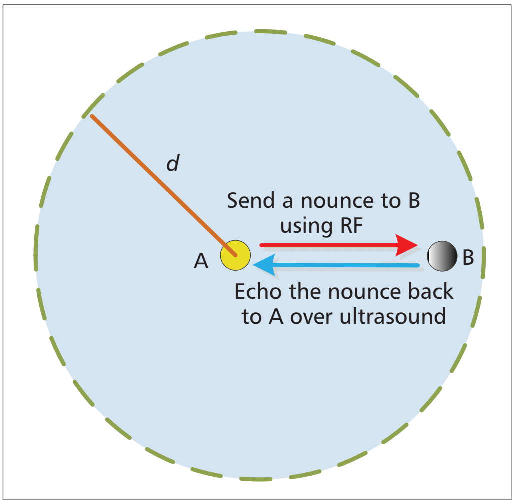

d
Send a nounce to B using RF
A
B
Echo the nounce back to A over ultrasound

Figure 5. If node B claims that it is located within the circle with a radius d centered at node A, the round-trip time should be less than $d/c + d/s$, where c is the speed of light and s is the speed of ultrasound.

Privacy Issues in Underwater Localization and Countermeasures

Besides the security attacks mentioned above, underwater localization also suffers from privacy vulnerabilities since a node must reveal certain information in order to be localized.

Identity Privacy of Underwater Localization

In most underwater localization schemes, different types of nodes may have different traffic patterns. For example, anchor nodes may periodically broadcast beacon messages while the node in charge of position estimation in a centralized localization mechanism may send the computed location information back to all the unknown nodes. Therefore, an attacker can thus simply sniff the traffic and then infer the identities of the nodes based on the traffic patterns. If a critical node is identified, it might be disabled by the attacker, causing the whole localization system to fail. Such an identity privacy disclosure may also result in other security vulnerabilities.

Location Privacy of Underwater Localization

In many critical applications, the positions of the sensor nodes are very sensitive. Location privacy leakage could bring many problems. For example, enemies can easily destroy the anchor nodes to disable the whole network if they harvest the locations of the anchors in a military reconnaissance application. Location privacy leakage may also lead to other security attacks such as the location spoofing attack. In the following, we discuss the location privacy issues in the location-related information collection phase and the location estimation phase.

Location Privacy in Location-Related Information Collection: In this phase, anchor nodes or unknown nodes broadcast beacon messages to measure the range, hop count, or network connectivity information. Attackers can passively sniff the beacon messages and estimate the distance to the sender based on the signal propagation model at different locations. Then the sender's location can be estimated using trilateration, as depicted in Fig. 6a. An attacker can also pretend to be an anchor node or unknown node and calculate its distances to authentic unknown/anchor nodes using TDoA, ToA, or RSSI at different locations; then it utilizes the computed range information to estimate the anchor's location by trilateration.

Location Privacy in Position Estimation: In both centralized and distributed localization, the positions of the anchor nodes are fed into the localization algorithms; thus, the anchors need to reveal their locations to the centralized node or all the unknown nodes, rendering such information potentially learnable by other nodes, as depicted in Fig. 6b. On the other hand, in centralized localization, the positions of all the unknown nodes are estimated by the centralized node, which implies that knowledge of the unknown nodes' locations is not limited to the unknown nodes themselves. If the centralized node is compromised, the unknown nodes' location privacy is violated.

Techniques to Preserve Privacy in Underwater Localization

Several schemes have been proposed in the past year to address the privacy issues in localization. The authors in [14] proposed a multi-lateral privacy-preserving localization mechanism in pervasive environments based on secure least squared error (LSE) estimation. In this scheme, privacy is protected as the position of an unknown node is calculated without the need to reveal anchors' locations, and the knowledge of the localization outcome is strictly limited to the unknown node itself. The authors in [15] developed a Privacy-Preserving WiFi Fingerprint Localization scheme (PriWFL), which utilizes homomorphic encryption to hide an unknown node's location during the localization process while preserving the location accuracy. These two schemes either involve a high communication overhead or require large computation capacity due to the use of computationally intensive cryptographic algorithms. Therefore, they cannot be used in underwater localization since the bandwidth of the acoustic links and the energy of underwater sensors are limited.

Open Research Issues and Challenges

Although many schemes have been proposed to secure the localization process and preserve the node privacy in terrestrial wireless sensor networks, they are not applicable underwater due to the unique characteristics of acoustic channels. In terrestrial wireless sensor networks, nodes use RF to establish the communication infrastructure. However, radio waves propagate at

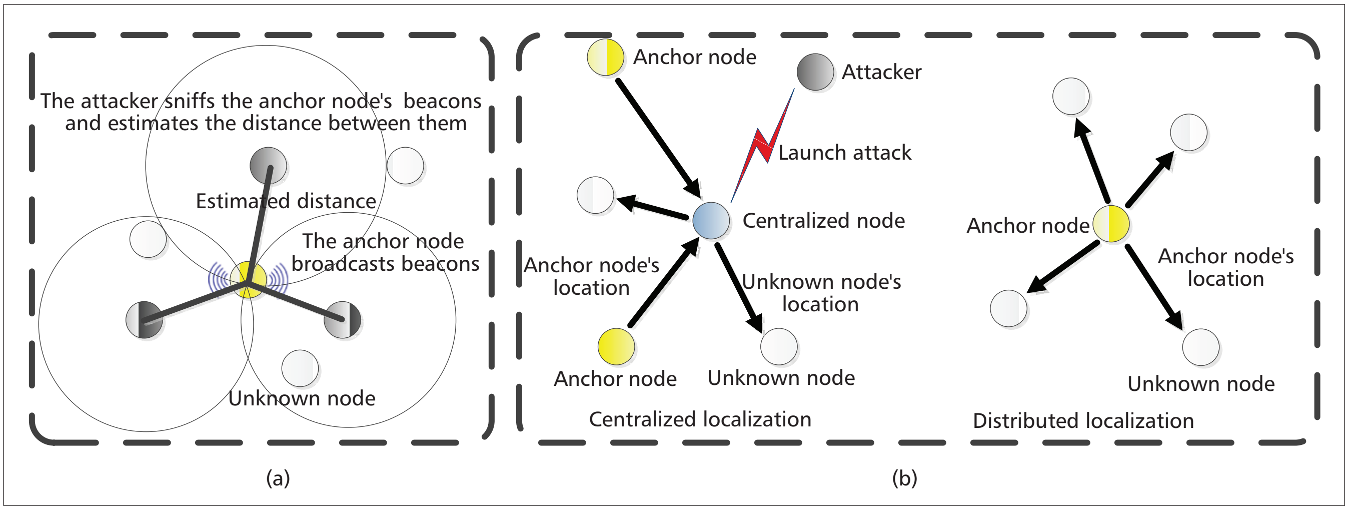


Figure 6. Location privacy leakages in underwater localization: a) location privacy leakage in the location-related information collection phase; b) location privacy leakages in the location estimation phase.

long distances through conductive sea water only at low frequencies (30–300 Hz), which requires large antennae and high transmission power. Therefore, underwater links are mainly based on acoustic wireless communication, which is characterized by high bit error rate, large and variable propagation delay, and low bandwidth. Securing underwater localization and preserving privacy are thus full of challenges. In the following, we summarize the major open research issues for securing localization and preserving privacy in underwater localization:

- Attack detection and location verification in a dynamic underwater environment. Under water, nodes can move with the water current and oceanographic animals, which complicates attack detection and location verification.
- Robust location computation in a noisy environment. The high bit error rate induced by fading, multipath, and refractive properties of the sound signal can cause transmission errors of critical security packets, which may lead to failure of secure localization schemes.
- Lightly weighted privacy-preserving schemes in resource-constrained underwater environments. Since acoustic channels are narrowband, and underwater nodes are energy-limited, the main challenge of designing a privacy-preserving scheme is how to make it lightly weighted without sacrificing location accuracy.

Conclusion

In this article, we discuss the security and privacy issues in underwater localization. We first review the major localization algorithms proposed for underwater sensor networks. Then we introduce the attacks against underwater localization and summarize a few secure localization schemes. We also analyze the privacy issues in underwater localization and discuss techniques for privacy preservation during the localization process. Finally, we outline the open research challenges of secure and privacy-preserving underwater localization.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 61472418), National High Technology Research and Development Program of China (Grant No. 2013AA011102), and the "Strategic Priority Research Program" of the Chinese Academy of Sciences (Grant No. XDA06040100).

References

[1] X. Cheng *et al.*, "Silent Positioning in Underwater Acoustic Sensor Networks," *IEEE Trans. Vehic. Tech.*, vol. 57, no. 3, 2008, pp. 1756–66.

[2] V. Chandrasekhar and W. Seah, "An Area Localization Scheme for Underwater Sensor Networks," *IEEE OCEANS 2006-Asia Pacific*, 2007, pp. 1–8.

[3] T. Bian, R. Venkatesan, and C. Li, "Design and Evaluation of a New Localization Scheme for Underwater Acoustic Sensor Networks," *IEEE GLOBECOM 2009*, pp. 1–5.

[4] M. Erol, L. F. Vieira, and M. Gerla, "Localization with *Dive'n'Rise* (DNR) Beacons for Underwater Acoustic Sensor Networks," *Proc. 2nd Wksp. Underwater Networks*, ACM, 2007, pp. 97–100.

[5] D. Niculescu and B. Nath, "Dv Based Positioning in Ad Hoc Networks," *Telecommun. Sys.*, vol. 22, no. 1–4, 2003, pp. 267–80.

[6] N. O. Tippenhauer *et al.*, "On the Requirements for Successful GPS Spoofing Attacks," *Proc. 18th ACM Conf. Comp. and Commun. Security*, ACM, 2011, pp. 75–86.

[7] D. Mirza and C. Schurgers, "Collaborative Localization for Fleets of Underwater Drifters," *IEEE OCEANS 2007*, 2007, pp. 1–6.

[8] D. Liu, P. Ning, and W. Du, "Detecting Malicious Beacon Nodes for Secure Location Discovery in Wireless Sensor Networks," *Proc. 25th IEEE Int'l. Conf. Distrib, Comp, Sys,*, 2005, pp. 609–19.

[9] W. Du, L. Fang, and P. Ning, "Lad: Localization Anomaly Detection Forwireless Sensor Networks," *Proc. 19th IEEE Int'l. Parallel and Distrib, Processing Symp.*, 2005, pp. 41a–41a.

[10] D. Liu, P. Ning, and W. K. Du, "Attack-Resistant Location Estimation in Sensor Networks," *Proc. 4th Int'l. Symp. Info. Processing in Sensor Networks*, IEEE Press, 2005, p. 13.

[11] L. Lazos and R. Poovendran, "Serloc: Robust Localization for Wireless Sensor Networks," *ACM Trans. Sensor Networks*, vol. 1, no. 1, 2005, pp. 73–100.

[12] S. Brands and D. Chaum, "Distance-Bounding Protocols," *Advances in CryptologyłEUROCRYPT'93*, 1994, Springer, pp. 344–59.

[13] N. Sastry, U. Shankar, and D. Wagner, "Secure Verification of Location Claims," *Proc. 2nd ACM Wksp. Wireless Security*, 2003, pp. 1–10.

[14] T. Shu *et al.*, "Multi-Lateral Privacy Preserving Localization in Pervasive Environments," *2014 Proc. IEEE INFOCOM*, Apr. 2014, pp. 2319–27.

[15] H. Li *et al.*, "Achieving Privacy Preservation in WiFi Fingerprint-based Localization," *2014 Proc. IEEE INFOCOM*, 2014, pp. 2337–45.

Biographies

HONG LI is a Ph.D. student at the University of the Chinese Academy of Sciences. He received his B.A. from Xi'an Jiaotong University. He currently works under Prof. Limin Sun in the Institute of Information Engineering, Chinese Academy of Sciences. His primary research interests include security and privacy in wireless networks, and localization.

YUNHUA HE is a Ph.D. student at Xidian University. He received his B.A. from Wuhan insititute of Technology. He currently works under Prof. Limin Sun at the Beijing Internet of Things Security Center. His primary research interests include location privacy for vehicular ad hoc network, and incentive mechanisms for mobile social networks.

LIMIN SUN [M] received his B.S., M.S., and Ph.D. degrees from the College of Computers, National University of Defense Technology in 1988, 1995, and 1998, respectively. Currently, he is a professor in the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include security and privacy in wireless networks, wireless sensor networks, and the Internet of Things. He is a Senior Member of the China Computer Federation (CCF).

XIUZHEN CHENG [F] received her M.S. and Ph.D. degrees in computer science from the University of Minnesota — Twin Cities in 2000 and 2002, respectively. She is a professor in the Department of Computer Science, The George Washington University, Washington, DC. Her current research interests include cyber physical systems, wireless and mobile computing, sensor networking, wireless and mobile security, and algorithm design and analysis. She worked as a program director for the U.S. National Science Foundation (NSF) for six months in 2006 and joined the NSF again as a part-time program director in April 2008. She received the NSF CAREER Award in 2004.

HONGSONG ZHU [M] (zhuhongsong@iie.ac.cn) received his Ph.D. degree from the Institute of Computing, Chinese Academy of Sciences. He is an associate professor in the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include security and privacy in wireless networks, indoor localization, wireless sensor networks, and the Internet of Things. He is a Senior Member of CCF.

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.

Software-Defined Underwater Acoustic Networks: Toward a High-Rate Real-Time Reconfigurable Modem

Emrecan Demirors, George Sklivanitis, Tommaso Melodia, Stella N. Batalama, and Dimitris A. Pados

Emrecan Demirors and Tommaso Melodia are with Northeastern University.

George Sklivanitis, Stella N. Batalama, and Dimitris A. Pados are with the State University of New York at Buffalo.

ABSTRACT

We review and discuss the challenges of adopting software-defined radio principles in underwater acoustic networks, and propose a software-defined acoustic modem prototype based on commercial off-the-shelf components. We first review current SDR-based architectures for underwater acoustic communications. Then we describe the architecture of a new software-defined acoustic modem prototype, and provide performance evaluation results in both indoor (water tank) and outdoor (lake) environments. We present three experimental testbed scenarios that demonstrate the real-time reconfigurable capabilities of the proposed prototype and show that it exhibits favorable characteristics toward spectrally efficient cognitive underwater networks, and high data rate underwater acoustic links. Finally, we discuss open research challenges for the implementation of next-generation software-defined underwater acoustic networks.

INTRODUCTION

Underwater acoustic networks (UANs) are an emerging research topic because of the key role that this technology will play in military and commercial applications including disaster prevention, tactical surveillance, offshore exploration, pollution monitoring, and oceanographic data collection. A key challenge in the design of UANs stems from the characteristics of the underwater acoustic (UW-A) channel, which exhibits high path loss, noise, multipath, high and variable propagation delay, and Doppler spread. Therefore, reliable communication links are practically feasible only at low data rates. Additionally, the propagation challenges in the underwater environment result in *temporally* and *spatially* varying UW-A channel coefficients, which drives research efforts toward the design of specialized protocols at different layers of the network protocol stack — often with a cross-layer approach.

As of today, the majority of existing deployed UANs are based on commercially available acoustic modems. Even though commercial modems enable a wide range of applications, they rely on inherently *fixed hardware* designs and proprietary protocol solutions that are regrettably far from satisfying the emerging reconfigurability needs of next-generation UANs. As a result, practical deployment of new protocols is either not feasible or prohibitive in terms of both implementation cost and time. Evidently, fixed hardware and closed software architectures that characterize commercial underwater modems prevent UAN applications from benefiting from the latest algorithmic developments.

Lack of standardization agreements for UANs impose additional hurdles in the design of reconfigurable networks. For example, different vendors equip underwater modems with proprietary communication protocols with different implementation requirements across different hardware and software platforms, which, in the end, prevents their integration in heterogeneous UANs. A sizable body of research work in UANs focuses on the development of software/protocol standards that resolve *interoperability* issues among different underwater modems (e.g., JANUS [1]). However, existing architectures are still far from being able to achieve the data rate performance and flexibility required by the next generation of UAN applications.

Finally, the spatial and temporal variations of the UW-A channel require reconfiguration of the communication parameters of UAN devices to provide stable performance in terms of bit error rate (BER) and packet error rate (PER). Currently, commercial modems address adaptation to channel variations by employing a set of predefined operational modes (pre-fixed set of communication parameter values). However, such ad hoc solutions lack i) the ability to switch in real time among a finite number of operational modes, ii) decision making mechanisms to decide and apply adaptation, and iii) the ability to dynamically adapt to all the possible environments because of the finite number of operational modes. Consequently, there is a need for UAN devices that can i) ease deployment and testing of new protocol

0163-6804/15/25.00 © 2015 IEEE

designs, ii) bridge the gap between different network devices and protocols to resolve the interoperability problem in heterogeneous UANs, and iii) intelligently decide and adapt their communication parameters or technology based on the environmental needs in real time.

Software-defined radio (SDR) has recently emerged as a technology platform that enables rapid prototyping of fully agile, intelligently adaptive, and reconfigurable networking devices to accommodate and test novel wireless networking protocols for RF communications. Considering the unique capabilities and features of SDR in RF and the need for flexible, easily reconfigurable UAN devices, in this article, we investigate the software/hardware challenges to build a software-defined acoustic modem (SDAM) and discuss its potential benefits for future UANs. To that end, we review existing software-defined efforts in UW-A and we propose an SDAM prototype based on commercial off-the-shelf (COTS) components. We design and implement physical layer schemes and decision algorithms that exploit both time and frequency degrees of freedom, and we demonstrate the real-time reconfigurable capabilities of the proposed modem with field experiments.

The remainder of this article is organized as follows. We first provide a brief overview of the SDR paradigm and current SDR-based architectures for UW-A communications. Then we propose and describe an SDAM architecture and provide experimental performance evaluation results in both indoor (water test tank) and outdoor (lake) environments. Finally, we discuss open research implementation challenges for future next-generation software-defined UANs.

Software-Defined Architectures for UW-A Communications

SDRs are now prevalent communication platforms in both commercial and military RF applications mainly because of the need for flexible and reconfigurable radio solutions and their ability to follow the rapid evolution of enabling technologies such as analog-to-digital converters (ADCs), digital-to-analog converters (DACs), general-purpose processors (GPPs), field-programmable gate arrays (FPGAs), and graphical processing units (GPUs). However, so far, SDR platforms have seen limited use in the context of underwater communications, primarily because of implementation and development cost challenges compared to RF technologies. A typical SDR architecture comprises a front-end that may include amplifiers, filters, and mixers, connected to an ADC/DAC and a softwarereprogrammable digital processing unit such as a digital signal processor (DSP), FPGA, GPP, or GPU. The SDR architecture provides the basis for software reconfigurable processing at the physical (PHY) and medium access control (MAC) layers, often through existing software tools, including GNU Radio, Simulink, and LabView, which are interoperable with higher-layer protocol implementations (e.g., TCP/IP).

SDR platforms have been adopted by several military programs including the U.S. Joint Tactical Radio System (JTRS) and NATO STANAG 5066, as well as by a variety of commercially available products that feature a mixture of hardware configurations at different cost, for example, USRP, RTL-SDR, HackRF, BladeRF, and PicoSDR. Open-source software packages, such as GNU Radio interface well with SDR systems, allow custom development of application-specific signal processing blocks, and have recently demonstrated their support for embedded devices too such as Xilinx Zedboard and Xilinx ZC702. As a result, a software-defined framework appears to be a powerful proposition for underwater communications that will be able to provide a low-cost and programmable/reconfigurable hardware alternative to commercially available underwater modems. However, SDR principles are not directly applicable to the underwater environment due to the particular (and challenging) characteristics of the underwater channel.

Current underwater networking research can be classified as follows:

- Efforts that involve hardware development of new experimental modems and pertinent software across some or all layers of the protocol stack
- Efforts that involve primarily software development for the network and application layers of the protocol stack of UANs that use commercially available modems

The majority of existing proposals offer only data logging capabilities for offline processing and do not report real-time reconfigurability and online performance evaluation through experimental field studies.

The main objective of high-layer software proposals in underwater networks is the development of a framework that facilitates integration of simulation, emulation, and testing policies and can interface with existing commercial or experimental underwater modems. The work in [2] reviews and compares two software frameworks, SUNSET and DESERT, implemented using the open source network simulators ns-2 and ns-2 miracle. SUNSET has been used as the standard plug-in in the SUNRISE [3] project. SUNRISE is a federation of testing infrastructures that is designed to represent different marine environments and support different applications. OceanTUNE [4] presents a similar testbed suite that provides flexibility in system and network configurations. Aqua-Net [5] is a software framework that uses WHOI micro-modems and Teledyne Benthos modems, and offers a layered architecture that enables cross-layer optimization. The work in [6] proposes an agent-based (service-oriented) architecture on a Java Virtual Machine (JVM), therefore allowing portability from simulation to real-time deployment. A more recent work [7] attempts to depart from OSI-layered implementations and follow a modular cross-layer software-defined architecture.

Lower-layer/hardware proposals in underwater networks concentrate on hardware development of underwater modems and explore their interface with hybrid software architectures that offer reconfigurability at the PHY and MAC layers. In particular, the work in [8] studies a modem architecture that is based on the open-source software tools GNU Radio, TinyOS, and TOSSIM, and the USRP hardware. The work in

There is a need for UAN devices that can i) ease deployment and testing of new protocol designs, ii) resolve the interoperability problem in heterogeneous UANs, and iii) intelligently decide and adapt their communication parameters in real time.

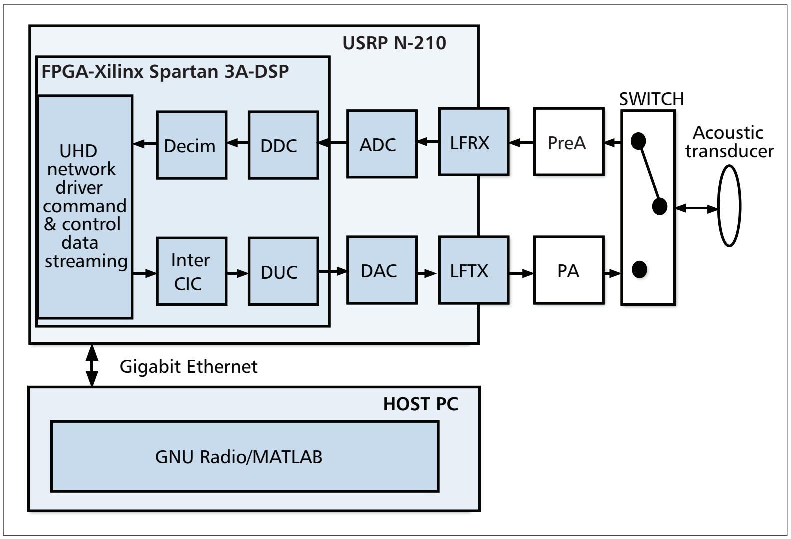


Figure 1. Hardware architecture of the proposed SDAM prototype.

[8] builds an underwater sensor network mote and parallels pertinent developments in the software radio and sensor network literature. Similarly, [9–11] provide early modem prototypes that are based on either FPGA/DSP or FPGA-only technologies and develop in-house software for controlling the PHY and MAC layer parameters.

Proposed SDAM Prototype

In this section, we present our developments toward the design and implementation of a variable-rate fully reconfigurable SDAM that is built from a collection of COTS components. We demonstrate and evaluate, in real-time underwater field experiments, the software-defined capabilities of the proposed modem/design. We provide a complete overview (from hardware to software) of the SDAM implementation, as well as our design considerations for the PHY, MAC, and network layers of the protocol stack.

Hardware Setup

The SDAM is at its core an SDR connected with wideband acoustic transducers through power amplifiers/preamplifiers. The SDAM exploits the unique capabilities and features of an SDR to fulfill the need for flexible, easily reconfigurable UAN devices [12].

The overall hardware architecture of the proposed SDAM prototype is illustrated in Fig. 1, while Fig. 2 depicts the prototype itself. The proposed SDAM is based on a USRP N210, which is a commercially available FPGA-based, SDR platform. We chose to work with LFTX and LFRX daughterboards (DC –30 MHz), which enable the development of a half-duplex transceiver operating in the frequency range of the selected omnidirectional acoustic transducer, Teledyne RESON TC4013, from 1 Hz to 170 kHz. To enhance the communication range of our SDAM, we used a linear wideband power amplifier (PA), Benthowave BII–5002, and a voltage preamplifier (PreA), Teledyne RESON VP2000. We also incorporated an electronic switch, Mini-Circuits ZX80–DR230+ to enable the operation of a single acoustic transducer as transmitter and receiver in a time-division duplex fashion. Baseband signal processing algorithms and protocols are mainly implemented in the host PC, which is connected to the USRP N210 through Gigabit Ethernet (GigE). Each prototype costs approximately $6000 (commercially available alternatives are almost 2× more expensive).

Physical Layer

We consider the design and implementation of two different signaling technologies for different applications based on zero-padded orthogonal frequency-division multiplexing (ZP-OFDM) and direct sequence spread-spectrum (DS-SS). Both technologies aim to maximize channel utilization and spectral efficiency in the challenging underwater environment by providing *online* adaptation at both PHY and MAC layers.

ZP-OFDM: We adopt a ZP-OFDM with a superimposed convolutional error correction coding scheme. In particular, we use a K-subcarrier ZP-OFDM where zero-padding works as a zero-power guard interval alternative to the conventional cyclic prefix. Before subcarrier allocation, data symbols are modulated with either binary phase shift keying (BPSK) or quadrature PSK (QPSK), while $K_P = K/4$ are used as pilot, K_D as data, and K_N as null subcarriers. A guard interval is added between each OFDM symbol to avoid interference among different OFDM symbols, and a pseudorandom noise (PN) sequence is transmitted at the beginning of each OFDM packet for frame detection and coarse synchronization purposes at the receiver. At the receiver side, a low-pass filter (LPF) is first used to reject out-of-band noise and interference. Null subcarriers are used for Doppler scale estimation and compensation, while pilot subcarriers are used for channel estimation and fine symbol synchronization in the frequency domain.

DS-SS: In this scheme, BPSK data symbols are modulated by a random binary code-waveform of length L and duration $1/(B - A)$; thus, transmitted waveforms occupy the whole available bandwidth (A Hz, B Hz) of the device, providing efficient utilization of the underwater acoustic spectrum resources. To avoid inter-symbol interference we ensure that the product between code length L and waveform duration is larger than the multipath spread; hence, we choose large code lengths, which leverage the capability of the selected technique for multiple access. More specifically, we adopt two different transmitter designs, one where a long PN sequence is used for frame detection, synchronization, and channel estimation at the receiver side, and a second where an all-1 sequence of unmodulated bits is used to achieve frame detection and coarse synchronization at the receiver [13]. For the second transmitter setup, we propose a blind receiver design that integrates and executes synchronization, channel estimation, and demodulation in a combined fashion with a low-complexity RAKE-type receiver [13]. The proposed receiver structure offers significant computational efficiency and resistance to multiple access interference, therefore sparking interest in future applications on heavily utilized spectrum resources.

User-Defined Decision Algorithms

Herein, we define the decision algorithms for adaptation in both the ZP-OFDM and DS-SS physical layer schemes. However, we first need to define a reliable feedback communication method to support real-time forward link adaptation, as potential unsuccessful feedback delivery may result in failure of the forward link as well. To that end, we propose and implement a feedback method based on binary chirp spread-spectrum (B-CSS) modulation, which is known to be resilient against the severe multipath and Doppler effects that characterize the UW-A channel. In addition, B-CSS provides a robust low data rate communication scheme with a low-complexity receiver design that very well fits the reliability and robustness requirements of an underwater feedback channel. More specifically, we leverage the quasi-orthogonality of up and down chirps by encoding a "1" bit with an up chirp and a "0" bit with a down chirp. Up and down chirp signals are generated by proper selection of the time-varying instantaneous frequency that ranges from f_0 to f_1, and the period T of a chirp signal. Therefore, if the chirp frequency variation rate defined as $\mu = (f_1 - f_0)/T$ is positive, we generate an up chirp; otherwise, if $\mu < 0$, we generate a down chirp.

Decision algorithms provide the SDAM with the capability to either adapt/change specific communication parameters (of a pre-selected communication technology) or switch between different communication technologies such as ZP-OFDM and DS-SS. Decisions are made by the receiver node based on user-defined algorithms, and are then communicated to the transmitter through the wireless feedback link. We consider three decision algorithms that best illustrate the adaptation capabilities of the proposed SDAM under preset performance constraints.

First, we consider adaptation of the modulation and the error correction coding rate in a ZP-OFDM link to solve a rate maximization problem under preset BER reliability constraints. We assume that the number of data subcarriers, as well as the symbol and guard interval duration in ZP-OFDM remain fixed, and data rate is a function of the modulation order and error correction coding rate. The receiver estimates the signal-to-interference-plus-noise ratio (SINR) per received packet, and jointly decides on the combination of a modulation (BPSK or QPSK) and error correction coding (1/2-rate convolutional code or uncoded) scheme to satisfy a predefined BER threshold constraint.

Second, we provide the SDAM with the capability to switch between ZP-OFDM and DS-SS communication technologies. This is achieved by designing a multiplexer-based adaptation mechanism that is activated upon successful decoding of an incoming packet. The SDAM is preprogrammed with all the primitive modules required for the implementation of both ZP-OFDM and DS-SS physical layer schemes. The enabling signal of the multiplexer is controlled by a Boolean variable that enables DS-SS transmission/reception upon successful reception of a ZP-OFDM packet and vice versa.

Finally, we consider real-time code-waveform adaptation in a DS-SS link to maximize the pre-detection SINR at the destination receiver of interest in the presence of multiple access interference. The receiver first senses the environment and calculates a disturbance (noise plus interference) autocorrelation matrix during the time that the user of interest is silent. Then the receiver estimates the pre-detection SINR for each incoming packet and looks for the best channel waveform of length L that is the solution to an SINR maximization problem. The new channel waveform is communicated to the transmitting source of interest through a chirp-based feedback link. Both the transmitter and receiver need to be synchronized. We implement two parallel processing receiver chains, one working with the updated and another with the preset channel waveform.

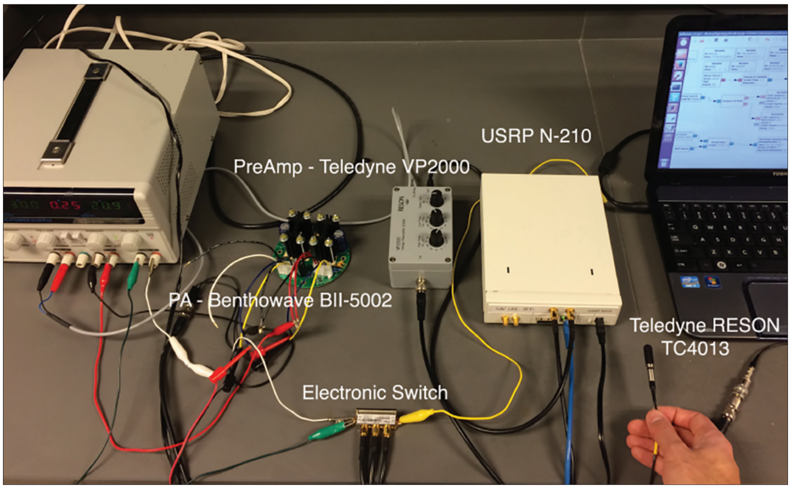


Figure 2. The proposed SDAM prototype.

Medium Access Control

We implement a simple MAC protocol that can support user-defined decision algorithms. Specifically, at the time an SDAM node has data available for transmission, it directly accesses the channel, transmits a data packet, and switches into receive mode, waiting for a feedback message from the destination node. A timeout time is also set for retransmission. In the absence of data, all nodes perform idle listening. Even though the current implementation includes one simple MAC protocol design, several others can be implemented by exploiting primitive built-in functionalities such as timer operations, idle listening, and retransmission.

Network Layer

The proposed SDAM prototype can support IPv4 and IPv6 protocols through an adaptation layer [14]. The adaptation layer offers a set of functionalities, i.e., IP header compression, IP packet fragmentation that enable the interfacing of the traditional IP network layer with the MAC layer of the prototype. Specifically, the traditional IPv4 and IPv6 headers are optimized for underwater acoustic communications to minimize the network delay and energy consumption.

Software Developments

We use an open source software framework called GNU Radio to drive adaptive baseband signal processing. GNU Radio offers a plethora of digi-

GNU Radio offers a plethora of digital signal processing blocks which are implemented in C++, and are usually wrapped into Python classes. Blocks are either instantiated by Python scripts, or used as building primitives of a communication flowgraph.

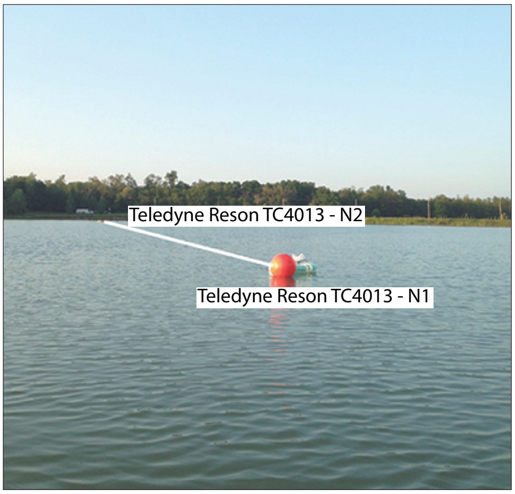


Figure 3. Underwater transducers are held by the red buoys and are deployed 322 ft apart.

tal signal processing blocks that are implemented in C++, and are usually wrapped into Python classes. Blocks are either instantiated by Python scripts or used as building primitives of a communication flowgraph. GNU Radio provides users with the ability to design custom-logic signal processing blocks, simulate them offline, and finally embed them into existing flowgraph designs. In this context, we design and implement custom signal processing blocks and build ZP-OFDM and DS-SS flowgraphs, while at the same time exploit the message passing capabilities of GNU Radio to support adaptive transmitter and receiver physical layer functionalities. Message passing allows the exchange of control messages between different blocks that are located either downstream or upstream in the GNU Radio flowgraph. Therefore, we achieve real-time adaptation by leveraging both GNU Radio's asynchronous messaging features and the properties of a chirp-based feedback link. Decoding of the feedback messages relies on two correlation filters equipped with an up chirp and a down chirp, respectively. We decide on the feedback message bits by comparing the outputs of the respective correlation filters at the receiver, and we update the PHY layer parameters in the forward link flowgraph accordingly. For the needs of communication technology adaptation between ZP-OFDM and DS-SS, we also implement a software multiplexer block.

Experiments

In this section, we present experimental results from three different testbed scenarios and evaluate the proposed SDAM prototype in both indoor and outdoor environments.

Physical Layer Online Adaptation

A series of experiments took place in Lake LaSalle at the State University of New York at Buffalo. Lake LaSalle has a depth of approximately 7 ft. We deployed two SDAM prototypes 322 ft apart from each other as illustrated in Fig. 3. We used ZP-OFDM signals that occupy a bandwidth of B = 24 kHz at a carrier frequency f_c = 100 kHz. We defined K = 1024 subcarriers for each OFDM symbol, where each subcarrier is either mapped with BPSK or QPSK modulation and is either coded (rate 1/2 convolutional error correction codes) or not coded. We also incorporated a guard time of T_g = 15 ms between each OFDM symbol.

The objective of this set of experiments is to demonstrate the real-time adaptation capabilities of our SDAM prototype at the physical layer in an outdoor lake environment. Figure 4 depicts experimental data rate and BER evaluation results of both an adaptive modulation/coding rate and a fixed (non-adaptive) scheme for 21 OFDM packet transmissions. The modulation scheme and coding rate of ZP-OFDM change according to a user-defined decision algorithm that aims to maximize data rate under preset BER threshold constraints. In Fig. 4 the BER threshold is empirically set to 10^{-3}, while SINR varies between 10 and 20 dB.

In both adaptive and fixed (non-adaptive) schemes, the SDAM, N_1, starts transmitting at the highest possible data rate, which requires uncoded QPSK modulation. We observe that since the non-adaptive scheme does not provide the SDAM with modulation/coding adaptation capabilities, data rate will remain constant. However, at the time that the SINR profile changes (6th packet), the preset BER threshold is not satisfied. On the other hand, in the adaptive scheme, the SDAM adapts the modulation and coding rate in real time as soon as the SINR decreases. Specifically, when the estimated SINR is lower than 10 dB, the SDAM incorporates 1/2-rate error correction coding and changes the modulation scheme from QPSK to BPSK. As a result, the data rate is adjusted to a lower value to satisfy the predefined BER constraints. Subsequently, as soon as the SINR increases (from packet 11 to 17), the SDAM changes into uncoded BPSK transmission to compensate for the data rate loss. Finally, the SDAM switches back to uncoded QPSK when the SINR reaches the initial level of 20 dB.

Cognitive Channelization

The following set of experiments considers the deployment of three SDAMs in a water test tank. Figure 5 depicts three SDAM nodes, which operate in the same frequency band, 91–99 kHz, and use random binary waveforms of length L = 63. N_1, and N_2 act as transmitters, and N_3 as a receiver, while feedback messages regarding the best channel waveform are exchanged between N_1 and N_3. Both transmitters use square-root raised cosine pulses of duration T_d = 0.16384 ms, and roll-off factor α = 0.35. The number of channel paths is set to N =20 and we measured a multipath spread of 31 waveform bits or 5.1 ms. Both N_1 and N_2 use the same transmission power and are positioned 5.5 ft apart from N_3. The distance between N_1 and N_2 is set at 1.5 ft.

Figure 6 illustrates SINR experimental studies for two different setups. The first setup considers a single transmitter node, N_1, that operates in the absence of N_2. The second setup considers two nodes, N_1 and N_2, that transmit simultaneously at the same frequency band, while a third node, N_3, uses a control channel at a different frequency band to periodically com-

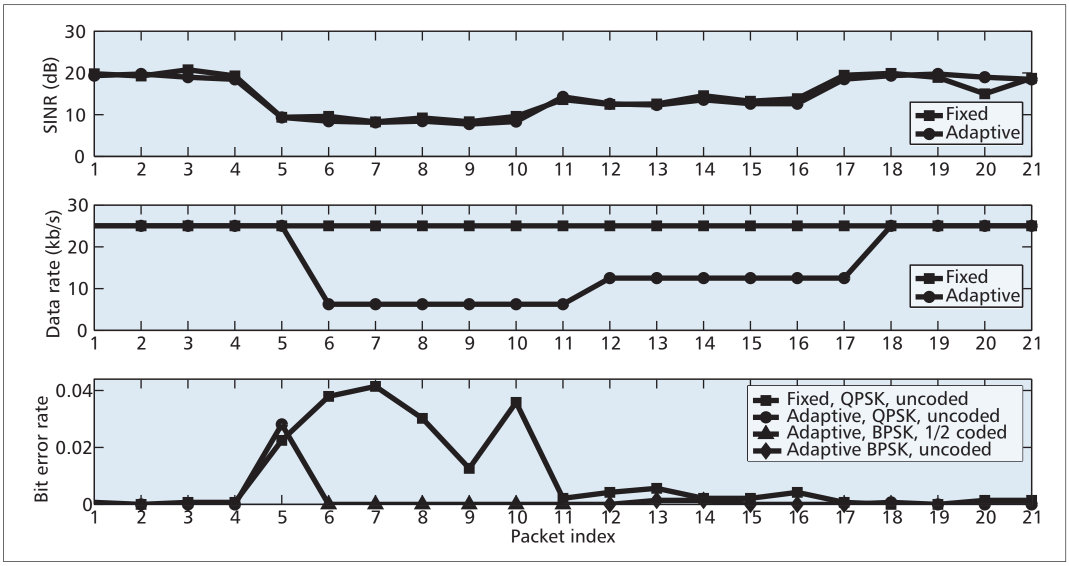


Figure 4. Comparison of adaptive with fixed (non-adaptive) scheme in terms of data rate, and BER for different SINR profiles.

Efforts towards reconfigurable underwater networks can certainly benefit from recent and parallel developments in the SDR literature. The adoption of SDR frameworks in underwater communications must carefully consider the challenges imposed both by the available technology and the underwater environment.

municate to N_1 a new channel waveform. For each setup we consider two different scenarios referred to as cognitive and fixed channelization, respectively. Cognitive channelization assumes that the transmitter node N_1 adapts its channel waveform based on the interference profile of N_2 and the additive noise. In the absence of N_2, waveform adaptation is based only on the noise profile. For the fixed channelization scenario, N_1 does not change the preset waveform. We observe that for both setups, cognitive channelization significantly outperforms fixed channelization in terms of receiver predetection SINR.

The experimental studies in this section show that underwater cognitive channelization is beneficial in both the presence and absence of interference from other signals/users due to the fact that the SDAM efficiently adapts to channel conditions [13].

Communication Technology Adaptation (Future Heterogeneous UW Networks)

Given the primitives of both ZP-OFDM and DS-SS, we also tested the capability of the proposed SDAM prototype to adapt/choose among different communication technologies. We designed an experimental scenario that aims at seamless switching between OFDM packet (BPSK modulation, no error correction coding, $K = 1024$ subcarriers in a bandwidth of 24 kHz) and DS-SS packet (BPSK modulation, no error correction coding, and waveform length $L = 31$ in a bandwidth of 6 kHz) transmissions at carrier frequency $f_c = 100$ kHz, and successfully demonstrated the dual mode capability of the prototype.

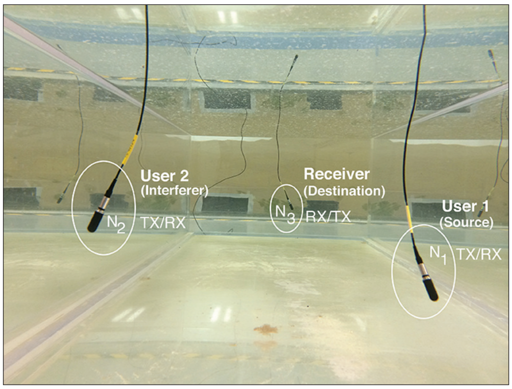


Figure 5. Cognitive channelization water test tank setup.

Software-Defined UANs: Open Research Challenges

Reconfigurability is an essential feature of next-generation underwater networks in early stages of study, especially in the context of real-time joint adaptation across all layers of the protocol stack. Efforts toward reconfigurable underwater networks can certainly benefit from recent and parallel developments in the SDR literature. The adoption of SDR frameworks in underwater communications must carefully consider the challenges imposed by both the available technology and the underwater environment. Below, we present an overview of the core challenges related to software-defined UAN implementations.

Hardware Challenges

Existing commercial or military SDR platforms provide broadband frequency support (DC — 6 GHz) with either fixed or configurable front-ends. However, the majority of the designs are optimized for RF terrestrial communications, and cannot address the challenges of the underwater environment (e.g., acoustic frequencies on the order of kilohertz). In addition, SDR front-ends that can be used directly for underwater communications utilizing the LFTX daughtercard on a USRP platform, which results in operating frequencies in the range of DC to 30 MHz,

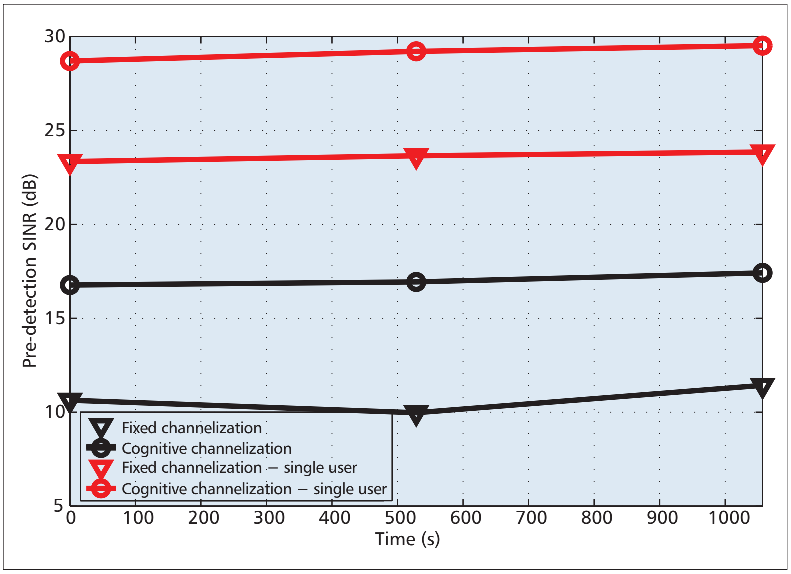


Figure 6. Cognitive channelization is compared to fixed channelization in terms of predetection SINR. The depicted SINR gains consider two different setups: a blind receiver configuration in a multi-user case and in a single-user (N_1–N_3) case.

do not include a power amplifier; thus, underwater communication is limited in terms of range/distance. Furthermore, commercial off-the-shelf (COTS) acoustic transducers are expensive, (their cost is tens of thousands of dollars) due to a proprietary and long process of characterization and analog circuitry optimization. COTS transducers also exhibit application-specific characteristics and low-volume production, and inherently have limited bandwidth capabilities. As a result, experimental prototyping requires careful review of the available COTS hardware components and specifications, and may also require the design of a custom circuit front-end.

Software Portability

One of the major design challenges of SDR or software-defined underwater systems is the development of platform-independent software. Effective proposals need to be able to demonstrate that pertinent software can easily be rebuilt for different SDR platforms with minimum effort/cost. Software interoperability is especially desirable in military systems where standardization is required for both the interconnection between application components, and between application components and system devices. To date, we do not have a complete structured (abstract) methodology/architecture available to implement reconfigurations across all layers of the protocol stack. Such an architecture could benefit future underwater communications by supporting, for example, cognitive functionalities that may lead to high efficiency in spectrum utilization. For instance, software components for static underwater waveform processing could be real-time swapped with cognitive components for adaptive processing to mitigate adverse external conditions (e.g., noise and multiple-access interference).

Computational Capacity

Future wideband high-bit-rate underwater communications require significant computational resources. It is well understood that data rate capabilities are coupled with the front-end design of the software-defined underwater network nodes (e.g., acoustic transducers and power amplifiers) and may vary according to the external environment as well as the role of the underwater software-defined unit in the network (e.g., surface station, AUV, etc.), which poses different weight, cost, and power consumption constraints. Similar to wired computer networks, software-defined underwater acoustic networks may benefit from a separation between data processing and control functionalities. As a result, future software-defined modem architectures could be able, for example, to distribute data processing between hardware and software and exploit parallelism for computation-intensive data processing components (e.g., finite impulse response filters, fast Fourier transform modules).

Energy Efficiency

Energy consumption of submerged units is a significant design challenge for future reconfigurable underwater modems. It is evident that algorithmic implementation on architectures based on application-specific integrated circuits (ASICs) or DSP-on-chip trade reconfigurability for computational and power efficiency. Recently, a new class of technologies such as programmable logic and IP (FPGAs), in-line vector instructions (ARM NEON), and vector execution units (modern GPUs) are breaking the boundaries of device functionalities and aim at creating a multicore blend of functions, as seen, for example, in several new products including TI KeyStone, Xilinx Zynq, and NVIDIA Tegra K1. Multicore architectures require software abstractions to control reconfiguration of tailored functional units or general processing components. For example, the work in [15] implements matching pursuit algorithms for underwater channel estimation in FPGAs (reconfigurable hardware platforms) and provides 210× and 52× reduction in energy consumption over the microcontroller and DSP implementations, respectively.

Real-Time Reconfigurable Processing

While considerable research work for the next-generation processors aim at increasing cycle clocks, mutli-core architecture proposals (i.e., single-instruction multiple-data and multiple-instruction multiple-data) allow a higher average number of instructions to be processed per clock cycle, and offer high performance and reconfiguration by exploiting parallelism. However, even though multi-core DSPs offer optimized features for digital signal processing, we still need to study the trade-off between high performance and low power consumption in software-defined underwater acoustic networks. More specifically, reconfiguration timing constraints can be significantly relaxed due to larger channel coherence times in underwater communication (on the order of seconds) compared to wireless LAN communications (on the order of milliseconds). Relaxed timing constraints imply low power consumption, which is beneficial to battery operated underwater deployments.

Conclusions

We have discussed SDR technology principles for UANs that need reconfigurable and intelligent network devices. Accordingly, we have proposed an SDAM prototype, based on COTS components, and have experimentally evaluated the proposed prototype in both indoor (water test tank) and outdoor (lake) environments. Then we have demonstrated, under three experimental scenarios, the real-time reconfigurable capabilities of the proposed SDAM prototype and highlighted its favorable characteristics toward spectrally efficient cognitive underwater networks and high data rate underwater acoustic links. Finally, we have reviewed open research implementation challenges for future next-generation software-defined UANs and presented an overview of existing SDR-based architectures for UW-A communications.

Acknowledgments

This work was supported in part by the National Science Foundation under grants CNS-1422874 and CNS-1126357.

References

[1] B. Tomasi, "JANUS: The Genesis, Propagation and Use of an Underwater Standard," *Euro. Conf. Underwater Acoustics*, Istanbul, Turkey, July 2010.

[2] R. Petroccia and D. Spaccini, "Comparing the Sunset and Desert Frameworks for In Field Experiments in Underwater Acoustic Networks," *OCEANS — Bergen, 2013 MTS/IEEE*, June 2013, pp. 1–10.

[3] C. Petrioli *et al.*, "The Sunrise Gate: Accessing the Sunrise Federation of Facilities to Test Solutions for the Internet of Underwater Things," *Underwater Commun. and Networking 2014*, Sept. 2014, pp. 1–4.

[4] J.-H. Cui *et al.*, "Ocean-TUNE: A Community Ocean Testbed for Underwater Wireless Networks," *ACM Int'l. Conf. Underwater Networks and Systems*, Los Angeles, CA, Nov. 2012, pp. 1–2.

[5] Z. Peng *et al.*, "Aqua-net: An Underwater Sensor Network Architecture: Design, Implementation, and Initial Testing," *OCEANS 2009, MTS/IEEE Biloxi — Marine Technology for Our Future: Global and Local Challenges*, Oct. 2009, pp. 1–8.

[6] M. Chitre, R. Bhatnagar, and W.-S. Soh, "Unetstack: An Agent-Based Software Stack and Simulator for Underwater Networks," *Oceans — St. John's 2014*, Sept 2014, pp. 1–10.

[7] J. Potter *et al.*, "Software Defined Open Architecture Modem Development at CMRE," *Underwater Commun. and Networking 2014*, Sept 2014, pp. 1–4.

[8] D. Torres *et al.*, "Software-Defined Underwater Acoustic Networking Platform," *Proc. 4th ACM Int'l. Wksp. UnderWater Networks*, ser. WUWNet '09, Nov 2009, pp. 7:1–7:8.

[9] E. M. Sözer and M. Stojanovic, "Reconfigurable Acoustic Modem for Underwater Sensor Networks," *Proc. 1st ACM Int'l. Wksp. Underwater Networks*, ser. WUWNet '06, Sept 2006, pp. 101–04.

[10] N. Nowsheen, C. Benson, and M. Frater, "A High Data-Rate, Software-Defined Underwater Acoustic Modem," *OCEANS 2010*, Sept 2010, pp. 1–5.

[11] M. Chitre, I. Topor, and T. Koay, "The UNET-2 Modem — An Extensible Tool for Underwater Networking Research," *OCEANS, 2012 — Yeosu*, May 2012, pp. 1–7.

[12] E. Demirors *et al.*, "Design of a Software-Defined Underwater Acoustic Modem with Real-Time Physical Layer Adaptation Capabilities," *Proc. Int'l. Conf. Underwater Networks & Systems*, ser. WUWNET '14, 2014, pp. 25:1–25:8.

[13] G. Sklivanitis *et al.*, "Receiver Configuration and Testbed Development for Underwater Cognitive Channelization," *2014 48th Asilomar Conf. Signals, Systems and Computers*, Nov 2014, pp. 1594–98.

[14] Y. Sun and T. Melodia, "The Internet Underwater: An IP-Compatible Protocol Stack for Commercial Undersea Modems," *Proc. 8th ACM Int'l. Conf. Underwater Networks and Systems*, ser. WUWNet '13, 2013, pp. 37:1–37:8.

[15] B. Benson *et al.*, "Energy Benefits of Reconfigurable Hardware for Use in Underwater Snesor Nets," *IEEE Int'l. Symp. Parallel Distrib. Processing 2009,* May 2009, pp. 1–7.

Biographies

Emrecan Demirors [S'11] (edemirors@ece.neu.edu) received his B.Sc. degree in electrical and electronics engineering from the Bilkent University, Ankara, Turkey, in 2009 and his M.Sc. degree in electrical engineering from the State University of New York at Buffalo, New York, in 2013. He is currently working toward his Ph.D. degree in electrical engineering at the Department of Electrical and Computer Engineering at Northeastern University, Boston, Massachusetts. His research interests include software-defined wireless communications and networks, cognitive radio networks, and underwater acoustic communications and networks. In 2014, he was the winner of the Nutaq Software-Defined Radio Academic US National Contest.

George Sklivanitis [S'11] (gsklivan@buffalo.edu) received his Diploma in electronic and computer engineering from the Technical University of Crete, Chania, Greece, in 2010. He is currently working toward his Ph.D. degree in electrical engineering at the State University of New York at Buffalo. His research interests span the areas of signal processing, software-defined wireless communications and networking, cognitive radio, and underwater acoustic communications. In 2014 he was the winner of the Nutaq Software-Defined Radio Academic US National Contest, and in 2015 he received the Graduate Student Excellence in Teaching award from University at Buffalo.

Tommaso Melodia [M'07] (melodia@ece.neu.edu) received his Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, in 2007. He is an associate professor with the Department of Electrical and Computer Engineering, Northeastern University, Boston, Massachusetts. He serves on the Editorial Boards of *IEEE Transactions on Mobile Computing*, *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Multimedia*, and *Computer Networks*. His research interests are in underwater sensor networks, cognitive and software-defined networks, and intra-body medical networks of implantable devices.

Stella N. Batalama [S'91, M'94] (batalama@buffalo.edu) received her Diploma degree in computer engineering and science from the University of Patras, Greece, in 1989 and her Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, in 1994. In 1995 she joined the Department of Electrical Engineering, State University of New York at Buffalo, where she is presently a professor. From 2009 to 2011, she served as the associate dean for research of the School of Engineering and Applied Sciences, and since 2010 she has served as chair of the Electrical Engineering Department. During the summers of 1997–2002 she was visiting faculty at the U.S. Air Force Research Laboratory (AFRL), Rome, New York. From August 2003 to July 2004 she served as acting director of the AFRL Center for Integrated Transmission and Exploitation (CITE). Her research interests include small-sample-support adaptive filtering and receiver design, cooperative communications, cognitive networks, underwater communications, covert communications, steganography, compressive sampling, adaptive multiuser detection, robust spread-spectrum communications, and supervised and unsupervised optimization. She was an Associate Editor for *IEEE Communications Letters* (2000–2005) and *IEEE Transactions on Communications* (2002–2008).

Dimitris A. Pados [M'95, SM'15] (pados@buffalo.edu) received his Diploma degree in computer science and engineering from the University of Patras and his Ph.D. degree in electrical engineering from the University of Virginia. Since August 1997, he has been with the Department of Electrical Engineering, State University of New York at Buffalo, where he currently holds the title of Clifford C. Furnas Professor of Electrical Engineering. His research interests are in the general areas of communication theory and systems, and adaptive signal processing with applications to interference channels and signal waveform design, secure wireless communications, and cognitive acoustic-to-RF modems and networks.

Reconfiguration timing constraints can be significantly relaxed due to larger channel coherence times in underwater communication (on the order of seconds) compared to wireless LAN communications (on the order of milliseconds). Relaxed timing constraints imply low power consumption, which is beneficial to battery operated underwater deployments.

Routing Protocols for Underwater Wireless Sensor Networks

Guangjie Han, Jinfang Jiang, Na Bao, Liangtian Wan, and Mohsen Guizani

Guangjie Han, Jinfang Jiang, Na Bao and Liangtian Wan are with the Hohai University, Changzhou.

Mohsen Guizani is with the University of Idaho.

ABSTRACT

Recently, underwater wireless sensor networks (UWSNs) have emerged as a promising networking technique for various underwater applications. An energy efficient routing protocol plays a vital role in data transmission and practical applications. However, due to the specific characteristics of UWSNs, such as dynamic structure, narrow bandwidth, rapid energy consumption, and high latency, it is difficult to build routing protocols for UWSNs. In this article we focus on surveying existing routing protocols in UWSNs. First, we classify existing routing protocols into two categories based on a route decision maker. Then the performance of existing routing protocols is compared in detail. Furthermore, future research issues of routing protocols in UWSNs are carefully analyzed.

INTRODUCTION

Over the past few years, there has been a rapidly growing amount of research on UWSNs owing to their wide applications in many underwater scenarios, e.g. marine climate observation, pollution tracking, assisted navigation, tactical underwater surveillance, disaster prevention, etc. Almost all the applications require underwater sensor nodes to effectively provide accurate sensed data. However, due to the complex underwater environment, how to quickly and effectively transmit the collected data to the sink node on the ocean surface is a very challenging research problem. Actually, there are many routing protocols that have been proposed for terrestrial wireless sensor networks (TWSNs). However, these are not suitable for UWSNs, mainly because of specific characteristics of UWSNs, such as dynamic structure, narrow bandwidth, rapid energy consumption, and high transmission latency. Therefore, many novel routing protocols have been specifically proposed for UWSNs. In this article we survey some of these routing protocols and provide a comparison table of the most important ones.

Usually, sensor nodes in UWSNs are mobile and freely float with the ocean current. Therefore, the established routing paths need regular updating and maintenance, which obviously introduces high energy consumption. However, it is generally known that all sensor nodes are energy limited, and hence it is challenging to build energy efficient routing protocols for UWSNs. In routing protocols for UWSNs, the route decision maker can be classified into two categories: sender and receiver. In the sender-based routing protocols, the sender node chooses its next hop node by itself, while in the receiver-based routing protocols, the next hop node is selected by the neighbor nodes of the sender. Comparatively, the receiver-based routing protocols are much more energy efficient than the sender-based protocols, as less communication overheads are required. Therefore, in this article the routing protocols are surveyed based on two categories, sender-based and receiver-based, as shown in Fig. 1.

The remainder of this article is organized as follows: the construction and design of the protocols are presented in detail. The performance of these protocols is compared in terms of energy efficiency, transmission delay, network throughput, etc. Finally, open issues and conclusions are drawn in the last section.

EXISTING UNDERWATER ROUTING PROTOCOLS

RECEIVER-BASED UNDERWATER ROUTING PROTOCOLS

In this section we classify protocols into three categories:

- Energy-based routing [1, 2].
- Geographic information-based routing [3–5].
- Hybrid routing protocols [6].

Each routing protocol is carefully analyzed.

Energy-Based Routing: In [1], an Energy optimized Path Unaware Layered Routing Protocol (E-PULRP) is proposed for UWSNs, where a sink node is assumed at the center of the network. E-PULRP consists of two phases: a layering phase and a communication phase. In the layering phase, a layered architecture is constructed based on the different energy levels of sensor nodes. The sensor nodes within a layer have the same hop count to the sink node. During the communication phase, the intermediate

0163-6804/15/$25.00 © 2015 IEEE

relay nodes are chosen based on their distances from the sink node. The sensor node that is closer to the sink and maximally away from the source is more likely to be selected as a next-hop node. In order to prolong the network lifetime, the sensor nodes that are not selected can go to sleep to save their energy. In addition, the layered network structure can well balance the energy consumption of the whole network. However, the mobility of sensor nodes is not considered in E-PULRP, therefore it is not suitable for real UWSNs' applications.

In [2] an energy-efficient and lifetime-aware routing protocol named QELAR is proposed based on reinforcement learning techniques. QELAR employs a Q-learning algorithm to deliver packets with maximum reward, i.e. minimum cost. In the Q-learning algorithm, an action-value function (Q-value), which gives the expected reward that can be received by taking an action in a given state, is adopted to calculate the reward function. Based on the reward function, the sensor node with the highest reward can be selected as an adequate packets' forwarder. Applying the Q-learning technique in the QELAR routing protocol can balance the workload among sensor nodes, reduce the networking overhead, and prolong the network lifetime. In addition, QELAR is specifically suitable for mobile UWSNs. However, each sensor node needs to record lots of information for the Q-value calculation, which introduces high demands on sensor nodes' storage space. Therefore, QELAR cannot be applied on a large scale of UWSNs.

Geographic Information-Based Routing: In order to handle the problem of node mobility, a Hop-by-Hop Dynamic Addressing Based (H2-DAB) routing protocol is proposed in [3], where each sensor node is assigned with a routable address that consists of two parts: node ID and hop ID. Based on the routable address, the H2-DAB routing protocol is available when even some sensor nodes come and leave the network along with the ocean current movement, since each sensor node can obtain its address dynamically, without the involvement of any static configuration. In addition, H2-DAB is a light and energy efficient routing protocol as it does not require any location information or maintain complex routing tables. However, how to update the routable address in a timely manner, which directly affects the effective transmission of information, is an urgent problem that needs further research.

In [4] a Depth-Based Routing (DBR) protocol is proposed for UWSNs. In DBR, each sensor node makes its own decision on packet forwarding based on its depth and the depth of the previous sender. As shown in Fig. 2, node S is a sender, and all the neighbor nodes, e.g. n_1, n_2, and n_3, will receive its packets. However, only n_1 and n_2 are chosen as candidate forwarding nodes since they are closer to the sink node on the water surface. In addition, node n_1 is preferred to forward the packets as compared to node n_2. The forwarding of node n_2 is prevented if it receives the packet from n_1 before its own scheduled sending time for the packet. DBR can handle network dynamics efficiently without requiring full-dimensional location information of sensor nodes. However, if there are many neighbor nodes in the network, it is very likely that multiple nodes forward the same packets and a sensor node may receive the same packet multiple times, which results in a high volume of packet collisions and high transmission delay and energy consumption. Therefore, in [5] a Delay-Sensitive Depth-Based Routing (DSDBR) protocol is proposed, which employs holding time to minimize end-to-end delay. Holding time is the residence time that the received packets can be kept on receiver nodes. The packets will be discarded beyond the holding time, which ultimately limits packet transmission delay.

Hybrid Routing Protocol: In [6] a novel routing protocol, called Link-state based Adaptive Feedback Routing (LAFR), is proposed for UWSNs. Different from the above mentioned routing protocols, the asymmetry of underwater acoustic communication links is taken into account in LAFR. Based on the recorded IDs in the downstream node table of a sensor node, the communication links can be detected as symmetric or asymmetric, which plays an important role in the process of route establishment. A sensor node can directly transmit packets without routing inquiry if the sensor node and its neighbor node are connected by a symmetric link, which can save a large amount of energy consumption for routing inquiry. In addition, a time-based priority forwarding mechanism is proposed to prevent flooding of routing request packets, and a

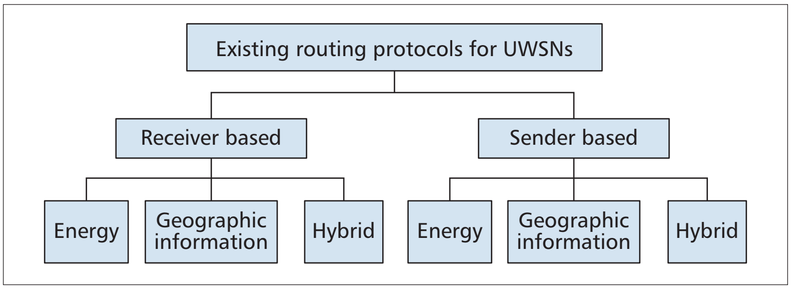


Figure 1. Classification of existing routing protocols in UWSNs.

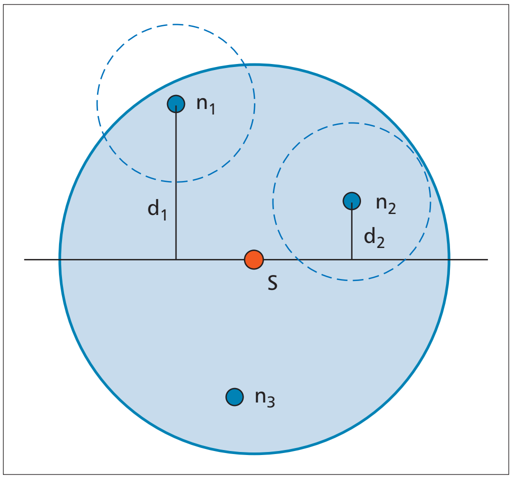


Figure 2. The selection of forwarding nodes in DBR.

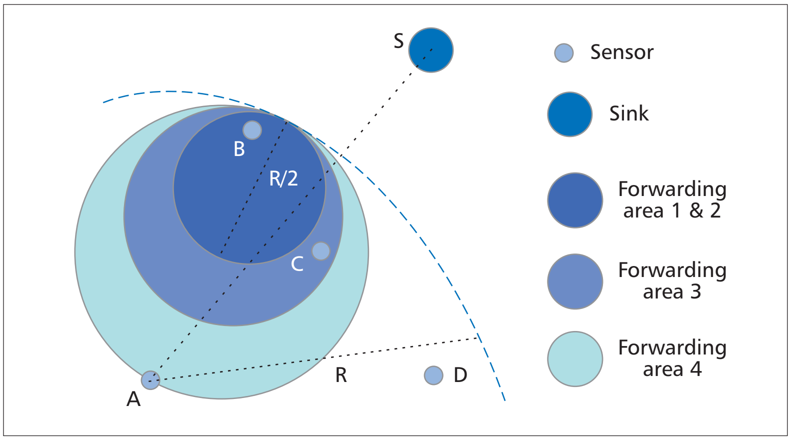


Figure 3. The forwarding area in ARP.

credit-based routing table update mechanism is adopted to avoid energy dissipation caused by frequent update of the routing table. However, the mobility of sensor nodes is not considered in LAFR, which directly impacts the asymmetry of underwater communication link.

Sender-Based Underwater Routing Protocols

In this section the routing protocols are also classified into three categories, as mentioned above.

Energy-based Routing: In [7] a novel Adaptive Routing Protocol (ARP) is proposed based on the types of messages and application requirements. The data packets are assigned with different delivery priority based on the packet characteristics, such as emergency level and age (i.e. residence time in the network), and the node status, such as residual battery and the density of the neighborhood. As shown in Fig. 3, a forwarding area is introduced. A higher delivery priority corresponds to a larger forwarding area. Clearly, "more important" packets can be delivered within shorter delays. ARP achieves a good trade-off among delivery ratio and end-to-end delay. However, the diameter of the forwarding area is still in need of further study.

In [8] a Mobicast Routing Protocol (MRP) is proposed for UWSNs. In MRP, a mobile sink or an autonomous underwater vehicle (AUV) is used as a data collector to collect data from the underwater environment. All the sensor nodes around the data collector form a 3-D geographic zone, which is called as a 3-D zone of reference (3-D ZOR). To save energy consumption, only the sensor nodes in a 3-D ZOR are required to wake up to transmit data to the data collector, while the other sensor nodes can go to sleep. MRP consists of two phases: collecting data within a 3-D ZOR, and waking up the sensor nodes in the next 3-D ZOR while trying to avoid topology holes caused by ocean currents. In order to solve the problem of the topology hole, an 'apple slice' technique is used to surround the hole. To the best of our knowledge, MRP is the first study to develop an underwater routing protocol to collect data from sensor nodes in sleep mode and avoid the routing holes caused by a dynamic topology.

Geographic Information-Based Routing: In [9] a Relative Distance Based Forwarding (RDBF) routing protocol is proposed for UWSNs, which is a location based routing protocol. In REBF, a fitness factor is defined based on the distance from a sink node to judge whether a sensor node is appropriate for packet forwarding.

The neighbor nodes that are closer to the sink are more likely to be selected as a next hop forwarder. In order to limit the number of forwarders and avoid packet collision, another parameter β is introduced, that is, if a node hears the β same packets, it will simply discard the reserved packet. Also, the energy balance issue is considered in REBF. A residual energy threshold is defined, and if a sensor node's residual energy level is below this threshold, it will stop forwarding any packet. It is obvious that only some of the sensor nodes participate in packet transmission, which significantly reduces energy consumption. In addition, using the selected and more appropriate sensor nodes to transmit data improves transmission efficiency and limits end-to-end transmission delay. However, REBF requires that each node accurately obtain its position information, which is hard to guarantee since it is difficult to obtain accurate location information in a dynamic underwater environment in a timely manner.

In [10] a novel Geocast technique is proposed with a hole detection mechanism for UWSNs. The proposed model is named Routing and Multicast Tree based Geocasting (RMTG), and consists of the following six parts:

- Neighbor table formation, which is used to provide all the sensor nodes with their neighbor nodes' location information.
- Route discovery, where the neighbor node that is closest to the destination is chosen as the next hop node.
- Route maintenance, which is used to address the issue that no neighbor node can be found as the next hop, and solve the link break problem.
- Multicast tree formation, which is responsible for constructing a multicast shortest path to transmit data packets.
- Hole detection.
- Boundary routing around the geocast region.

RMTG provides an efficient routing protocol for UWSNs in terms of node memory saving, mobility handling, and an acceptable end to end latency. However, the greedy forwarding method that is used is not sufficiently energy efficient, which may cause a large amount of packet blocking. In addition and similar to the RDBF protocol discussed in [9], all the sensor nodes in RMTG are also required to know their geographic coordinates.

In [11] a Diagonal and Vertical Routing Protocol (DVRP) is proposed for UWSNs. Since horizontal communication between the sensor nodes on the same depth levels is always useless and wastes a large amount of communication overhead, DVRP adopts the triangular inequality theorem to avoid horizontal communication.

As shown in Fig. 4, when transmitting packets, source node S will choose neighbor node N instead of M as the forwarder node. However, accurate 3-D coordinate information is required, which is challenging in underwater environments. In [12] a depth-based routing protocol, named the Void-Aware Pressure Routing (VAPR) protocol, is proposed for UWSNs, which uses hop count and depth information to decide next hop nodes. Comparatively, obtaining depth information is much easier than coordinate information in UWSNs. VAPR consists of two phases: enhanced beaconing and opportunistic directional data forwarding. In the first phase, each sensor node is initialed with beacon information, e.g. its depth, hop count, data forwarding direction, etc. In the second phase, sensor nodes forward data packets solely based on the data forwarding direction and next-hop data forwarding direction, which can ensure that the packets can be forwarded upward to the water surface. Based on periodic beacon signals, VAPR is robust to network dynamics such as node mobility and failure.

Hybrid Routing Protocol: In [13] a power-efficient routing protocol (PER) is proposed for UWSNs. PER consists of two phases: the selection of forwarding nodes and the trimming of a forwarding tree. In the first phase, based on the distance, the angle between two neighboring sensor nodes, and the remaining energy of the sensor nodes, two candidate forwarding nodes are selected. In the second phase, the trimming of a forwarding tree is conducted based on the number of duplicated packets received by each sensor node, which can be used to constrain excessive packet forwarding and prevent unnecessary power consumption. However, PER needs the information of all the neighbor nodes to select forwarding nodes, which introduces higher memory utilization and more communication overheads.

In [14] a Channel-Aware Routing (CARP) Protocol is proposed for UWSNs, in which both link quality and hop count are considered for routing selection. The link quality of a sensor node is evaluated based on its past successful transmissions to its neighbor nodes. Generally, the communication link with the least hop count and the best link quality is selected for packet transmission. To the best of our knowledge, CARP is the first underwater routing protocol that takes the link status into account. However, due to the dynamic characteristic of UWSNs, the link quality of the same transmission changes frequently. It is hard to maintain and update in a timely manner the link status information during the process of data transmission.

In [15] the DBR protocol [4] is extended to an Energy Efficient Depth-Based Routing (EEDBR) protocol. In DBR, only the depth information of sensor nodes is used, while in EEDBR, both the depth information and the residual energy of sensor nodes are taken into account to select the next hop node. In addition, DBR is a receiver-based routing protocol, that is, a sender broadcasts its packets to its neighbor nodes, then the neighbor nodes decide whether to forward the received data packets or not. In this case, there are likely large amounts of redundant transmissions due to the lack of information about neighboring nodes. While in EEDBR, based on the neighboring nodes' depths and residual energy information, the sender can select a proper set of neighbor nodes to forward data packets to avoid unnecessary transmissions. Therefore, compared with DBR, EEDBR is much more energy efficient in terms of redundant data transmission. However, in order to efficiently transmit data packets, the neighbor nodes' information, e.g. the depth information and the residual energy level, needs to be periodically updated.

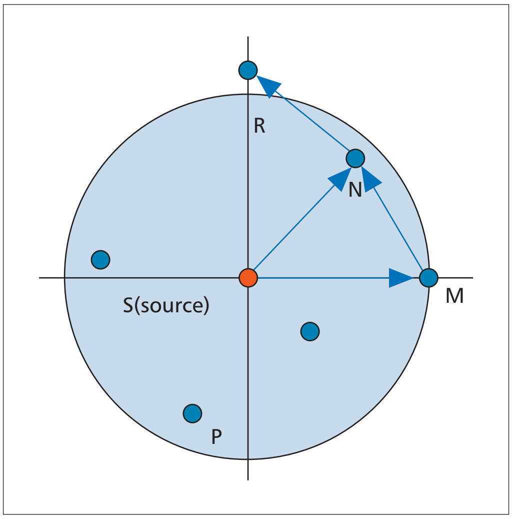


Figure 4. The selection of next hop nodes in DVRP.

In the receiver-based routing protocols, since there is not enough information exchange between sensor nodes, the sensor nodes cannot obtain their neighborhood information in time. This ultimately can cause that packets cannot be transmitted with the highest efficiency.

Discussion and Open Research Issues

As analyzed earlier, in the sender-based routing protocols, the sender node chooses its next-hop nodes by itself. Once there are packets waiting for transmission, the sender first needs to collect its neighbor nodes' information and then inform the chosen nodes. Therefore, many information communications are required between the sender and its neighbor nodes, which introduces high communication overhead. In particular, the senders consume relatively higher energy. In contrast, in the receiver-based routing protocols, a receiver node can decide whether it can be a forwarder node by itself. The redundant communications between the sender and its neighbor nodes are not needed. Generally speaking, the receiver-based routing protocols require less communication overhead, and are much more energy efficient than the sender-based protocols. However, in the receiver-based routing protocols, since there is not enough information exchange between sensor nodes, the sensor nodes cannot obtain their neighborhood information in time. This ultimately can lead to packets not being transmitted with the highest efficiency. The packet transmission efficiency is measured by energy efficiency, bounded latency, load balancing, hole bypassing, and dynamic robustness, as shown in Table 1.

Routing protocol	Energy efficiency	Bounded latency	Multi-path	Load balancing	Loop free	Hole bypassing	Dynamic robustness	Geographic information		Communication overhead	Time complexity
								Location	Depth		
E-PULRP[1]	✓				✓					Low	$O(n)$
QELAR [2]	✓				✓					Low	$O(m_1)$
H2-DAB [3]	✓				✓		✓		✓	Low	$O(n^2)$
DBR [4]	✓		✓		✓				✓	Low	$O(m_1)$
DSDBR [5]	✓	✓	✓		✓				✓	Low	$O(m_1)$
LAFR [6]	✓		✓		✓					Low	$O(n \times m_1)$
ARP [7]	✓	✓	✓		✓			✓		High	$O(m_1)$
MRP [8]			✓		✓	✓	✓			Low	$O(m_2)$
RDBF [9]	✓		✓	✓	✓			✓		High	$O(n \times m_1)$
RMTG [10]			✓		✓	✓	✓	✓		High	$O(n^2)$
DVRP [11]					✓			✓		Median	$O(m_1)$
VAPR [12]			✓		✓		✓		✓	Median	$O(m_1)$
PER [13]	✓		✓		✓					High	$O(n^2)$
CARP [14]					✓					High	$O(m_1)$
EEDBR [15]	✓		✓		✓					High	$O(m_1)$

Table 1. Comparison of the underwater routing protocols.

It is generally known that the sensor nodes in UWSNs are all resource constrained and energy limited, therefore in the design of routing protocols, energy efficiency has been a significant design objective. An energy efficient routing protocol means that it consumes the least possible energy for packet transmission. As compared in Table 1, most routing protocols can be energy efficient by adopting the duty-cycle mechanism, e.g. E-PULRP [1] and MRP [8], where only some of the sensor nodes are required to participate in packet transmission, choosing the minimum energy cost route, e.g. QELAR [2], or using the link state information, e.g. LAFR [6]. In addition, in order to save energy consumption and prolong network lifetime, the traffic load in the network should be balanced to avoid some sensor nodes becoming exhausted too early. However, in current underwater routing protocols, few of them take load balancing into account.

As a data centric network, the sensed data in UWSNs should be accurately transmitted to the sink node in a bounded latency since the propagation delay of acoustic signals is relatively high while the sensed data always has a certain time-validity. Especially for some emergency applications, such as ocean pollution surveillance, a packet that reports pollution should be delivered to the base station as quickly as possible. However, not all the sensed data are delay-sensitive. For example, the packets that record normal environmental parameters, e.g. temperature and salinity, can tolerate a longer transmission delay than the urgent packets. Therefore, the underwater routing protocols should be designed to handle different environmental requirements adaptively. In the above surveyed routing protocols, only a few routing protocols take data transmission delay into consideration, and only in ARP [7] is the transmission delay carefully analyzed, where different delivery priorities are assigned to ordinary packets and emergency packets. However, the detail of delivery priority needs further study.

Efficient data transmission needs not only less energy consumption and end-to-end delay, but also higher delivery ratio and delivery reliability. As shown in Table 1, most current routing protocols use multi-path routings to ensure higher delivery ratio and the reliability of data. The performance of multi-path and single-path routings is compared in Fig. 5. It obviously shows that multi-path routings, e.g. DBR [4], MRP [8], and EEDBR [15], outperform single-path routings, e.g. QELAR [2] and H2-DAB [3], in terms of successful delivery ratio. However, when the number of sensor nodes reaches 700, the delivery ratios of multi-path routings, e.g. DBR [4] and EEDBR [15], abruptly drop, since the number of collisions also increases with the increase in the number of nodes. In addition, multi-path routings consume much more energy than a single path routing. Therefore, how to achieve a good trade-off among delivery ratio, average end-to-end delay, and energy consumption is still a significant problem that needs urgent resolution. Furthermore, a well designed routing protocol should be loop free. A looped path is useless for data transmission. In this case, the packets cannot be transmitted to the sink node, and a large amount of energy will be wasted.

From the comparison we can also find that most underwater routing protocols take advantage of depth information instead of 3D coordinates to establish routes, since in the underwater environment, obtaining depth information is much easier and less costly than obtaining 3D coordinates. Depth information can be easily obtained by pressure sensors, while the coordinates of underwater sensor nodes are calculated based on complex localization algorithms. Due to node mobility and sparse deployment, how to design an accurate localization algorithm is also an unresolved issue. Also, for an efficient routing protocol, hole bypassing is essential. A route hole in underwater environments is inevitable because of node movement or disconnected links, which may cause sensor nodes spending extra delay or additional energy. In addition, the inevitable route hole directly reduces the data delivery ratio. However, most current routing protocols do not consider how to overcome the hole bypassing problem.

Last but not least, the time complexities of the surveyed routing algorithms are compared, where n is the number of sensor nodes in the network, m_1 is the maximum number of a sender's neighbor nodes, and m_2 is the maximum number of neighbor nodes in the 3-D ZOR.

Based on the survey and the comparison of existing underwater routing protocols, we can conclude that most current research is only suitable for small and static UWSNs. However, underwater sensor nodes freely float with ocean current and most UWSNs are mobile. Thus, a novel dynamic routing protocol is needed for dynamic and large-scale UWSNs. Furthermore, current routing protocols do not consider the security problem, even though UWSNs are always deployed in unattended and even hostile environments. There are many malicious attacks against data transmission in UWSNs. Therefore, the secure routing protocol can be viewed as another interesting research issue that should be further studied.

Conclusion

Routing for UWSNs has been one of the most important issues in underwater applications. Over the past few years, many routing protocols have been proposed for UWSNs based on the unique characteristics of UWSNs. In this article we present a detailed survey of underwater routing protocols. Each routing protocol is carefully analyzed, and its advantages, disadvantages, and performance issues are highlighted. In addition, we compare the protocols in terms of energy efficiency, path latency, multi-path capability, reliability, dynamic robustness, hole bypassing, etc.

From the comparison it is found that there are still many research challenges not yet solved. Therefore, further work should be performed in order to investigate the available solutions in greater detail, and propose new approaches to achieve a better routing protocol for UWSNs.

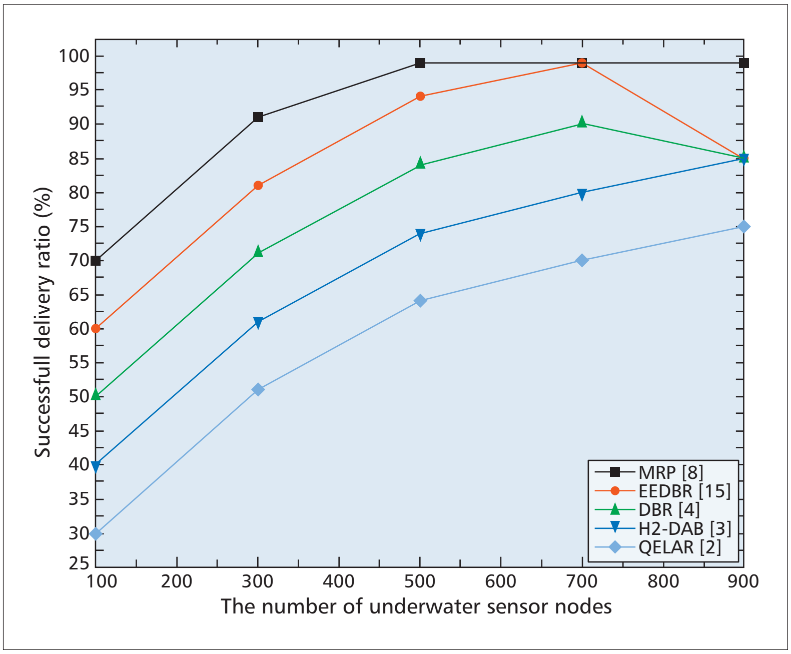


Figure 5. Different data delivery ratios of multi-path and single-path routings.

Acknowledgment

The work is supported by the Jiangsu Province Ordinary University Graduate Innovation Project, No. CXZZ13_0240, the Qing Lan Project, the National Natural Science Foundation of China under Grant 61572172, and the Natural Science Foundation of Jiangsu Province of China, No. BK20131137 and No. BK20140248.

References

[1] S. Gopi *et al.*, "E-PULRP: Energy Optimized Path Unaware Layered Routing Protocol for Underwater Sensor Networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, 2010, pp. 3391–401.

[2] T. Hu and Y. Fei, "QELAR: A Machine-Learning-Based Adaptive Routing Protocol for Energy-Efficient and Lifetime-Extended Underwater Sensor Networks," *IEEE Trans. Mobile Comp.*, vol. 9, no. 6, 2010, pp. 796–809.

[3] M. Ayaz *et al.*, "An Efficient Dynamic Addressing Based Routing Protocol for Underwater Wireless Sensor Networks," *Comput. Commun.*, vol. 35, no. 4, 2012, pp. 475–86.

[4] H. Yan *et al.*, "DBR: Depth-Based Routing for Underwater Sensor Networks," *Proc. Net.*, vol. 4982, 2008, pp. 72–86.

[5] M. R. Jafri *et al.*, "Towards Delay-Sensitive Routing in Underwater Wireless Sensor Networks," *5th Conf. EUSPN*, vol. 37, 2014, pp. 228–35.

[6] S. Zhang *et al.*, "A Link-State Based Adaptive Feedback Routing for Underwater Acoustic Sensor Networks," *IEEE Sensors J.*, vol. 13, no. 11, 2013, pp. 4402–12.

[7] Z. Guo *et al.*, "Adaptive Routing in Underwater Delay/Disruption Tolerant Sensor Networks," *5th Annu. Conf. on WONS*, 23-25 Jan. 2008, pp. 31–39.

[8] Y. S. Chen and Y. W. Lin, "Mobicast Routing Protocol for Underwater Sensor Networks," *IEEE Sensors J.*, vol. 13, no. 2, 2013, pp. 737–49.

[9] Z. Li *et al.*, "Relative Distance Based Forwarding Protocol for Underwater Wireless Networks," *Int. J. Distrib. Sensor Networks*, 2014, pp. 1–11.

[10] S. K. Dhurandher *et al.*, "A Novel Geocast Technique with Hole Detection in Underwater Sensor Networks," AICCSA, May 2010, pp. 1–8.

[11] T. Ali *et al.*, "Diagonal and Vertical Routing Protocol for Underwater Wireless Sensor Network," *Procedia Soc. Behav. Sci.*, vol. 129, May 2014, pp. 372–79.

[12] Y. Noh *et al.*, "VAPR: Void-Aware Pressure Routing for Underwater Sensor Networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 5, May 2013, pp. 895–908.

[13] C. J. Huanga *et al.*, "A Power-Efficient Routing Protocol for Underwater Wireless Sensor Networks," *Appl. Soft Comp.*, vol. 11, no. 2, 2011, pp. 2348–55.

[14] S. Basagni *et al.*, "Channel-Aware Routing for Underwater Wireless Networks," *OCEANS*, May 2012, pp. 1–9.

[15] A. Wahid and D. Kim, "An Energy Efficient Localization-Free Routing Protocol for Underwater Wireless Sensor Networks," *Int'l. J. Distrib. Sensor Networks*, 2012, pp. 1–11.

BIOGRAPHIES

GUANGJIE HAN [S'01, M'05] (hanguangjie@gmail.com) is a professor in the Department of Information & Communication Systems at Hohai University, China. He finished his work as a postdoctoral researcher with the Department of Computer Science at Chonnam National University, Korea, in 2008. From October 2010 to 2011 he was a visiting research scholar with Osaka University, Suita, Japan. He has served as an editor of IJAHUC, KSII, and JIT. His current research interests include sensor networks, green computing, cloud computing, and mobile computing.

JINFANG JIANG (jiangjinfang1989@gmail.com) is pursuing her Ph.D. degree from the Department of Information & Communication Systems at Hohai University. She received her B.S. degree in information and communication engineering from Hohai University, China in 2009. Her current research interests are security, trust management, routing, and localization for underwater acoustic sensor networks.

NA BAO (baon1401@163.com) is pursuing a master degree from the Department of Information & Communication Engineering at Hohai University, China. She received her B.S. degree in information and communication engineering from Anhui Xinhua University, China in 2011. Her current research interests include underwater acoustic sensor networks and wireless sensor networks.

LIANGTIAN WAN [M'15] (wanliangtian1@163.com) received the B.S. degree and the Ph.D. degree from the College of Information and Communication Engineering from Harbin Engineering University, Harbin, China in 2011 and 2015, respectively. He is a visiting scholar with Hohai University, Changzhou, China. His research interests include sensor networks, signal processing, and compressed sensing.

MOHSEN GUIZANI [S'85, M'89, SM'99, F'09] (mguizani@ieee.org) is a professor and the Department Chair of the Electrical and Computer Engineering Department at the University of Idaho. He served as the Associate Vice President for Graduate Studies at Qatar University, Doha, Qatar. He serves on the editorial boards of six technical journals, and he is the founder and EIC of *Wireless Communications and Mobile Computing Journal*, published by John Wiley. He served as the chair of the IEEE Communications Society Wireless Technical Committee (WTC) and chair of the TAOS Technical Committee. He was an IEEE Computer Society Distinguished Lecturer from 2003 to 2005. He is a senior member of the ACM and a fellow of the IEEE. His research interests include computer networks, wireless communications and mobile computing, and optical networking.

Turbo Equalization for Single-Carrier Underwater Acoustic Communications

Yahong Rosa Zheng, Jingxian Wu, and Chengshan Xiao

ABSTRACT

Recent research in underwater acoustic communications has taken advantage of MIMO technologies to achieve reliable communication with 10–100 times increase of data rate in comparison to traditional systems. The powerful turbo equalization and FEC coding techniques enable both single-carrier modulation and OFDM systems to combat triply selective UWA channels. This article reviews the time-domain and frequency-domain turbo equalizer schemes for MIMO SCM systems. Low-complexity techniques are presented with both turbo linear equalizers and turbo soft decision feedback equalizers in both the time and frequency domains. Although results are shown specifically for UWA channels, these turbo equalizer techniques are also suitable for terrestrial RF communication systems.

INTRODUCTION

Underwater acoustic (UWA) communication, when integrated with underwater robots, unmanned vehicles, or wireless sensor networks, plays a critical role in ocean exploration, underwater infrastructure monitoring, disaster prevention and rescue, and military surveillance. Acoustic wave is the most energy-effective means to achieve medium (1–10 km) and long (10–1000 km) range communications. However, UWA channels exhibit significant technical challenges for achieving high throughput because of their limited bandwidth, slow wave propagation, severe noise and interference, and time-varying dispersion. In contrast to radio frequency communications, common horizontal UWA channels have about 100 kHz bandwidth for 1 km range or 40 kHz for 10 km range due to frequency-dependent attenuation; while an RF wireless channel usually occupies 1–20 MHz bandwidth. Over the last decade, coherent UWA communication has achieved data rates on the order of 10 kb/s [1], and recent application of multiple-input multiple-output (MIMO) technologies in UWA has boosted the rate in excess of 100 kb/s for medium-range communications.

MIMO UWA channels experience severe triply-selective fading. Extremely long delay spread is often on the order of 10 ms, which results in inter-symbol interference (ISI) of 100 taps at 10 ksym/s). The maximum Doppler spread is on the order of 1–10 Hz due to the motion of water as well as transceiver platforms. Severe Doppler spread causes fast time-variation in channel impulse responses (CIRs). Inversely proportional to the maximal Doppler frequency, the coherence time of the UWA fading channels can be as short as 100 ms or 1000 symbols at 10 ksym/s. Simultaneous transmission by multiple elements at the transmitter can boost data rate on one hand, but also causes co-channel interference (CCI) on the other hand. Therefore, powerful MIMO receiver techniques are required to enable reliable reception of high data rate communication.

Turbo equalization is one of the enabling technologies for wideband RF communication and has recently been applied to UWA transceivers [1]. In contrast to conventional equalization and detection schemes, which make hard decisions in one iteration, turbo equalizers exchange the extrinsic information of soft symbols/bits with the soft decision decoders iteratively. Since forward error correction (FEC) coding embeds correlation among the coded bits, and the fading channel introduces correlation among received symbols, turbo equalizers can glean more information about each transmitted bit at each iteration by exploiting the correlation property via adaptive filters.

Due to severe Doppler spread and multipath delay spread in UWA channels, turbo equalizers are often required to achieve satisfactory performance in both single-carrier modulation (SCM) and (multi-carrier) orthogonal frequency-division multiplexing (OFDM) systems [2, 3]. In an SCM system, a symbol is represented by a baseband pulse that occupies the entire bandwidth of the channel and is modulated to a single passband carrier; in an OFDM system, the channel bandwidth is divided into hundreds of frequency bins, and a baseband symbol is modulated onto a narrowband carrier. The block of modulated multi-carriers is then converted to time-domain via inverse fast Fourier transform (IFFT), and a transmitted OFDM symbol occupies the entire time duration of the channel.

Yahong Rosa Zheng and Chengshan Xiao are with the Missouri University of Science and Technology.

Jingxian Wu is with the University of Arkansas, Fayetteville.

0163-6804/15/$25.00 © 2015 IEEE

Accurate channel estimation is critical to the performance of the CE-based Turbo equalizers and is achieved by two modes: the training mode uses the pilot block to estimate the channel iteratively, and the decision-directed mode uses the decisions of the previous and current sub-blocks to estimate and track the MIMO channel.

This article focuses on turbo equalization techniques for SCM systems that utilize continuous transmission rather than block transmission [4–7]. Since UWA channels are quite long in comparison to channel coherence time, continuous transmission eliminates zero padding (ZP) or cyclic prefix (CP) required in block transmission. Therefore, continuous transmission achieves better data efficiency than block transmission. In the receiver, the continuous signal stream may be divided into overlapped sub-blocks for equalization in either the time or frequency domain.

Both time-domain and frequency-domain turbo equalizers are divided into two categories: linear equalizer (LE) and decision feedback equalizer (DFE) [8]. Turbo LEs use only feedforward (FF) filters to filter the received signals, and the filter coefficients are adapted to minimize the mean squared error (MSE) between the equalized symbols and the transmitted ones. Turbo DFEs utilize FF filters to process the received signal and feedback (FB) filters to filter the soft decision symbols of the soft FEC decoders from the previous iteration. The minimum MSE (MMSE) criterion is used for turbo DFE to jointly design the FF and FB filters. Turbo DFE utilizes the soft decisions to reconstruct and remove the residual ISI in the FF filter output, thus achieving better performance than the turbo LE. If designed properly, turbo DFE also reduces catastrophic error propagation often encountered in hard-decision DFE at low signal-to-noise ratio (SNR).

Furthermore, depending on how the adaptation is implemented, turbo equalizers can be divided into channel estimation (CE)-based and direct adaptation (DA)-based ones [9]. The CE-based turbo equalizer explicitly estimates the fading channel coefficients, and then designs the equalizer filters using the channel coefficients and training sequence. The DA-based turbo equalizer adapts the filters directly through the training sequence without estimating the channel coefficients. The CE-based approach requires slightly higher computational complexity but a shorter training sequence and faster convergence than the DA-based approach. In both approaches, the training sequence consists of known pilot symbols and detected symbols from previous sub-blocks.

As DA-based turbo equalizers have been well covered in the literature [1, 9], this article aims to introduce CE-based turbo equalizers that have been applied to MIMO UWA successfully in recent years. For time-domain turbo equalizers, we review the turbo LE [4] and three turbo DFE schemes: the block decision feedback equalizer (BDFE) [5], soft decision feedback equalizer (SDFE) [6], and soft interference cancellation equalizer (SICE) [7]. For frequency-domain turbo equalizers, we first introduce the CP reconstruction technique that enables the low-complexity implementation of FD turbo LE for continuous transmission [7], and then review three FD turbo DFE approaches: frequency domain equalization with time-domain decision feedback (FDE-TDDF), with frequency-domain decision feedback (FDE-FDDF), and time-domain equalizer with FD decision feedback (TDE-FDDF) [8]. We demonstrate the performance of these schemes by experimental results using the data recorded in the SPACE08 ocean experiment.

In a short article like this, it is difficult to exhaustively cover all research progress in the literature, or cover the technical details in great depths. Interested readers are referred to the references listed here and the references therein. It is our hope that this article provides a study guide for learning and analyzing the complicated MIMO turbo equalization schemes.

Single-Carrier MIMO Systems

Assume an $N \times M$ MIMO UWA ccommunication system, as shown in Fig. 1a, where N and M are the numbers of transmit transducers and receive hydrophones, respectively. At the transmitter side, each bitstream $\mathbf{a}_n$ is independently encoded by an FEC encoder. The coded bits $\mathbf{b}_n$ are interleaved and mapped to Q-ary quadrature amplitude modulation (QAM) or phase shift keying (PSK) symbols, and then transmitted after frame construction and carrier modulation. Note that $\mathbf{c}_n$ denotes the interleaved bit vector that maps to symbol s_n, and $\mathbf{s}_k$ denotes the symbol vector at time instant k containing the N transmit symbols.

The received baseband signals are processed iteratively by the turbo MIMO equalizer and decoders. The equalizer outputs the extrinsic log-likelyhood ratios (LLRs) of the interleaved bits for each transmit branch $\lambda_e(\mathbf{c}_n)$, for $n = 1, ..., N$, which becomes the a priori information for the soft-decision (SD) decoder after passing through the de-interleaver. The nth SD decoder outputs the extrinsic information about the coded bits $\lambda_e(\mathbf{b}_n)$, which in turn becomes the a priori LLR of the equalizer after the interleaver. The a priori LLR $\lambda_a(\mathbf{c}_n)$ is used by the turbo equalizer for the next iteration. The iterative process allows the turbo equalizer and the SD decoders to exploit the correlation among the received signals and glean soft information of the transmitted bits back and forth. If the turbo equalizer is designed correctly, the LLRs converge after a few iterations, and the SD decoders make hard decisions $\hat{\mathbf{a}}_n$ on the information bits.

The structure of the transmitted data packet is depicted in Fig. 1b, where the linear frequency modulation (LFM) blocks at the beginning and end of the packet are used for coarse synchronization and Doppler estimation. The m-sequence is used for fine synchronization and initial channel estimation. The data packet is divided into data frames with each frame consisting of a pilot block and a large payload block. In the receiver, the payload block is divided into overlapped sub-blocks for equalization. Accurate channel estimation is critical to the performance of the CE-based turbo equalizers and is achieved in two modes: the training mode uses the pilot block to estimate the channel iteratively, and the decision-directed mode uses the decisions of the previous and current sub-blocks to estimate and track the MIMO channel.

For single-carrier systems, it is unnecessary to insert gaps or cyclic prefixes between the sub-blocks; thus, the data efficiency is higher than

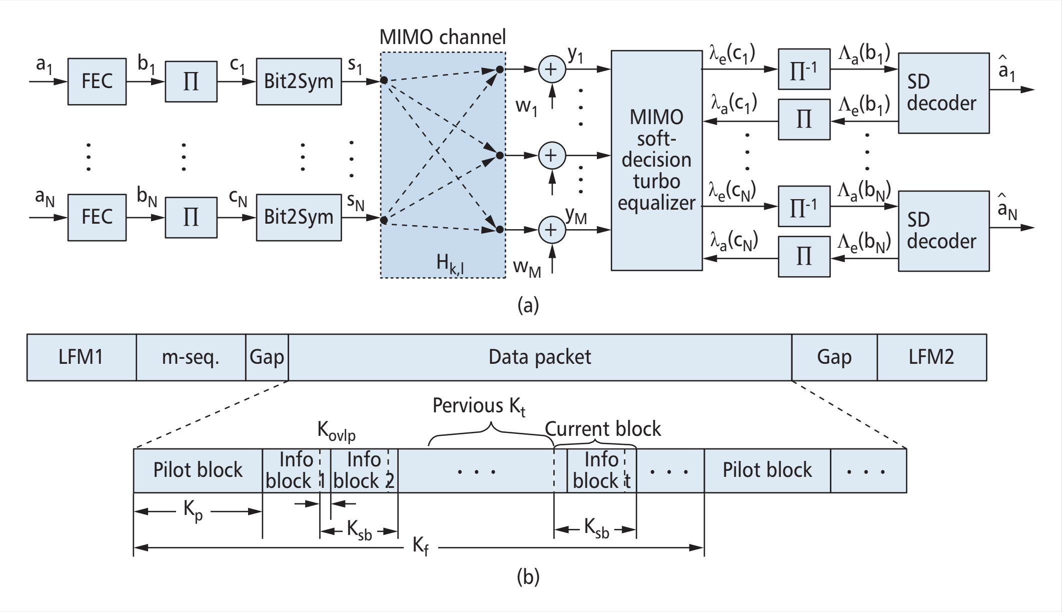


Figure 1. a) Structure of a MIMO communication system with turbo equalizer and maximum *a posteriori* probability (MAP) decoder; b) signaling structure of each transmit branch and subblock format used by the turbo receiver.

The optimal solution often requires the auto-covariance matrix of the equalized/detected soft symbol vector and the cross-covariance of the detected symbol vector and its corresponding transmitted vector. However, the statistics are often difficult to estimate and the receiver has to make reasonable approximations and assumptions.

the block transmission commonly used in OFDM systems. Since no gaps are inserted between subblocks, the inter-block interference (IBI) has to be removed from the received signals, and the IBI-free signals corresponding to each sub-block are fed to the turbo equalizer.

The triply-selective MIMO channel at time instance k is modeled as a set of tapped delay lines (TDLs) in which the lth delay tap is an $M \times N$ matrix $\mathbf{H}_{k,l}$. MIMO UWA channels are often noncausal, as shown in the example in Fig. 2. The channel impulse response (CIR) has a main tap that corresponds to the tap of the strongest gain. The main tap divides the overall CIR into causal and anticausal portions. Denoting the lengths of the overall channel, the anticausal CIR, and the causal CIR as L, L_1, and L_2, respectively, we have $L = L_1 + L_2 + 1$.

The received baseband signal vector $\mathbf{y}_k$ is the sum of the N transmit branches convolved with the channel and the additive Gaussian noise

$$\mathbf{y}_k = \sum_{l=-L_1}^{L_2} \mathbf{H}_{k,l}\, \mathbf{s}_{k-l} + \mathbf{w}_k, \tag{1}$$

where $\mathbf{w}_k$ is the vector of M noise components having zero mean and variance σ_w^2. The long multipath CIRs result in severe ISI at the receiver. The causal CIR causes postcursor ISI, which comes from the past symbols, and the anticausal CIR causes precursor ISI, which is due to the future symbols.

Both time-domain (TD) and frequency-domain (FD) turbo equalizers have been developed for single-carrier UWA MIMO systems. These equalizers are further divided into two types: soft decision linear equalizers (SD-LEs) and soft decision feedback equalizers (SD-DFEs). Based on how the LLRs are calculated in the turbo equalizers, both SD-LEs and SD-DFEs may compute the LLRs using a sequence estimator that computes the *a posteriori* probabilities based on a sequence of equalized symbols, or a symbol estimator that only uses the current soft symbol to compute the extrinsic LLR of the current symbol. The next two subsections review the TD and FD turbo equalizers that have been successfully applied to UWA.

TIME-DOMAIN TURBO EQUALIZATION

To equalize the MIMO channel in the time domain, the receiver has to first align the current cursor with the main tap CIR, and then design adaptive transversal filters to minimize the MSE between the equalized symbols and the transmitted ground truth. The optimal solution often requires the auto-covariance matrix of the equalized/detected soft symbol vector, and the cross-covariance of the detected symbol vector and its corresponding transmitted vector. However, the statistics are often difficult to estimate, and the receiver has to make reasonable approximations and assumptions. Therefore, it is challenging to ensure convergence and achieve good performance, especially when the receiver has low SNR. Four time-domain turbo equalizers have been developed for UWA communications, as shown in Fig. 3, which can effectively combat the long multipath and severe Doppler channels. This section details each of the TD turbo equalizers, and then compares their complexity and performance.

TD TURBO LE

A time-domain linear equalizer is an FF filter that adapts the filter coefficients to minimize the MSE between the transmitted true symbols and the equalized symbols. Figure 3a depicts a

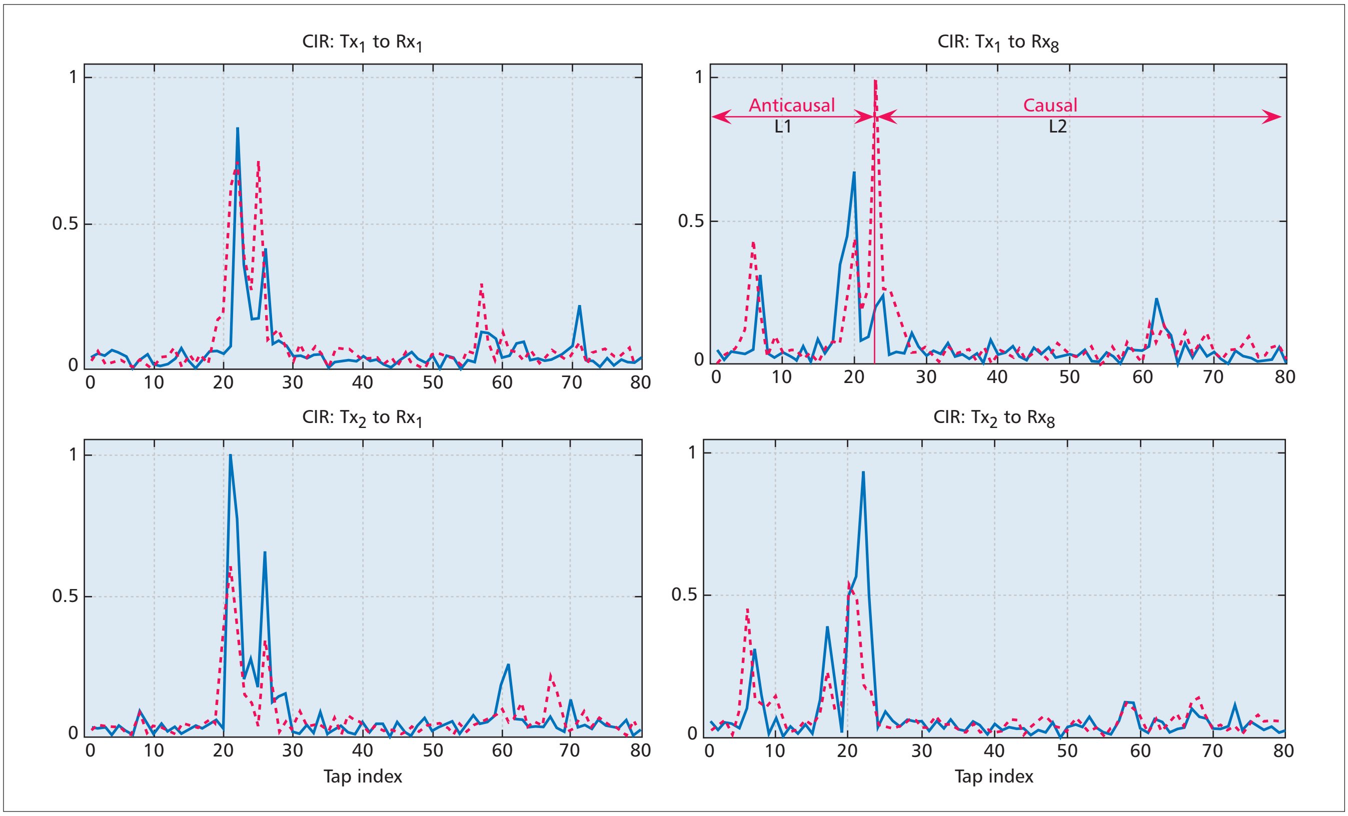


Figure 2. Magnitude of MIMO UWA channel impulse responses at two different time instants, estimated from the data recorded in the GOMEX2008 experiment.

block-based MIMO turbo LE [4] that utilizes hybrid soft interference cancellation. The received signal **y** of each sub-block is preprocessed to remove the interference from the previous sub-block. Assume the sub-block length is L_s and the equalizer filter for the nth transmitter branch is $\mathbf{f}_n$, which has $L_f = \mathrm{L}_a + \mathrm{L}_c + 1$ coefficients, with L_a and L_c being the lengths of the anticausal and causal taps, respectively. The IBI-free signal $\mathbf{y}_k$ at time instant k is a concatenated vector of length ML_f. To equalize the symbol corresponding to the nth transmit branch and the kth time instant, the soft interference from other transmitted branches and time instants is reconstructed by convolving the estimated MIMO channel $\hat{\mathbf{H}}$ with the masked soft symbol vector $\bar{\mathbf{s}}_{n,k}$, the elements of which are the soft symbols previously detected. The precursor (or future) symbols $\tilde{\mathbf{s}}$ are calculated by $\lambda_a^{(i-1)}$, the a priori LLR of the $(i - 1)$th iteration. The post-cursor (or past) symbols $\breve{\mathbf{s}}$ are calculated by combining the extrinsic LLR and the a priori LLR of the current iteration as $\lambda_s^{(i)} = \lambda_e^{(i)} + \lambda_a^{(i)}$, where $\lambda_a^{(i)}$ is the interleaved decoder LLR of the current iteration. The output of the HSIC unit is filtered through the equalizer to yield the (n, k)th soft symbol $\hat{s}_{n,k}$.

Using the equalized symbol to compute the extrinsic LLR requires knowledge of the distribution of $\hat{s}_{n,k}$ which is often unknown. The common approach is to simply assume that the fading channel cascaded with the equalizer is equivalent to a Gaussian noise channel between the transmit symbol and the equalized symbol, such that

$$\hat{s}_{n,k} = \mu_{n,k}\, s_{n,k} + \upsilon_{n,k}, \qquad (2)$$

with $\upsilon_{n,k}$ being a zero-mean Gaussian noise. Conditioned on a given transmitted symbol in the modulation scheme, the equalized symbol $\hat{s}_{n,k}$ follows a Gaussian distribution with mean and variance computed from the equalizer vector and the corresponding channel vector.

The optimal equalizer coefficients are computed through the covariance matrices of the received signal vector $\mathbf{y}_k$ and the a priori soft symbols $\bar{\mathbf{s}}$, as well as the MIMO fading channel [4]. Theoretically, the optimal equalizer vector can adapt for each n and k. However, due to the high computational complexity of matrix inversion, the low-complexity approach is to design a time-invariant equalizer vector for each transmit branch n and adapt the equalizer after each Turbo iteration.

The turbo LE in [4] also incorporates a reliability ordering scheme to enhance the performance. The reliability of a soft decision is measured as the inverse of the variance of the soft decision. The idea of reliability ordering extends the layering approach in the V-BLAST to order the detection of symbols in both space and time. The V-BLAST selects the transmit branch with the strongest receiving power to detect first, while the turbo LE selects the symbol with the highest reliability to detect first. The interference of the detected symbol to other symbols in the block is then removed through the HSIC. The turbo LE with HSIC also enjoys low computational complexity and stable performance besides high data efficiency.

TD Turbo BDFE

A reliability-based block DFE (BDFE) is illustrated in Fig. 3b, where the IBI removal, channel estimation, and SIC units are similar to those used in the TD turbo LE. However, the BDFE uses two filter matrices — an FF matrix and an FB matrix — to output a block of equalized symbols **r** simultaneously for all N branches. The FF filter **F** in the BDFE has coefficients for all NL_f rows and ML_f columns, rather than the one-column FF filter vector $\mathbf{f}_n$, which is designed separately for each branch in the turbo LE scheme. Meanwhile, the FB filer **B** is an NL_f by NL_f matrix that reconstructs the residual ISI vector **x** from the *a posteriori* soft symbols $\breve{\mathbf{s}}$ and the masked a priori soft symbols $\bar{\mathbf{s}}_{n,k}$. The overall equalized symbol vector is then $\hat{\mathbf{s}} = \mathbf{r} - \mathbf{x} + \bar{\mathbf{s}}$.

Taking advantage of the block of symbols available at the equalizer output, the BDFE utilizes the sequence-based LLR to calculate the *a posteriori* LLR $\lambda_T^{(i)}$ of each symbol and to detect the *a posteriori* soft symbols $\breve{\mathbf{s}}$. Note that we use the subscript T in the total LLR $L_T^{(i)}$ to distinguish from the sum LLR $L_s^{(i)}$ in other schemes. The post-cursor *a posteriori* symbols are also fed

BDFE suffers from huge computational complexity, especially when the block size is large, which is often the case in UWA. To reduce the complexity, two symbol-based SDFEs are developed by replacing the matrix filters with vector filters, replacing the sequence-based LLR by symbol-based LLR, and changing the FB filter structures.

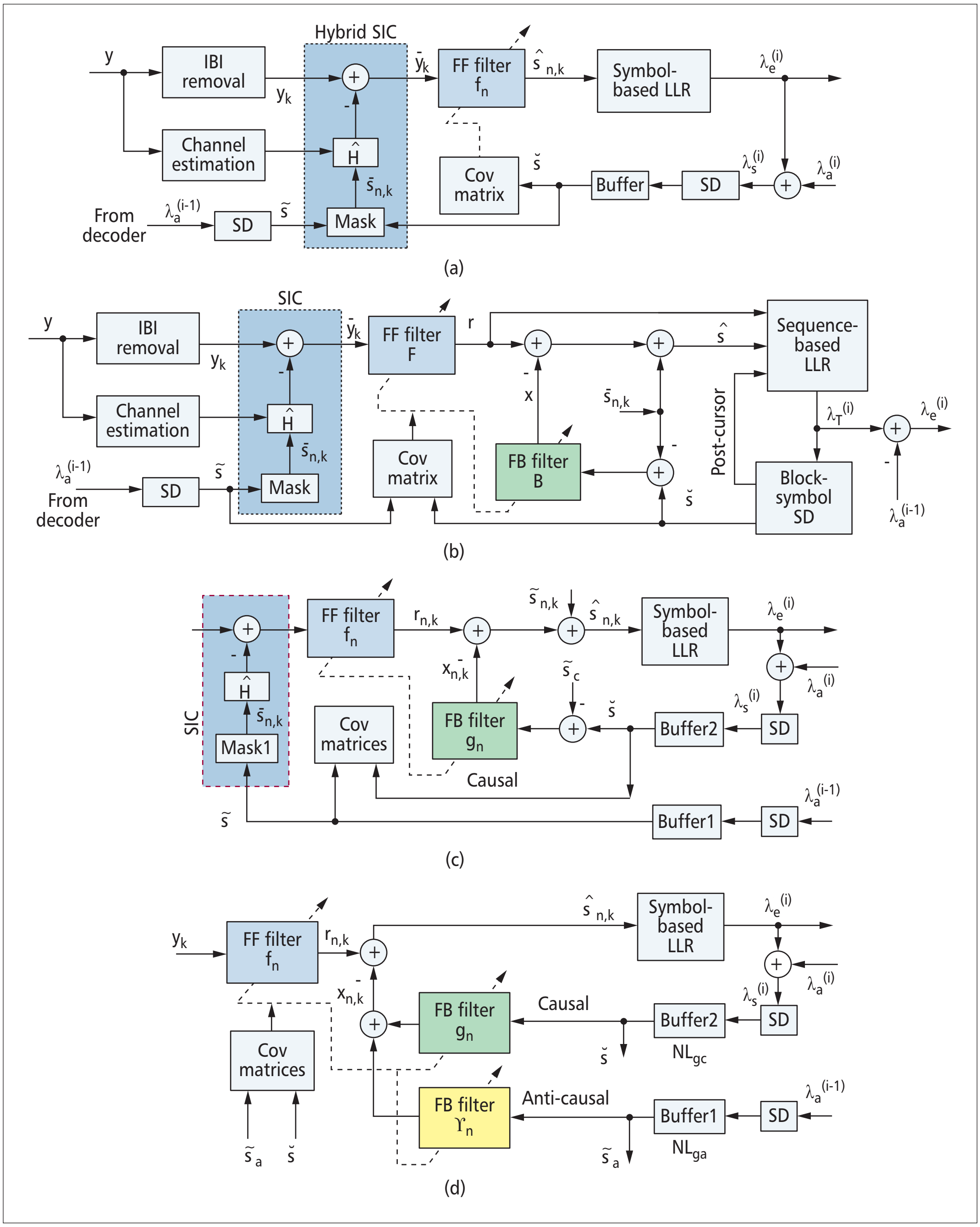


Figure 3. Time-domain turbo equalizers with the mask block set the (n, k)th element to zero. Note that turbo LE, SDFE, and SICE have N sets of FF and FB filters $\mathbf{f}_n$, $\mathbf{g}_n$, and $\boldsymbol{\gamma}_n$ separately designed for each branch; while the BDFE has concatenated FF and FB filters, which can suppress CCIs from other transmit branches: a) turbo LE with symbol-based LLR and HSIC [4]; b) BDFE with sequence-based LLR [5]; c) SDFE with symbol-based LLR and SIC [6]; d) SICE with symbol-based LLR [7].

Matrix inversion and Cholesky decomposition are required to compute the optimal FF and FB filter matrices, leading to high computational complexity. The complexity can be reduced by using the same FF and FB filter matrices for all symbols in one iteration, and such a simplification results in only marginal performance degradation compared to the full adaptive design.

back to the LLR unit to aid the calculation of the total LLR $\lambda_T^{(i)}$. The extrinsic LLR $\lambda_e^{(i)}$ of the ith iteration is then $\lambda_e^{(i)} = \lambda_T^{(i)} - \lambda_a^{(i-1)}$, the difference between the total LLR of the ith iteration and the a priori LLR of the $(i - 1)$th iteration.

The optimal solution of the BDFE filters is computed from the covariance matrix of the error vector $\mathbf{e} = \mathbf{r} - \mathbf{x}$ and the covariance matrix of the a priori soft symbol vector $\bar{\mathbf{s}}$. Matrix inversion and Cholesky decomposition are required to compute the optimal FF and FB filter matrices, leading to high computational complexity. The complexity can be reduced by using the same FF and FB filter matrices for all symbols in one iteration, and such a simplification results in only marginal performance degradation compared to the full adaptive design. In addition, the calculation of the FF and FB filters are based on the common assumption of error-free decision feedback, which eliminates the need for cross-covariance between the transmitted and equalized symbol vectors. The BDFE also calculates the accurate total LLR without the approximated Gaussian assumption (Eq. 2) of the soft decisions, which is commonly used in the turbo LE case. Compared to the TD turbo LE, the BDFE significantly reduces error propagation through the inclusion of the FB filter, and reduces processing latency through the block processing approach.

TIME-DOMAIN SDFE AND SICE

Despite its excellent performance, BDFE suffers from huge computational complexity, especially when the block size is large, which is often the case in UWA. To reduce the complexity, two symbol-based SDFEs are developed by replacing the matrix filters with vector filters, replacing the sequence-based LLR by symbol-based LLR, and changing the FB filter structures.

Figure 3c shows the structure of the SDFE, where the nth FF filter $\mathbf{f}_n$ has a length NL_f with $L_f = L_a + L_c + 1$, which is similar to the turbo LE. The FB filter $\mathbf{g}_n$ of the SDFE has a length NL_{gc} with $L_{gc} = L_2 + L_c + 1$ and L_2 being the length of the causal CIR. The a priori soft symbols $\tilde{\mathbf{s}}$ are soft decisions from the a priori LLRs of the $(i - 1)$th iteration, which are input to the SIC and covariance estimator. The a priori soft symbols of the ith iteration are available for causal taps and are input to the FB filter after subtracting $\tilde{\mathbf{s}}_c$, the causal portion of $\tilde{\mathbf{s}}$. The equalized soft symbol is then

$$\hat{s}_{n,k} = \mathbf{f}_n^H(\mathbf{y}_k - \hat{\mathbf{H}}\bar{\mathbf{s}}_{n,k}) - \mathbf{g}_n^H(\breve{\mathbf{s}} - \tilde{\mathbf{s}}_c) + \tilde{s}_{n,k}.$$

Intuitively, the FB filter reconstructs the postcursor interference to the (n, k)th symbol, while the precursor interference is suppressed partially by the SIC and the FF filter. The optimal solutions to the FF and FB filter vectors are derived by assuming Gaussian channel between $\hat{s}_{n,k}$ and the ground-truth symbol $s_{n,k}$ as in the turbo LE case, but with non-perfect soft decisions. The low-complexity time-invariant solutions require the auto-covariance matrices of $\mathbf{s}$ and $\breve{\mathbf{s}}$, and the cross-covariance matrix between $\mathbf{s}$ and $\breve{\mathbf{s}}$. However, the equalizer has no access to the transmit symbol vector $\mathbf{s}$, and the statistics of the a priori symbol vector $\tilde{\mathbf{s}}$ are used instead.

Although it has fast initial convergence, the SDFE in Fig. 3c suffers from an error floor that prevents it from reaching the matched filter bound. The main reason is that the precursor interference cancellation is intertwined in the SIC and the FF filter. This motivates the SICE [7], which uses an additional transversal filter to explicitly reconstruct the precursor interference. As shown in Fig. 3d, the FB filter γ_n is a length-NL_{ga} filter for reconstructing the interference from the precursor soft symbols detected by the a priori LLR of the $(i - 1)$th iteration $\lambda^{(i-1)}$, where $L_{ga} = L_1 + L_a + 1$. The FF filter $\mathbf{f}_n$ and FB filter $\mathbf{g}_n$ are the same as those in the SDFE, but the SIC unit is removed.

Again, the filters are jointly adapted to minimize the MSE between the equalized symbol $\hat{s}_{n,k}$ and the transmitted symbol $s_{n,k}$, and the equivalent Gaussian channel assumption is used to approximate the statistics. The solution requires the auto-covariance matrices $\Phi_{\breve{\mathbf{s}}\breve{\mathbf{s}}}$ and $\Phi_{\mathbf{s}\tilde{\mathbf{s}}}$, and the cross-covariance matrices $\Phi_{\mathbf{s}\breve{\mathbf{s}}}$ and $\Phi_{\mathbf{s}\tilde{\mathbf{s}}}$. Without the ground-truth $\mathbf{s}$, the SICE uses the turbo LE only in the first iteration, and then uses the a priori soft symbols $\tilde{\mathbf{s}}$ in place of the ground-truth symbols to compute the cross-covariance matrices $\Phi_{\tilde{\mathbf{s}}\breve{\mathbf{s}}}$ and $\Phi_{\tilde{\mathbf{s}}\tilde{\mathbf{s}}}$.

EXPERIMENTAL RESULTS OF TD TURBO EQUALIZERS

Many ocean experiments were conducted with single-carrier MIMO underwater communication systems during 2007–2010, and results were reported in several publications. Here, the performance of the four TD equalizers were evaluated by the ocean experiment SPACE08, which was conducted on the coast of Martha's Vineyard, Edgartown, Massachusetts, in October 2008. The carrier frequency was $f_c = 13$ kHz, and the symbol period was $T_s = 0.1024$ ms. The transmit pulse shaping was a square-root raised cosine filter with a roll-off factor 0.2; thus, the occupied channel bandwidth was 11.7188 kHz. The transmit equipment had four transducers, and the receiver had 12 hydrophones. The transmission frame began with a linear frequency modulation (LFM) signal, followed by three packets with quadrature PSK (QPSK), 8PSK, and 16-quadrature amplitude modulation (QAM), respectively, and ended with a trailing LFM. Each packet included a maximum-length sequence (m-sequence) of length 511, and a data payload consisting of 30,000 symbols. The coding scheme was a rate-1/2 convolutional code. An analysis of the channel scattering function revealed that the maximum Doppler spread was about 5 Hz, corresponding to a channel coherence time about 100 ms or an approximate length of 1000 taps in terms of the symbol period $T_s = 0.1024$ ms. The receiver was located at 200 m or 1000 m away from the transmitter, and the recorded signals had 45 packets for 200 m and 34 packets for 1000 m. The multipath channel lengths of both ranges were estimated as $L = 100$ taps.

The bit error rate (BER) performance of the TD equalizers are compared in Fig. 4 for 8PSK and 16-QAM transmissions. The BDFE achieved the best performance, followed by the SICE.

The SDFE sometimes performed similar to the turbo LE if not worse. The EXIT chart simulation, not shown here due to page limits, indicates that the SDFE had trouble converging with the matched filter bound, while other TD turbo equalizers do.

Frequency-Domain Turbo Equalization

When the channel length is large, single-carrier frequency-domain equalizers (SC-FDEs) can effectively reduce computational complexity by converting the received signals into FD via fast Fourier transform (FFT). With the overhead of FFT and inverse FFT (IFFT), the complexity of FDE is lower than that of the TDE when the equalizer length is larger than 10 [8]. Since the UWA channels are often longer than 50 taps and the equalizer lengths are longer than the channels, the computational saving of FDEs is significant. Additionally, SC-FDEs can be applied to continuous transmission without ZP or CP, which is required in OFDM systems. Since ZP or CP has to be longer than the channel length in OFDM block transmission, SC-FDEs can also achieve higher data efficiency than OFDM. This section will discuss the FD turbo LE and DFE schemes suitable for MIMO UWA channels, as shown in Fig. 5, where the FFT length K is the same as the sub-block length K_{sb} in Fig. 1b. The overlap length K_{ovlp} for receiver processing is set to be the same as the channel length l, and the channel is assumed to be time-invariant for each sub-block.

Although not shown in Fig. 5, inter-block interference (IBI) removal and a CP reconstruction block are used to preprocess the received signals when the transmitted data have no ZP or CP. The IBI from the previous sub-block is removed from the current sub-block, and the CP effect is added back to the current sub-block by reconstructing the received signals due to the last K_{ovlp} transmitted symbols of the current sub-block. The CP reconstruction renders a block-circulant channel matrix in the TD, hence enabling the FD equalizers to reduce the computational complexity. It is also assumed that the FFT unit is preceded by a serial-to-parallel buffer, and the IFFT unit is followed by a parallel-to-serial buffer. The channel estimation is done in the time domain.

Frequency-Domain Turbo LE

Single-carrier frequency-domain turbo linear equalizers (FD-LE) have been successfully applied to UWA MIMO communications. The basic structure of an FD turbo LE is shown in Fig. 5a, where the frequency domain channel matrix $\hat{\mathcal{H}}$ becomes block-diagonal after the FFT, and the FD equalizer is computed by k matrix inversions of $N \times N$ small matrices instead of a huge TD channel matrix of size $(NK) \times (NK)$. However, the transmitted symbols of the current sub-block are unknown in the first iteration. Therefore, the equalizer has to treat them as zeros and their contribution in the received signal as interfering noise [7]. After the first iteration, the CP is reconstructed from the soft

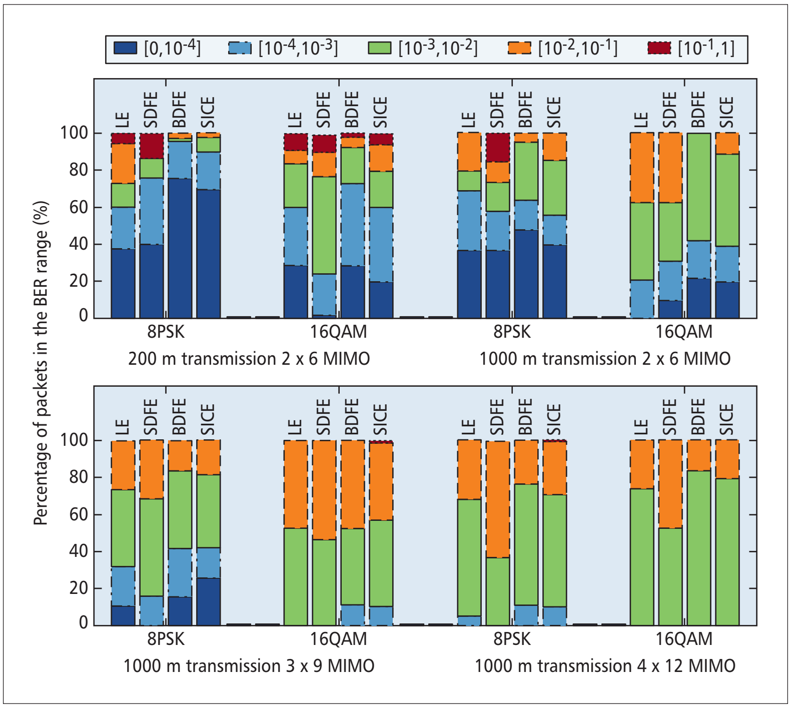


Figure 4. Error performance of the TD turbo equalizers.

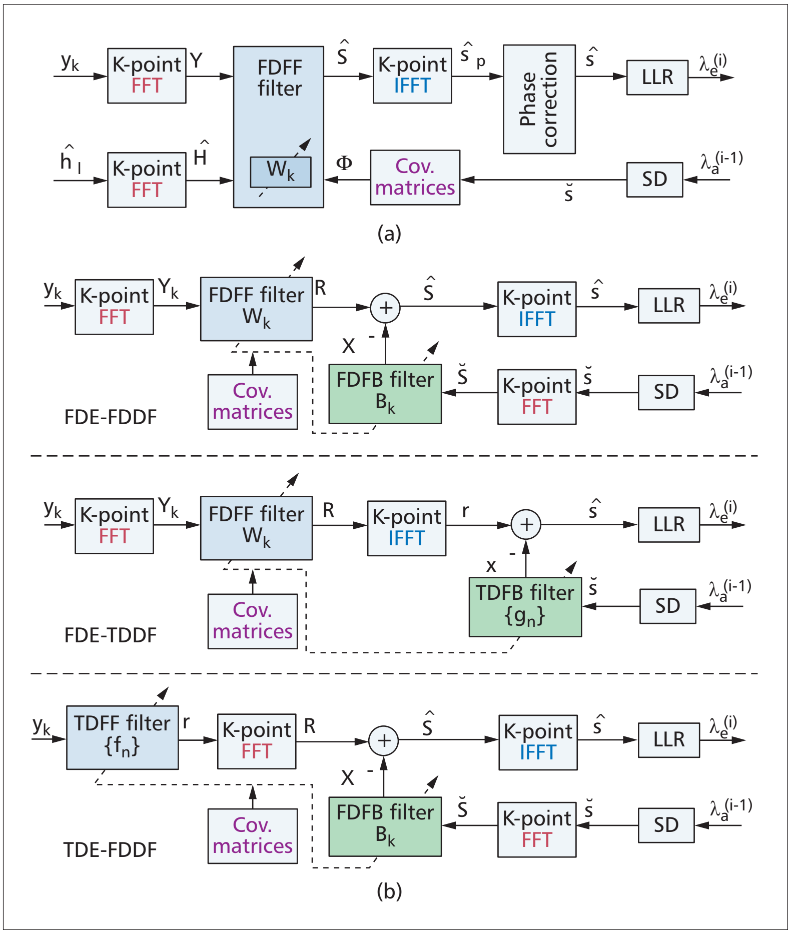
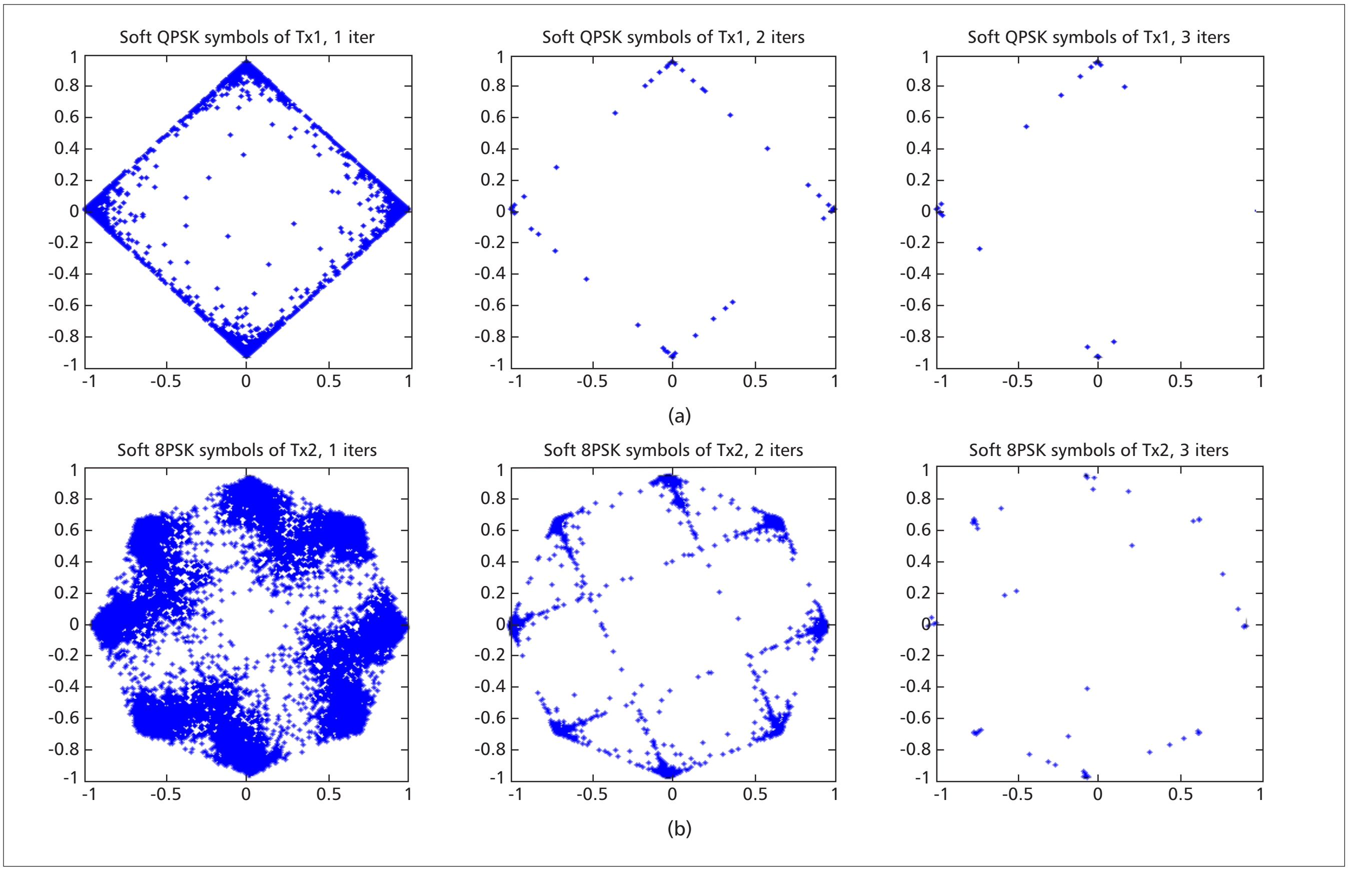


Figure 5. FD turbo equalizers, where upper case letters denote FD signals and lower case letters denote TD signals: a) FD turbo linear equalizer; b) FD turbo equalizer with decision feedback.



(a)

(b)

Figure 6. Scatter plots of estimated soft symbols of a packet recorded by the SPACE08 experiment with 1000 m transmission: a) QPSK; b) 8PSK.

symbols $\check{\mathbf{s}}$ of the previous iteration. The FD equalizers $\mathbf{W}_k$ are designed for each frequency bin $k = 1, \cdots, K$ using the FD channel matrix and the covariance matrix of the a priori soft symbols.

After the FD equalizer, the TD soft symbols $\hat{\mathbf{s}}_p$ exhibit rapid phase rotation due to the high instantaneous Doppler spread in most UWA channels. The Doppler effect leads to non-zero entries on some of the off diagonal blocks of the FD channel matrix, and the simplified implementation of the FD equalizer results in uncompensated phase in the estimated soft symbols. A phase estimation and correction unit is added to remove the phase rotation before LLR calculation. This unit is unique for severe Doppler UWA channels, and is easy to implement by a few multiply-add operations. Other performance enhancement techniques are also available, such as layered detection with soft successive interference cancellation, time-reversal filtering, and sequence-based LLR calculation, which also increase the computational complexity.

Frequency-Domain Turbo DFE

The FD turbo DFEs (FD-DFEs) have two common forms: a frequency-domain equalizer with frequency-domain decision feedback (FDE-FDDF) has both FF and FB filters in the frequency domain, as shown in Fig. 5b; a frequency-domain equalizer with time-domain decision feedback (FDE-TDDF) has an FF filter in FD followed by a TD feedback filter, as shown in Fig. 5b. It is interesting to note that a block-based TDE with FDDF, shown in Fig. 5c, is also available in the literature, and the time-domain BDFE or SDFE scheme is similar to the TDE-TDDF scheme.

For low-complexity implementation, the FF and FB filters in the FDE-FDDF scheme are designed for each frequency bin, and they require inversion of small matrices. In the FDE-TDDF scheme, the $M\times N$ FF filter $\mathbf{W}_k$ is designed for each frequency bin $k = 1, \cdots, K$, and the $(NL_{gc}) \times 1$ FB filter $\mathbf{g}_n$ is designed for each transmit branch $n = 1, ..., N$. In the TDE-FDDF scheme, the FF filters are N sets of $(ML_f)\times 1$ vectors $\mathbf{f}_n$, and the FB filters are K sets of $M \times N$ matrices $\mathbf{b}_k$. The adaptation of the filters at each iteration uses the estimated FD or TD channel matrices and the covariance matrices of the soft a priori symbols from the previous iteration.

Experimental Results for FD Turbo Equalizers

Many data sets from ocean experiments have been processed by the FDE-LE and FDE-FDDF schemes. To illustrate the gain of iterative detection, the scatter plots of soft QPSK and 8PSK symbols after soft decision decoding at different turbo iterations are shown in Fig. 6. It is clearly seen that turbo equalization and detection gradually separate the soft symbols and push them closer to valid constellation points. The BER performance of the FD turbo LE is similar to that of the TD turbo LE [10]. It is also noted that all the FDE soft decision feedback schemes yield similar performance as the FD turbo LE with interference cancellation.

It is clearly seen that turbo equalization and detection gradually separated the soft symbols and pushed them closer to valid constellation points. The BER performance of the FD turbo LE is similar to the TD Turbo LE. It is also noted that all the FDE soft decision feedback schemes yield similar performance as the FD turbo LE with interference cancellation.

Conclusion

This article reviews four TD turbo equalizer schemes and four FD turbo equalizer schemes for single-carrier MIMO UWA communications. Their performance is compared using the SPACE08 experimental data. The turbo LEs in both TD and FD achieved satisfactory performance in severe ISI channels, and their computational complexity is affordable. Low-complexity turbo SICE and FDE-TDDF are also promising for practical implementation and can achieve better performance than the turbo LE in severe non-causal channels. The time-domain BDFE scheme achieves the best performance at the cost of high computational complexity. Future work in this area includes analysis of complexity-performance trade-offs and real-time implementation.

Acknowledgment

This work was supported in part by the Office of Naval Research under grants N00014-07-1-0219 and N00014-10-1-0174, and by the National Science Foundation under grants ECCS-0846486, CCF-0915846, ECCS-1202075, ECCS-1231848, and ECCS-1408316. The authors would like to thank their Ph.D. students Jun Tao, Jian Zhang, Huang Lou, Amirhossein Rafati, Longbao Wang, Zengli Yang, and Weimin Duan for producing some of the results presented in the article.

References

[1] A. Singer, J. Nelson, and S. Kozat, "Signal Processing for Underwater Acoustic Communications," *IEEE Commun. Mag.*, vol. 47, no. 1, Jan. 2009, pp. 90–96.

[2] C. Laot, A. Glavieux, and J. Labat, "Turbo Equalization: Adaptive Equalization and Channel Decoding Jointly Optimized," *IEEE JSAC*, vol. 19, no. 9, 2001, pp. 1744–52.

[3] B. Li *et al.*, "Multicarrier Communication Over Underwater Acoustic Channels with Nonuniform Doppler Shifts," *IEEE J. Ocean. Eng.*, vol. 33, no. 2, 2008, pp. 198–209.

[4] J. Tao *et al.*, "Enhanced MIMO LMMSE Turbo Equalization: Algorithm, Simulations, and Undersea Experimental Results," *IEEE Trans. Signal Processing*, vol. 59, no. 8, Aug. 2011, pp. 3813–23.

[5] J. Tao *et al.*, "Robust MIMO Underwater Acoustic Communications Using Turbo Block Decision-Feedback Equalization," *IEEE J. Ocean. Eng.*, vol. 35, no. 4, Oct. 2010, pp. 948–60.

[6] A. Rafati, H. Lou, and C. Xiao, "Soft-Decision Feedback Turbo Equalization for LDPC-Coded MIMO Underwater Acoustic Communications," *IEEE J. Oceanic Eng.*, vol. 39, no. 1, Jan. 2014, pp. 90–99.

[7] J. Wu, L. Wang, and C. Xiao, "Low Complexity Soft-Interference Cancelation Turbo Equalization for MIMO systems with Multilevel Modulations," *IET Commun.*, vol. 9, no. 5, 2015, pp. 728–35.

[8] N. Benvenuto *et al.*, "Single Carrier Modulation with Nonlinear Frequency Domain Equalization: An idea Whose Time Has Come Again," *Proc. IEEE*, vol. 98, no. 1, Jan 2010, pp. 69–96.

[9] J. W. Choi *et al.*, "Adaptive Linear Turbo Equalization over Doubly Selective Channels," *IEEE J. Ocean. Eng.*, vol. 36, no. 4, Oct. 2011, pp. 473–89.

[10] J. Zhang and Y. Zheng, "Frequency-Domain Turbo Equalization with Soft Successive Interference Cancellation for Single Carrier MIMO Underwater Acoustic Communications," *IEEE Trans. Wireless Commun.*, vol. 10, no. 9, Sept. 2011, pp. 2872–82.

Biographies

Yahong Rosa Zheng [M'03, SM'07, F'15] (zhengyr@mst.edu) received her B.S. degree in electrical engineering from the University of Electronic Science and Technology of China, Chengdu, in 1987, her M.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 1989, and her Ph.D. degree from Carleton University, Ottawa, Ontario, Canada, in 2002. She was an NSERC postdoctoral fellow from January 2003 to April 2005 at the University of Missouri-Columbia. In fall 2005, she joined the Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, where she is currently a professor. Her research interests include array signal processing, wireless communications, and wireless sensor networks. She has served as a Technical Program Committee (TPC) member for many IEEE international conferences, including IEEE Vehicular Technology, IEEE GLOBECOM, IEEE ICC, IEEE Wireless Communications and Networking Conference, and others. She served as a Wireless Communications Symposium Co-Chair for ICC 2014 and GLOBECOM 2012. She also served as an Associate Editor of *IEEE Transactions on Wireless Communications*, 2006–2008. She is currently an Associate Editor of *IEEE Transactions on Vehicular Technology*. She was the recipient of an NSF CAREER Award in 2009.

Jingxian Wu [S'02, M'06] (wuj@uark.edu) received his B.S. degree from Beijing University of Aeronautics and Astronautics, China, in 1998, his M.S. degree from Tsinghua University, Beijing, in 2001, and his Ph.D. degree from the University of Missouri at Columbia in 2005, all in electrical engineering. He is currently an associate professor with the Department of Electrical Engineering, University of Arkansas, Fayetteville. His research interests mainly focus on signal processing for large-scale networks and wireless communications, including energy-efficient information sensing and processing, green communications, and statistical data analytics. He served as Co-Chair of the 2012 Wireless Communication Symposium of IEEE ICC, and the 2009 and 2015 Wireless Communication Symposia of IEEE GLOBECOM. He served as an Associate Editor of *IEEE Transactions on Vehicular Technology* from 2007 to 2011. He is also an Editor of *IEEE Transactions on Wireless Communications* and an Associate Editor of *IEEE Access*.

Chengshan Xiao [M'99, SM'02, F'10] (xiaoc@mst.edu) earned a B.S. degree in electronic engineering from the University of Electronic Science and Technology of China in 1987, an M.S. degree in electronic engineering from Tsinghua University in 1989, and a Ph.D. in electrical engineering from the University of Sydney in 1997. Since 2010, he has been a professor with the Department of Electrical and Computer Engineering at Missouri University of Science and Technology. He is currently serving as a program director at USA National Science Foundation through intergovernmental personnel assignment. Previously, he was a senior member of scientific staff with Nortel Networks, Ottawa, Canada, a faculty member at Tsinghua University, the University of Alberta, Edmonton, Canada, and the University of Missouri — Columbia. He also held visiting professor positions in Germany and Hong Kong. His research interests include wireless communications, signal processing, and underwater acoustic communications. He is the holder of three U.S. patents. He is the Director of Conference Publications of IEEE ComSoc and a member of the IEEE ComSoc Fellow Evaluation Committee. Previously, he served as an elected member of the IEEE ComSoc Board of Governors, and a Distinguished Lecturer of both IEEE ComSoc and the IEEE Vehicular Technology Society. He has also served as an Editor, Area Editor, and Editor-in-Chief of *IEEE Transactions on Wireless Communications*, and an Associate Editor of *IEEE Transactions on Vehicular Technology* and *IEEE Transactions on Circuits and Systems-I*. He was Technical Program Chair of IEEE ICC 2010, Cape Town, South Africa. He served as the founding Chair of the IEEE Technical Committee on Wireless Communications. He is a recipient of the 2014 Humboldt Research Award from the Alexander von Humboldt Foundation, and the 2014 Joseph LoCicero Award from IEEE ComSoc.

Structured Sparse Methods for Active Ocean Observation Systems with Communication Constraints

Urbashi Mitra, Sunav Choudhary, Franz Hover, Robert Hummel, Naveen Kumar, Shrikanth Naryanan, Milica Stojanovic, and Gaurav Sukhatme

Urbashi Mitra, Sunav Choudhary, Naveen Kumar, Shrikanth Naryanan, and Gaurav Sukhatme are with the University of Southern California.

Franz Hover and Robert Hummel are with the Massachusetts Institute of Technology.

Milica Stojanovic is with Northeastern University.

ABSTRACT

Actuated sensor networks enabled by underwater acoustic communications can be efficiently used to sense over large marine expanses that are typically challenged by a paucity of resources (energy, communication bandwidth, number of sensor nodes). Many marine phenomena of interest admit sparse representations, which, coupled with actuation and cooperation, can compensate for being data starved. Herein, new methods of field reconstruction, target tracking, and exploration-exploitation are provided, which adopt sparse approximation, compressed sensing, and matrix completion algorithms. The needed underlying structure (sparsity/low-rank) is quite general. The unique constraints posed by underwater acoustic communications and vehicle kinematics are explicitly considered. Results show that solutions can be practically implemented, even over large ocean spaces.

INTRODUCTION

Future advances in ocean monitoring, offshore industry, and basic marine sciences will rely heavily on our ability to jointly consider communication, actuation, and sensing in a unified system that includes remote instruments, underwater vehicles, human operators, and sensors of all types. These tasks will require methods to detect and track large-scale ocean phenomena such as algal blooms, oil spills, ocean currents, and hydrothermal vents, as well as man-made signals such as those emanating from an airplane's black box. We envision a scenario as depicted in Fig. 1, where multiple autonomous underwater vehicles (AUVs) interact and coordinate via acoustic communications with a network of sensors to detect and track a phenomenon of interest. The underlying network architecture includes both static, communication-enabled sensor nodes, as well as actuated nodes in the form of AUVs. Thus, our system needs to control and move some of the nodes to achieve its sensing and communication goals. Moreover, the choices made regarding communication, control, and sensing are interdependent [1].

These goals require active consideration of the underlying problem of exploration and exploitation. A critical feature of this problem is sparsity: our sensors will be sparsely deployed due to cost and complexity considerations; our fleet of AUVs will be modest in number. As a result, our observations of a vast ocean can only be a scant sampling of the phenomena of interest. Satellite remote sensing and traditional ship operations in a large-scale underwater observation system are similarly limited, and many ocean measurements are made point-wise or confined to a local area. Furthermore, we often need to first find (explore for) a feature and then track (exploit) it, as illustrated in Fig. 2.

While there has been strong attention paid to exploiting the sparse nature of underwater acoustic channels (e.g., [2]), our problems of interest are very different. We wish to examine how to task sensors, static and/or mobile, toward achieving an underwater mission goal that will be dependent on the sensor network's understanding of its environment. Prior work related to field estimation and target tracking suggests that sparse representations exist for many features of interest, and therefore, we may not be irreparably data-starved. At the same time, a mission has to be accomplished by coordinated sensing and action, typically enabled by acoustic communication between energy-constrained and kinematically constrained agents. In this article, we review how sparsity can be leveraged in the design of exploration-exploitation methods tailored to underwater sensing systems. Our particular approach does not require detailed prior information about the fields or targets of interest.

In particular, we tackle three related problems: field reconstruction [1, 10, 11, 12], target tracking [5, 13], and methods for exploration-exploitation [14, 15], which build on the insights derived from the previous approaches. Field

0163-6804/15/$25.00 © 2015 IEEE

reconstruction is an important issue for many scientific, industrial, and tactical applications. Examples include geographical, chemical, and temperature map building. In some cases, reconstruction is the end goal (e.g., climate recording), while in other cases it represents an intermediate step to support the mission at hand, for example, one in which mobile assets need to learn about the environment in which they operate and adapt to it. Sensor measurements of the field are typically correlated in some domain, and this correlation forms the basis of compressive sensing theory: when the field to be reconstructed is sparse (with respect to some basis), it suffices to collect data only at some of the points to reconstruct the full view of the field.

In recent years, there has been a surge of new results in the area of sparse field reconstruction using sensor networks. Although harsh environments, energy management, and limited communications have been considered, most of these methods have been optimized for terrestrial conditions. Underwater systems pose new challenges and constraints. For example, sparse methods have been developed predominantly ignoring the cost of sampling, while in oceanographic field applications, constraints such as limited energy and maneuverability can dominate, and the number of samples may be quite low. The interplay between such resource constraints and the fundamental sparse algorithms is a general problem we refer to as kinematically constrained sparse approximation (KCSA) [10, 11].

The extension from field reconstruction to target tracking is a natural one, and we provide methods that explicitly consider the dynamics of the targets and the nature of the underwater acoustic communication channel. An example is random access compressed sensing (RACS) [12, 13], which notably does not require synchronization. Target detection is a related application that addresses whether a specific phenomenon of interest has occurred. The target usually cannot be observed directly, however, and the task then is to infer the presence or location of the target from indirect observations, with an eye to the cost of acquiring each sample, thus resulting in the exploration-exploitation [14, 15] optimization trade-off.

We summarize key challenges of underwater sensing and communication networks that are enabled by actuation, including the unique features of the acoustic channel and the constraints of underwater vehicles. We lay out the key ideas in compressive sensing, sparse approximation, and matrix completion. We examine how to reconstruct a large field when limited to a few mobile vehicles operating with constraints over a large space. We change the perspective by considering field reconstruction when using static underwater sensors. Here, the impact of using the underwater acoustic medium to transmit sensor data to a fusion center introduces unique complexities; and the work is extended to the tracking of mobile sources. Finally, the insights of the prior two research areas are employed to derive novel methods based on matrix completion to solve a realistic exploration-exploitation problem.

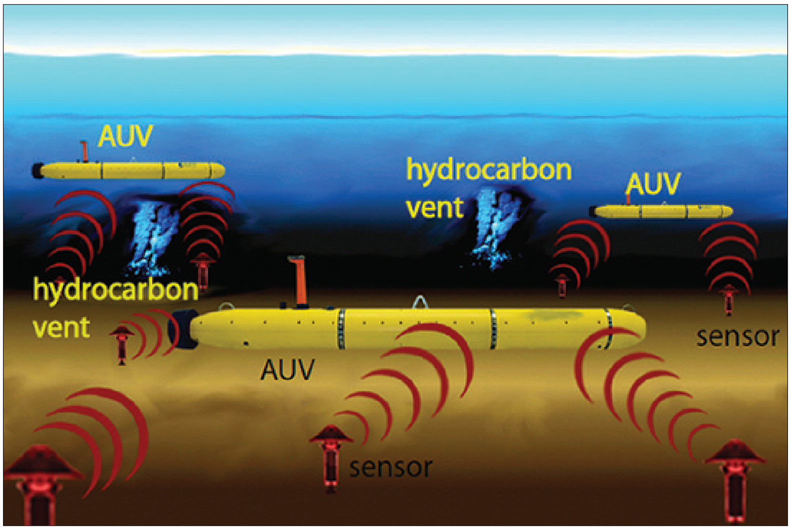


Figure 1. The envisioned network architecture for future ocean observing systems, comprising both fixed and moving nodes capable of advanced sensing and wireless communications.

The Challenges of Underwater Actuated Networks

The boundaries between communications, networking, navigation, and sensing are blurred in underwater actuated sensor networks, where multiple levels of dynamics exist: moving (possibly actuated) nodes, time-varying communication channels, and evolving phenomena to be detected, identified, and tracked.

Underwater Acoustic Communication Channels

Challenges in the design of underwater acoustic communication systems include a severely limited, range-dependent bandwidth, extensive time-varying multipath propagation, and long propagation delays caused by the low speed of sound underwater (1500 m/s) [3]. Absorption leads to exponential path losses with respect to frequency. Propagation delays result in significant delay spreads (on the order of hundreds of symbols). There are very short channel coherence times that are exacerbated in mobile environments and orders of magnitude shorter than in RF systems. In terms of the channel, we focus on the sparse nature of multipath propagation.

Underwater Sensors and Vehicles

Ocean vehicle systems face obstacles not shared in the terrestrial and aerial domains. Most of the ocean remains uncharted, with detailed information about the sea floor available only at specific and scientifically important sites, or areas of specific interest such as oil and gas fields. Underwater systems are expensive to build and instrument, and ship time also introduces a huge cost that encourages remote operations. Unlike land-based vehicles, underwater vehicles are exposed to extreme pressure, corrosion, and fouling, and strong variable wave, current, and wind disturbances. Such disturbances make operating a vehicle difficult, if not impossible. At the

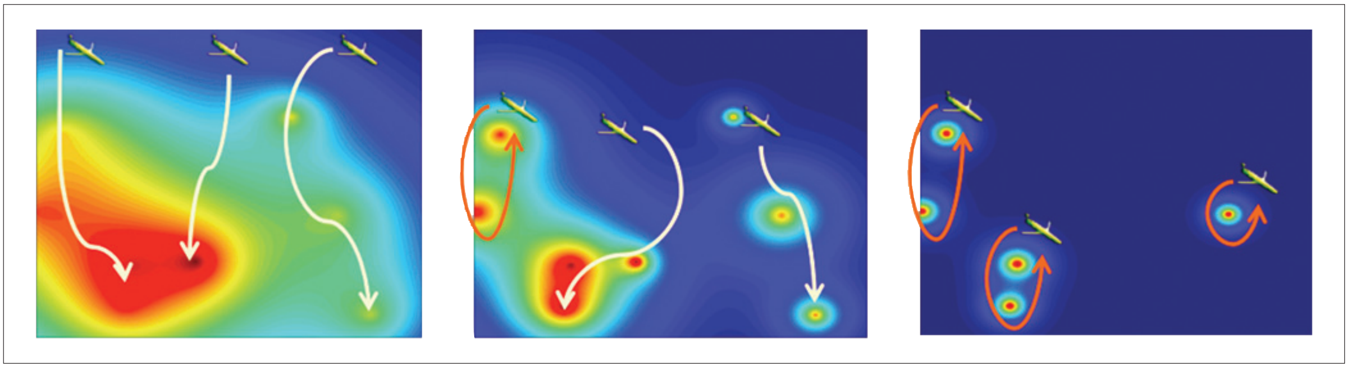


Figure 2. Exploration-exploitation at sea: autonomous agents first explore in a coordinated way to identify phenomena of interest (left); as the field evolves, the AUV network differentiates to allow individual agents to follow particular targets (exploitation).

same time, it should be noted that currents can sometimes be exploited opportunistically, since energy consumption is a major concern from the point of view of propulsion, acoustic communications, and hotel load. Aerial vehicles face similar energy issues, but usually have two key advantages over underwater systems: low-power communication over an RF channel with other vehicles or a base station, and the ability to measure their own location with precision using GPS. Underwater positioning methods are often acoustic and subject to the challenges noted above for acoustic communications. Today, ocean vehicles typically break the surface for precision localization via GPS — a hazardous and time-consuming operation.

As one specific example of these challenges, we carried out an experiment with an autonomous surface vehicle (ASV) station-keeping via commands sent through an acoustic modem. The effects of wind in such a setting are to both push the boat off the reference point and degrade the acoustic channel because of surface chop. Indeed, we observed a full order of magnitude increase in positioning error as the wind speed increased from 1 to 8 m/s [4]. Another experiment coordinating vehicles through acoustic communication is detailed in [5]. Overall, the costs of actuation, communication, and sensing can be comparable for marine systems [3], and this parity strongly affects how underwater acoustic networks are to be optimized in support of a mission goal [1].

Sparse Approximation and Matrix Completion Review

Here we review key elements of sparse approximation and matrix completion needed for understanding the methods presented in this article.

Any family of signals with a linear algebraic representation requiring many more parameters than the actual number of unknowns in a given instance is called a low-dimensional signal. Two low-dimensional signals that have been the subject of active research are sparse vectors and low-rank matrices.

Considering sparse vectors first, let $\Psi \in \mathbb{R}^{n \times n}$ be an orthonormal basis so that any signal $f \in \mathbb{R}^n$ can be written as $f = \Psi x$ for a coefficient vector $x \in \mathbb{R}^n$. x is called *s-sparse* if it has at most s nonzero elements, and if a sparse x suffices, the signal f is said to admit a sparse representation in Ψ. Intuitively, an s-sparse signal has only s unknowns, many fewer than the full representation [6]. f is called compressible if it is well approximated — but not constructed exactly — by a sparse x. Many natural phenomena are considered to be compressible in domains such as frequency, wavelets, or total variation. For low-rank matrices, a key classical result is that any $m \times n$ matrix X admits the singular value decomposition $X = U\Sigma V$, and that the number of nonzero elements in Σ denotes the rank of X. The Eckart-Young-Mirsky result says further that reduced-rank approximations to Σ, obtained simply by keeping only the r largest elements in Σ, are optimal in the Frobenius norm [7]. Low-rank matrices arise frequently in second-order datasets (Netflix) and data generated from bilinear observations [8].

Incoherence is a critical requirement for efficiently sampling low-dimensional models, and measures the dissimilarity between the signal basis and the measurement basis. Information measured in one basis will "spread out" when measured in a second (incoherent) basis, enabling efficient reconstruction even if few measurements are made. In point-wise compressed sensing, the measurement $y = R\Phi f$ involves a "fat" matrix $R\Phi$, where R is a fat projection (a row-reduced identity matrix), and Φ an orthonormal measurement basis. Coherence of the measurement and signal bases is defined as

$$\mu(\Phi,\Psi)=\sqrt{n}\max_{1 \le j,k \le n}\left|<\phi_j,\psi_k>\right|,$$

where ϕ_j and ψ_k denote the j^{th} and kth atoms of Φ and Ψ, respectively, and $<\cdot,\cdot>$ denotes an inner product. For a 2D field represented in the discrete cosine basis, and sampled with delta functions, $\mu(\Phi, \Psi) = 2$, a highly incoherent basis pairing is indicated. The reconstruction task can be formulated as a convex optimization [6]

$$\hat{x} = \text{argmin } ||z||_1 \text{ subject to } y = R\Phi\Psi z,$$

where $||\cdot||_1$ is the one-norm. Pulling together sparsity and incoherence, a key theorem of compressed sensing states that if f is s-sparse in Ψ, then $\hat{x} = x$ with high probability if the number of measurements (i.e., the number of rows in R) exceeds $C\mu^2(\Phi, \Psi)s \log n$ for some positive con-

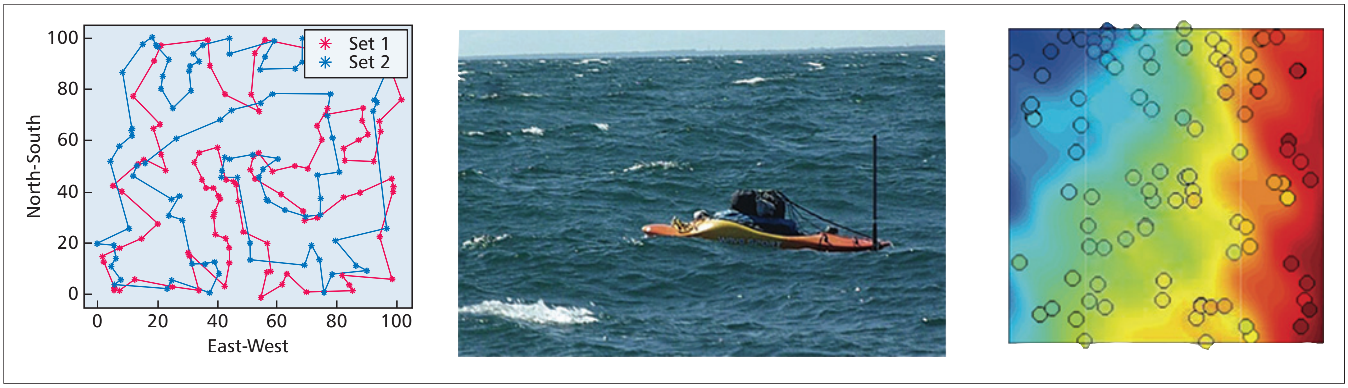


Figure 3. Kinematically-constrained sparse approximation in action at sea. The left image shows two "random TSP" paths generated and executed by an autonomous surface vehicle (center); the right image shows a depth-field reconstruction based on sparse samples and the discrete-cosine transform.

stant C. Coherence, and thus the associated sampling rate, determine the feasibility of the reconstruction.

The measurement ideology in low-rank matrix completion bears a strong resemblance to that of compressed sensing. Here $y = \Omega(X)$, where Ω denotes a small set of indices, and hence the measurement basis is the set of delta functions over the vector space of $m \times n$ matrices. Incoherence is defined as the sum of the inner products, squared, between delta functions and the left and right singular vectors [7]. The reconstruction is formulated as the convex problem

$$\hat{X} = \text{argmin } ||Z||_* \text{ subject to } y = \Omega(Z),$$

where $||\cdot||_*$ is the sum of singular values. Akin to compressed sensing, if Ω is sampled uniformly at random, $\hat{X} = X$ with high probability when the number of measurements exceeds $Cv(X)rn \log^2 n$, where r is the rank of X, and $n > m$.

Turning to the selection of points via R, random sampling gives strong stability guarantees for both compressed sensing and low-rank matrix completion in the presence of measurement noise as well as model mismatch. Mismatch occurs when the unknown signal is compressible instead of sparse, or is approximately low-rank instead of exactly low-rank. In compressed sensing, we appeal to the restricted isometry property (RIP) of the measurement matrix of order s, by which the measurement process approximately preserves pairwise distances between s-sparse vectors. In low-rank matrix completion, a weaker condition called *restricted strong convexity* is desirable, and similarly sufficient to show stability, given $r \ll n$ [9].

Random sampling may be infeasible or unsuitable, and it is desired in such cases to find new projection matrices that:

- Result in efficient reconstruction
- Can be implemented efficiently by the sampling system (e.g., mobile robots in a large domain)

In general, it is not easy to show that structured projections satisfy the RIP, although there have been recent efforts directed toward the design of deterministic projection matrices. These approaches control the structure of the sensing matrix $R\Phi\Psi$; however, in our case we can only control the measurement matrix $R\Phi$. To the best of our knowledge, there are no techniques for structured random construction of the measurement matrix, especially given the motion constraints of robots.

Field Reconstruction with Mobile Sensors via KCSA

Since kinematic constraints are one of our main divergences from standard sparse approximation, a natural step is the construction of a cost related to resources. We can identify three key aspects of such a cost:

- The Euclidean length of a transit between two sample sites
- A heading change (maneuvering) required to move between two successive sites
- The resources required for an individual sample itself

Working with a large sea-surface temperature dataset and the discrete cosine transform basis, we optimized statistics of randomly drawn transit lengths and maneuvers to minimize the sparse approximation reconstruction error in a simple random walk scheme [10]. This optimized random walk, surprisingly, was outperformed in sparse approximation error by an even simpler procedure of randomly choosing a set of sites and connecting them with a nearest-neighbor traveling salesperson problem (TSP) solution. Figure 3 depicts an experimental test with this method. More generally, reconstruction errors can be bounded by ensuring that the RIP in the sampling basis is achieved, suggesting a second class of lightweight KCSA methods. For a given dictionary and a set of sample points, restricted isometry is trivial to compute incrementally, so a vehicle can be instructed to move to randomly selected, as yet unvisited points based simply on what the next point will do to the RIP. This is a purely greedy approach that typically offers an improvement of 10–20 percent relative to points selected without considering the RIP [11].

Maintaining restricted isometry and minimizing path length could be tackled as well in a more formal optimization setting. With the spike basis, however, the projection (sampling) design is a binary programming problem. Furthermore,

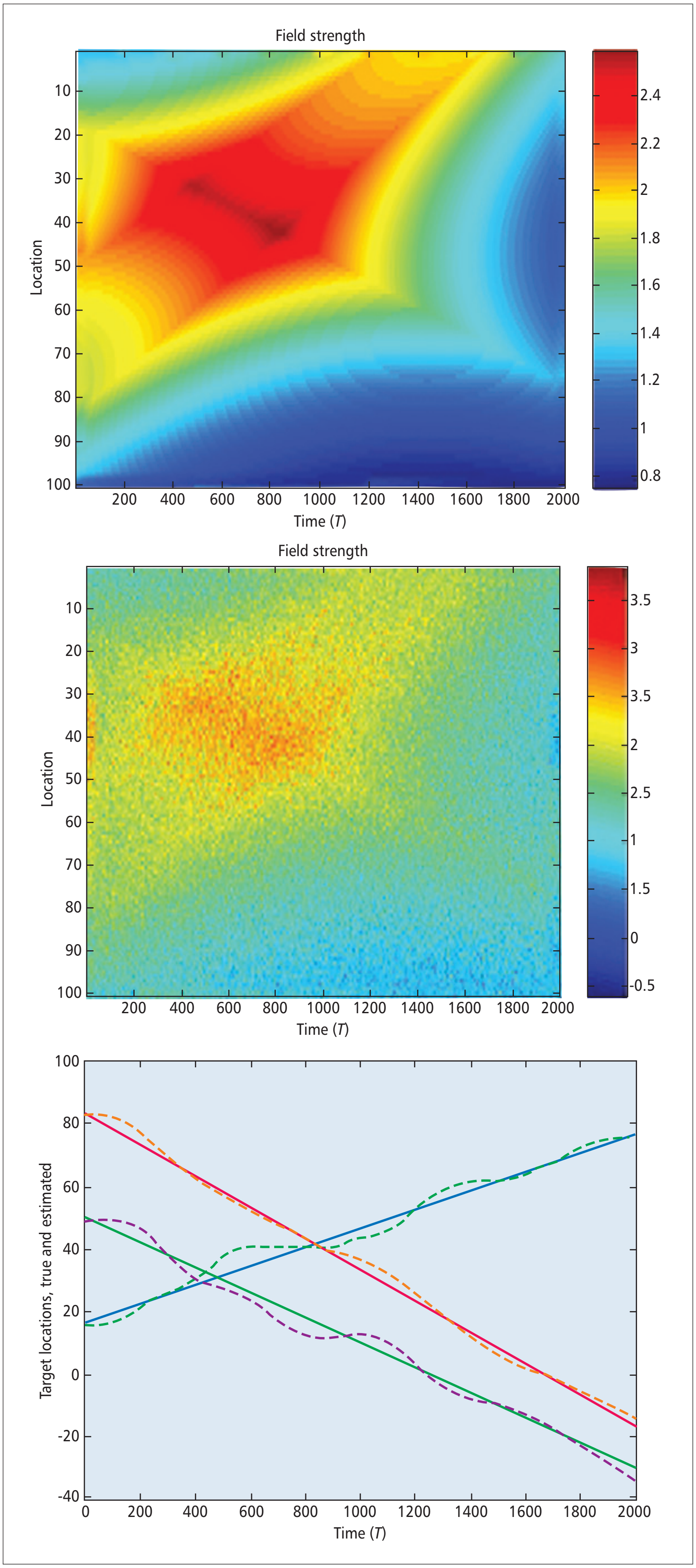


Figure 4. Target tracking without field reconstruction for three mobile sources. Instaneous observations of the noise-free (top) and noisy (center) fields; (bottom) true and estimated paths.

we have found that when the number of samples is small, there is virtually no advantage in the designed plan over random points; with a larger number of samples, the designed points have a coherence typically 20–30 percent lower than the random points. The formal sampling problem can also be merged with the TSP, but complexity scaling is very poor since they are not well coupled. An alternative is to generate a pool of optimal sample sets without attention to the path, and then apply strong TSP solvers to members of the pool. Applying such a process, we find that a pool typically identifies a Pareto-like frontier of TSP cost vs. coverage (e.g., as characterized by the star discrepancy), and thus a trade-off space between sampling efficacy for the purpose of reconstruction and the path length. This multi-way trade-off is developed further later.

Field Reconstruction in an Underwater Sensor Network

As noted earlier, in underwater acoustic systems, key physical constraints include limited bandwidth and the high rate of packet loss. Moreover, synchronization is difficult, often mandating a design that does not rely on accurate timing. To address these issues, we draw on the sparse (compressible) nature of the sensing field, through a design that capitalizes on random channel access and compressive sensing (RACS), using multiple sensors interacting with a fusion center. In this approach, a sensor takes local field measurements at random instants in time [12]. Each measurement is encoded into a data packet, which is immediately transmitted to the fusion center. The fusion center collects all the packets over a certain interval of time, and then attempts to reconstruct the field.

The key (and perhaps counterintuive) idea behind combining random access and compressive sensing is that the fusion center does not care which sensors provide the data packets as long as the sensors are selected uniformly at random, and there are sufficient packets. With random access, packet collisions are inevitable. However, they occur randomly and thus do not alter the way in which the FC perceives packet arrivals. Consequently, the fusion center can simply discard the packets that have collided (as well as those that were received in error due to noise, fading or interference), and perform field reconstruction using the remaining useful packets. Because of the random nature of this architecture, a probabilistic approach to system design becomes necessary. Our approach relies on the notion of sufficient sensing probability, the probability with which full field reconstruction is achieved in a certain interval of time. Setting this probability to a desired target value, system optimization under a relevant criterion (e.g., minimum energy per bit or minimum reconstruction time) yields the necessary parameters: the packet generation rate (per-node sensing rate), required bandwidth, and power. Many fields of interest are sparse with respect to fairly general dictionaries; thus, the reconstruction method is relatively non-parametric.

The concept of random access compressive sensing can also be extended to the tracking of vehicles each of which emanates a particular signature [13]. Based on the knowledge of signatures, a reconstruction algorithm is designed that side-steps the field reconstruction, and instead focuses directly on object tracking. Figure 4 provides an example of the multi-vehicle tracking capabilities of RACS. Clearly, there is correlation in the map data (i.e., the field is sparse). This particular spatio-temporal field is created by three moving objects, each having an exponentially decaying signature; the level of sparsity is three. Shown on the right are the objects' true trajectories (solid) and estimates obtained by the tracking algorithm (dashed), applied to samples of the noisy field (–20 dB signal-to-noise ratio, SNR, 30 percent packet loss) obtained from 10 sensors distributed uniformly along a track. As with field reconstruction, effective target tracking based on RACS does not require complex models of vehicle behavior (vs. field behavior). As such, our methods are highly robust.

Due to the time-varying nature of the field, one needs to optimize the sensing duration — the key is to acquire a sufficient number of samples for reliable reconstruction, while ensuring that the sensing duration is short enough so that the observed field does not change significantly within a single sensing period. From the perspective of target detection using the acquired image, there is another notion of sensing rate optimality corresponding to the trade-off between detection accuracy and energy efficiency. One must also consider the rate at which the field is changing due to the speed of the vehicles to be tracked. To address these issues, a feedback scheme can be employed, resulting in adaptive sensing. The feedback signal is generated by performing target detection on the acquired field, which is then compared to a model of the field, resulting in a model-based error. Figure 5 depicts the evolution of the sensing rate (T) based on such an algorithm. If there is a large model-based error, the sensing rate is increased (state A). To compensate for too high a sensing rate, a motion-dependent error measure is also computed (state D). If the vehicles move slowly, the fusion center can respond by decreasing the sensing rate to conserve energy. Thresholds on the two error measures are optimized to maximize performance while constraining the sensing energy. States B and C correspond to conflicts in the decisions resulting from the two different sensing metrics.

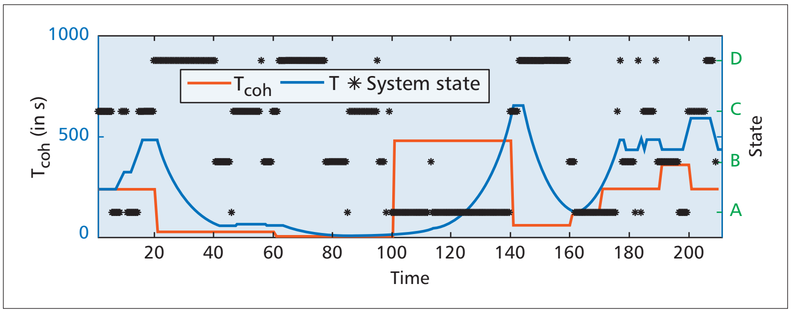


Figure 5. Controlled sensing rate T using the proposed method; states A and D correspond to undersensing and oversensing, respectively; states B and C correspond to the cases when the model error and sensing error metrics lead to conflicting decisions. (T_{coh} is the coherence time of the time-varying field).

Tackling the Integrated Exploration-Exploitation Problem

Our prior work on field reconstruction shows that kinematic and communication constraints can readily be employed to develop practical reconstruction methods exploiting new results in sparse approximation. Furthermore, a large family of fields admits the needed sparsity. We can extend these notions to the exploration-exploitation problem. As a concrete example of integrated search with sparse methods, we consider an underwater target that emits a field of observable intensity, as in Fig. 6a; close to the target, the intensity is high, but the intensity decays as we move away from the target location. The aggregate field shown here is a side-scan sonar image, and the target signature could be due to an acoustic signal from a flight data recorder lost in some sector of the ocean. A deployed AUV is restricted to collecting a few samples along some navigation path (Fig. 6b), with the hope that these samples contain enough information to localize the target to a smaller region of interest (Fig. 6c). The intensity field in Fig. 6a is approximately separable, however, and therefore admits a low-rank matrix structure with respect to the image matrix. This property allows the reduction of the number of samples required for target localization [14], subsequently leading to reduced cost of navigation [15]. On the other hand, a decaying low-rank target field is also a very generic structural assumption not specific to this particular dataset. Thus, a rich set of target signatures meet the two key assumptions; additionally, we can extend our results to allow for multiple targets, different adaptive and non-adaptive sampling approaches, and even distributed implementations using multiple AUVs with communication costs. With our approach, we do not need to know the target signature *a priori*.

We developed a new methodology for this problem based on low rank matrix completion and the high-dimensional properties of the traveling salesperson problem, and one appealing outcome is a characterization of a multi-way trade-off between sampling, navigation, computation, and localization [15]. Our algorithm takes a hierarchical exploration-exploitation approach that combines the strengths of low-rank matrix completion and binary search, viz. noise robustness and computational speedup. In each stage, based on the current field of view, a set of random locations is selected; the AUV determines an optimized path covering these locations, at which sample observations are collected. The field is then reconstructed using low-rank matrix completion [7], computing the best rank one approximation. After post-processing, the target is localized, a smaller field of view is determined, and the steps are then repeated. We assume that the algorithm terminates after *k*

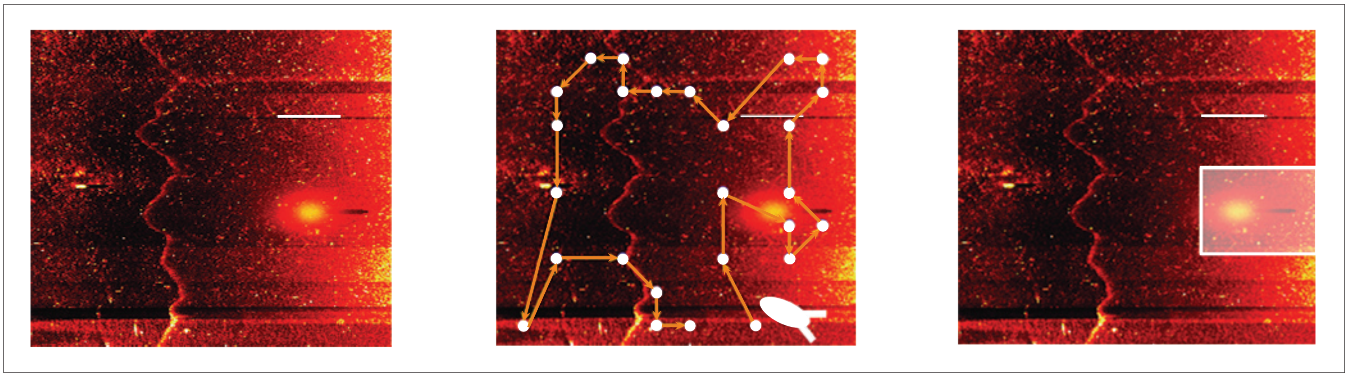


Figure 6. (left) Original intensity map of an underwater target (sonar data); (center)initial sampling path of the AUV; (right) the localized region of interest after several stages of the matrix completion exploration-exploitation algorithm.

iterations, define the total search space of size $n \times n$, and the field of view to be of size $m \times m$ representing a uniformly decimated sub-matrix of the total search space. With these definitions, theoretical results on low rank matrix completion and Euclidean TSP can be used to show that our approach takes $O(km \log^2 m)$ samples and $O(\mathrm{km}^{3.5})$ computational time and achieves $O(k\sqrt{m} \log m)$ cost of navigation with high probability, where the number of iterations increases with noise level and satisfies $\log_m n \leq k \leq n/m$.

A design parameter is the field of view. Intuitively, the size of the field of view is chosen inversely proportional to the spread in the target field since a larger spread enables collecting sample intensities above the noise floor farther from the target location, thus reducing sampling requirements. Decaying separable fields admit good rank-one approximation via singular vectors that are also decaying. The decaying property of the singular vectors can be used for further processing via unimodal function fitting (a form of isotonic regression), which results in greater noise reduction and improved target localization. Numerical results show that there is an optimum spreading factor leading to the best detection performance. One can additionally characterize the best balance between existence of a significant gradient in the signature (after noise corruption) and the target signature occupying a larger footprint; that is, fields with large support of the gradient show better performance for a fixed sample complexity. We can also provide a theoretical analysis that provides the trade-off between our ability to localize the target within a stage and our ability to reconstruct the current windowed field. Finally, the overall multi-way trade-off can be characterized by including navigation and kinematic costs (or constraints) as well as the incorporation of the randomness induced by packet dropping in the underwater acoustic communication channel. Thus, the features of our field reconstruction, tracking, and exploration-exploitation work can be seamlessly combined.

Conclusions

We have argued that mobile underwater sensor networks enabled with control and acoustic communication will play a major role in future ocean observing infrastructure. Such autonomous systems need novel mechanisms for handling underwater communication constraints, and need to integrate sensing and classification in providing solutions for key exploration-exploitation trade-offs. The unavoidable burden of limited resources in these operations (e.g., physical energy, time, communications, and accurate synchronization) are in a sense balanced by the promise of compressed sensing and low-rank approximation theory over huge spatial domains, coupled with standard tools including solvers for the traveling salesperson problem. Thus, despite the paucity of resources and limited number of deployed sensors, whether they be static or mobile, the underlying sparsity (or low rank property) of the processes of interest enables us to overcome the challenge of being data-starved. Our structures of interest are general and do not require complex a priori model information for our fields or targets. We have specifically considered constraints inherent to our underwater system: the kinematic constraints of vehicles as well as the unique characteristics of the underwater acoustic channel. These constraints have led to novel problem formulations and attendant solutions. In fact, some of our most recent outcomes can bypass weaknesses of more classical approaches to solving the exploitation-exploration problem that are not flexible to incorporating multiple resource constraints such as navigation, communication, and sensing while allowing for theoretical analysis. The successful implementation of these methods at sea will have an impact on many applications and industries, including environmental monitoring, aquatic ecosystems, ocean accident remediation, surveillance for defense applications, homeland security, the oil and gas industry, aquaculture, geological and oceanographic science, and marine biology.

References

[1] G. Hollinger *et al.*, "Underwater Data Collection Using Robotic Sensor Networks," *IEEE JSAC*, Special Issue on Communications Challenges and Dynamics for Unmanned Autonomous Vehicles, vol. 30, no. 5, June 2012, pp. 899–911.

[2] C. Berger *et al.*, "Sparse Channel Estimation for Multicarrier Underwater Acoustic Communication: From Subspace Methods to Compressed Sensing," *IEEE Trans. Signal Processing*, vol. 58 no. 3, Mar. 2010, pp. 1708–21.

[3] M. Stojanovic and J. Preisig, "Underwater Acoustic Communication Channels: Propagation Models and Statistical Characterization," *IEEE Commun. Mag.*, vol. 47, no. 1, Jan. 2009, pp. 84–89.

[4] F. Arrichiello *et al.*, "Effects of Underwater Communication Constraints on the Control of Marine Robot Teams," *Proc. IEEE Int'l. Conf. Robot Commun. and Coordination*, Odense, Denmark, Mar. 2009.

[5] B. Reed *et al.*, "Multi-Vehicle Dynamic Pursuit Using Underwater Acoustics," *Proc. Int'l. Symp. Robotics Research*, Dec. 2013, Singapore.

[6] E. Candes and T. Tau, "Decoding by Linear Programming," *IEEE Trans. Info. Theory*, vol. 51, no. 125, Dec. 2005, pp. 4203–21.

[7] D. Gross, "Recovering Low-Rank Matrices From Few Coefficients in Any Basis," *IEEE Trans. Info. Theory*, vol. 57, no. 3, Mar. 2011, pp.1548–66.

[8] S. Choudhary and U. Mitra, "On Identifiability in Bilinear Inverse Problems," *Proc. IEEE Int'l. Conf. Acoustics, Speech and Signal Processing*, May 2013, Vancouver, Canada, pp. 4325–29.

[9] S. Negahban and M. J. Wainwright, "Restricted Strong Convexity and Weighted Matrix Completion: Optimal Bounds with Noise," *J. Machine Learning Research*, vol. 13, no. 1, Jan. 2012, pp. 1665–97.

[10] R. Hummel *et al.*, "Mission Design for Compressive Sensing with Mobile Robots," *Proc. Int'l. Conf. Robotics and Automation*, May 2011, Shanghai China, pp. 2362–67.

[11] F. Hover *et al.*, "One-step-Ahead Kinematic Compressive Sensing," *Proc. IEEE GLOBECOM Wi-UAV Wksp.*, Dec. 2011, Houston, TX, pp. 1314–19.

[12] F. Fazel, M. Fazel, and M. Stojanovic, "Random Access Compressed Sensing over Fading and Noisy Communication Channels," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, May 2013, pp. 2114–25.

[13] K. Kerse, F. Fazel, and M. Stojanovic, "Target Localization and Tracking in a Random Access Sensor Network," invited paper, *47th Asilomar Conf. Signals, Systems and Computers*, Nov. 2013, Pacific Grove, CA, pp. 103–07.

[14] S. Choudhary *et al.*, "Active Target Detection with Mobile Agents," IEEE Int'l. Conf. Acoustics, Speech and Signal Processing, May 2014, Florence, Italy, pp. 4185–89.

[15] S. Choudhary *et al.*, "Active Target Detection with Navigation Costs: A Randomized Benchmark," invited paper, *52nd Annual Allerton Conference on Commun., Control, and Computing*,Oct. 2014, Monticello, IL, pp. 109–15.

Biographies

Urbashi Mitra [S'88, M'98, SM'04, F'07] received B.S. and M.S. degrees from the University of California at Berkeley and her Ph.D. from Princeton University. Prior to her Ph.D. studies, she was a member of technical staff at Bellcore. After a six-year stint at the Ohio State University (OSU), she joined the Department of Electrical Engineering at the University of Southern California (USC), Los Angeles, where she is currently a professor. She is the inaugural Editor-in-Chief of *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*. She is a Distinguished Lecturere for the IEEE Communications Society for 2015–2016. She was a member of the IEEE Information Theory Society's Board of Governors (2002–2007, 2012–2014) and the IEEE Signal Processing Society's Technical Committee on Signal Processing for Communications and Networks (2012–2014). She is the recipient of: a 2015 Insight Magazine STEM Diversity Award, 2012 GLOBECOM Signal Processing for Communications Symposium Best Paper Award, 2012 U.S. National Academy of Engineering Lillian Gilbreth Lectureship, USC Center for Excellence in Research Fellowship (2010–2013), the 2009 DCOSS Applications & Systems Best Paper Award, Texas Instruments Visiting Professor (Fall 2002, Rice University), 2001 Okawa Foundation Award, 2000 OSU College of Engineering Lumley Award for Research, 1997 OSU College of Engineering MacQuigg Award for Teaching, and a 1996 National Science Foundation CAREER Award. She currently serves on the following IEEE award committees: Fourier Award for Signal Processing, James H. Mulligan, Jr. Education Medal, and the Paper Prize Award. She has been/is an Associate Editor for *IEEE Transactions on Signal Processing* (2012–), *IEEE Transactions on Information Theory* (2007–2011), *IEEE Journal of Oceanic Engineering* (2006–2011), and *IEEE Transactions on Communications* (1996–2001). She has co-chaired the Technical Program Committees of the 2014 IEEE International Symposium on Information Theory in Honolulu, Hawaii, 2014 IEEE Information Theory Workshop in Hobart, Tasmania, the IEEE 2012 International Conference on Signal Processing and Communications, Bangalore India, and the IEEE Communication Theory Symposium at ICC 2003 in Anchorage, Alaska. She served as General Co-Chair of the first ACM Workshop on Underwater Networks at Mobicom 2006, Los Angeles, California. She was Tutorials Chair for IEEE ISIT 2007 in Nice, France, and Finance Chair for IEEE ICASSP 2008 in Las Vegas, Nevada. She has held visiting appointments at the Delft University of Technology, Stanford University, Rice University, and the Eurecom Institute. She served as co-Director of the Communication Sciences Institute at USC from 2004 to 2007. Her research interests are in wireless communications, communication and sensor networks, underwater acoustic communication, biological communication systems, detection and estimation, and the interface of communication, sensing, and control.

Sunav Choudhary is a Ph.D. candidate in the Ming Hsieh Department of Electrical Engineering at USC. He received his B.S. degree in 2010 from the Indian Institute of Technology, Kharagpur. He is the recipient of the following awards and honors: 2015 USC Center for Applied Mathematical Sciences Prize Winner, 2014 IEEE International Symposium on Information Theory Student Travel Grant, 2013 Ming Hsieh Institute Electrical Engineering Research Festival Best Poster Award, 2010–2014 USC Annenberg Graduate Fellowship, 2006–2010 IIT Kharagpur Goralal Syngal Memorial Scholarship, and the 20016–2010 CBSE Merit Scholarship for Professional Studies. His present research interests are in the field of sparse signal approximation and its applications to underwater acoustic communications.

Franz S. Hover [M'92] received his B.S.M.E. degree from Ohio Northern University, and his M.S. and Sc.D. degrees from the Joint Program in Applied Ocean Physics and Engineering at the Woods Hole Oceanographic Institution and Massachusetts Institute of Technology (MIT), Cambridge. He was a consultant to industry and then a member of the research staff at MIT, where he worked in fluid mechanics, biomimetics, and marine robotics. His research has led to commercial development of the HAUV platform for autonomous ship hull inspection, advances in computational tools for power systems, and innovations in subsea flow control technology. He is currently an associate professor at the MIT Department of Mechanical Engineering, with research focusing on design methodologies and complex marine systems. He has authored or co-authored over 100 refereed papers.

Robert Hummel [S'11] received his B.S. and M.S. degrees in Mechanical Engineering from MIT in 2009 and 2012, respectively, with a focus on robotics, mechatronics, and autonomous sampling techniques. He experience includes designing and building autonomous vehicles, machines, and mechanisms, and various electro-mechanical systems. He has also worked with numerical simulations and optimization. Currently, he is co-founder and design engineer at ATSE in Northborough, Massachusetts, an engineering consulting firm specializing in electric motors, generators, and drives for industrial, defense, and renewable energy applications.

Naveen Kumar [S'10] received his B.Tech. degree in instrumentation engineering from the Indian Institute of Technology, Kharagpur, in 2009. He is currently working toward his Ph.D. degree at the Department of Electrical Engineering, USC. His research interests include machine learning and signal/image processing for applications to speech and multimedia problems. He was awarded the Viterbi School Dean's Doctoral Fellowship at USC in 2009. He has also worked on teams that have won the Interspeech 2012 Speaker Trait Challenge and the Northern Digital Inc. Excellence Awards at the 2014 International Seminar on Speech Production (ISSP).

Shrikanth S. Narayanan [S'88, M'95, SM'02, F'09] received his M.S., Engineer, and Ph.D. degrees in electrical engineering from the University of California Los Angeles in 1990, 1992, and 1995, respectively. From 1995 to 2000 he was with AT&T Labs-Research, Florham Park, New Jersey, and AT&T Bell Labs,Murray Hill, New Jersey, first as a senior member and later as a principal member of technical staff. He is currently a professor of engineering at USC, where he directs the Signal Analysis and Interpretation Laboratory. He holds appointments as a professor of electrical engineering, computer science, linguistics, and psychology, and is the founding director of the Ming Hsieh Institute, Ming

The successful implementation of these methods at sea will have an impact on many applications and industries, including environmental monitoring, aquatic eco-systems, ocean accident remediation, surveillance for defense applications, homeland security, the oil and gas industry, aquaculture, geological and oceanographic science, and marine biology.

Hsieh Department of Electrical Engineering, USC. He has published more than 600 papers and has been granted 16 U.S. patents. His research interests include human-centered signal and information processing and systems modeling with an interdisciplinary emphasis on speech, audio, language, multimodal and biomedical problems, and applications with direct societal relevance. He is a Fellow of the Acoustical Society of America and the American Association for the Advancement of Science, and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is also an Editor of the *Computer Speech and Language Journal* and an Associate Editor of *IEEE Transactions on Affective Computing*, *APSIPA Transactions on Signal and Information Processing*, and the *Journal of the Acoustical Society of America*. He was also previously an Associate Editor of *IEEE Transactions on Speech and Audio Proseccing* (2000–2004), *IEEE Signal Processing Magazin*e (2005– 2008), and *IEEE Transactions on Multimedia* (2008–2011). He has received a number of honors, including Best Transactions Paper awards from the IEEE Signal Processing Society in 2005 (with A. Potamianos) and 2009 (with C. M. Lee), and selection as an IEEE Signal Processing Society Distinguished Lecturer for 2010–2011. Papers coauthored with his students have won awards at Interspeech 2014 Cognitive and Physical Load Challenge, Interspeech 2013 Social Signal Challenge, Interspeech 2012 Speaker Trait Challenge, Interspeech 2011 Speaker State Challenge, Interspeech 2013 and 2010, Interspeech 2009 Emotion Challenge, IEEE DCOSS 2009, IEEE MMSP 2007, IEEE MMSP 2006, ICASSP 2005, and ICSLP 2002.

MILICA STOJANOVIC [(SM'08, F'10] graduated from the University of Belgrade, Serbia, in 1988, and received M.S. ('91) and Ph.D. ('93) degrees in electrical engineering from Northeastern University, Boston, Massachusetts. She was a principal scientist at MIT, and in 2008 joined Northeastern University, where she is currently a professor of electrical and computer engineering. She is also a guest investigator at the Woods Hole Oceanographic Institution and a visiting scientist at MIT. Dr. Stojanovic is the recipient of the 2015 IEEE/OES Distinguished Technical Achievement Award. Her research interests include digital communications theory, statistical signal processing and wireless networks, and their applications to underwater acoustic systems. She is an Associate Editor for the *IEEE Journal of Oceanic Engineering* and a past Associate Editor for *IEEE Transactions on Signal Processing* and *IEEE Transactions on Vehicular Technology*. She also serves on the Advisory Board of *IEEE Communication Letters*, and chairs the IEEE Ocean Engineering Society's Technical Committee for Underwater Communication, Navigation and Positioning.

GAURAV S. SUKHATME [SM'05, F'11] is a professor of computer science at USC. He received his undergraduate education at IIT Bombay in computer science and engineering, and M.S. and Ph.D. degrees in computer science from USC. He is co-director of the USC Robotics Research Laboratory and director of the USC Robotic Embedded Systems Laboratory. His research interests are in robot networks with applications to environmental monitoring. He has served as Principal Investigator (PI) on numerous NSF, DARPA, and NASA grants. He is a co-PI on the Center for Embedded Networked Sensing, an NSF Science and Technology Center. He is a recipient of the NSF CAREER award and the Okawa Foundation research award. He is one of the founders of the Robotics: Science and Systems conference. He was Program Chair of the 2008 IEEE International Conference on Robotics and Automation and the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. He is Editor-in-Chief of *Autonomous Robots* and has served as Associate Editor of *IEEE Transactions on Robotics and Automation* and *IEEE Transactions on Mobile Computing*, and is on the Editorial Board of *IEEE Pervasive Computing*.

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, 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 behaviours 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 & 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 web site 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

Scheduled Publication Dates: Twice per year, May and November

Series Editors

Jinsong Wu, Alcatel-Lucent, China, wujs@ieee.org
John Thompson, University of Edinburgh, UK, john.thompson@ed.ac.uk
Honggang Zhang, UEB/Supelec, France; Zhejiang Univ., China, honggangzhang@zju.edu.cn
Daniel C. Kilper, University of Arizona, USA, dkilper@optics.arizona.edu

Underwater Sensor Networks: A New Challenge for Opportunistic Routing Protocols

Amir Darehshoorzadeh and Azzedine Boukerche

The authors are with the University of Ottawa.

ABSTRACT

Opportunistic routing (OR) is a promising paradigm that selects the next-hop forwarder on the fly. OR has gained a lot of attention from the research community for its ability to increase the performance of wireless networks. In OR a potential group of nodes (candidates) is selected to help as the next-hop forwarder. Each candidate that receives the packet can continue forwarding the packet. In OR, by using a dynamic relay node to forward the packet, the transmission reliability and network throughput are increased. Underwater sensor networks (UWSNs) collect data from the environment and transfer them to the sonobuoys on the surface to send them to a center for further processing. Because of the acoustic channels common to UWSNs, they have low bandwidth, high error probability, and longer propagation delay compared to radio channels. These properties of UWSNs make them good potential candidates for using OR concepts to deliver packets to the destination. This article reviews and compares different OR protocols proposed for UWSNs. We classify the existing approaches in different categories, discuss representative examples for each class of protocols, and uncover the requirements considered by the different protocols, as well as the design requirements and limitations under which they operate. Finally, we discuss potential future research directions for UWSNs using the OR paradigm.

INTRODUCTION

Recently, underwater sensor networks (UWSNs) [1–3] have received a great deal of attention from the wireless communication and networking communities. Applications for UWSNs include offshore exploration, ocean monitoring, marine and wildlife studies, and controlling underwater mineral extraction. The underwater sensor nodes collect data from the environment and transfer it to sonobuoys on the surface, which send it back to a center for further processing. While there are some similarities between UWSNs and terrestrial sensor networks, UWSNs exhibit some unique particularities, such as acoustic communication, limited energy and data storage, large number of deployed sensors, low bandwidth, high latency, and high bit error rate, all characteristics of UWSNs that differentiate them from terrestrial sensor networks.

Radio frequency (RF), common in the physical layer technology of terrestrial sensor networks, is not a feasible technology for UWSNs since they face a rapid attenuation of radio signals. Therefore, acoustic channels are used in UWSNs for communication purposes. Acoustic channels have lower bandwidth and longer propagation delays compared to radio channels [1]. UWSNs therefore face an unreliable environment compared to wireless sensor networks (WSNs). Not only do acoustic channels have limited bandwidth due to the water current, but the underwater sensor nodes are highly dynamic, resulting in a network topology that is also dynamic. Furthermore, because acoustic channels are affected by path loss, noise, and Doppler spread, the links between nodes in UWSNs are highly error-prone [4]. Low channel bandwidth, high latency, and high packet loss probability lead to increased retransmissions and high energy consumption. Because UWSNs have these properties, the design of routing protocols for these types of networks is particularly challenging. Minimizing the number of packet transmissions and increasing packet progress while a simple and fast routing protocol is applied results in better performance in these networks. Many routing protocols have been proposed for terrestrial wireless sensor networks, but they are not applicable to UWSNs due to the unique characteristics of acoustic channels [2].

Opportunistic routing (OR) [5, 6] is proposed to take advantage of the broadcast nature of the wireless medium. When sending a packet from a source to a specific destination in traditional routing protocols, one node is selected as the next-hop forwarder, and all packets should be forwarded to the destination using the designated next hop [7]. As we can see in the traditional

0163-6804/15/$25.00 © 2015 IEEE

routing approach, if a node other than the designated next hop receives the packet, it is not allowed to continue forwarding the packet. In contrast to traditional routing protocols, OR selects the next-hop forwarder on the fly by taking advantage of the broadcast nature of the wireless medium. A set of potential next-hop forwarders (candidates), usually referred to as the candidate set (CS), is selected. Priority for each selected candidate is determined according to their ability to act as the next-hop forwarder. Different metrics are used in OR protocols to assign priority (see [6] for more details). The highest priority is given to the candidate that can reach the destination at the lowest cost.

By using OR, each candidate can act as a potential next-hop forwarder. Therefore, the sender puts the CS in the header of a data packet and broadcasts it. The candidates that have successfully received the packet coordinate with each other to decide which has a higher priority and should forward the packet. The candidate with the highest priority that has received the packet continues forwarding the packet, and the others discard it. This process of deciding which candidate should forward and which should discard the packet is usually referred to as *candidate coordination*.

As explained above, in environments that have unreliable transmissions due to various factors such as interference, OR can be a good solution by selecting the next node forwarder on the fly. Furthermore, OR combines weak links and creates a virtual strong link, which can increase path reliability. Therefore, UWSNs, which transmit packets in an unreliable environment, are a good potential candidate to use the OR concept. The main objectives of OR are to increase path reliability by creating a virtual strong link and minimizing the number of retransmissions needed to deliver a packet from a source to a given destination. Therefore, by using OR in UWSNs, the packet can progress toward its destination using multiple potential candidates that provide better path to the destination, which results in fewer retransmissions. Furthermore, lower numbers of transmissions result in energy conservation, which is one of the main challenges in UWSNs.

In this article, a comprehensive review of OR protocols used in UWSNs is presented. We review the important issues of using OR in different kinds of networks. We classify the proposed OR protocols in UWSNs into different categories. Based on our classification, we explain each OR protocol in UWSNs in greater detail, emphasizing its advantages and disadvantages. We summarize and compare all explained protocols. We conclude by discussing future research directions that must be further investigated for the application of OR in UWSNs.

Opportunistic Routing

In this section we give an overview of the concept of OR, its advantages, and different issues and parameters that should be considered for its use in different types of networks.

When a node sends a packet to a designated next-hop forwarder in a wireless environment with unreliable and lossy links, the next-hop node may not receive the packet. However, due to the broadcast nature of the wireless medium, a packet that is transmitted to the next-hop forwarder may be received by neighboring nodes which will be unable to assume the role of designated next-hop forwarder. Therefore, using OR, each packet travels along different paths toward the destination. This property of OR makes it suitable to be used in UWSNs, which face low link delivery probability.

The performance of OR protocols is affected by three main factors:

- The algorithm that selects the candidates
- The metric used to select and prioritize the candidates
- The method used by nodes in the CS to communicate with each other in order to do the coordination

In the following sections we explain each of these three factors in more detail.

Metrics in OR

The view taken by each node from the network topology determines the type of OR metric. In this sense, OR metrics are classified into two categories: *end-to-end* and *local* metrics. In end-to-end metrics, a node needs the topology information of the entire network. However, local metrics depend only on local information derived from nodes' neighbors. Obviously, end-to-end metrics outperform local ones in terms of the cost of delivering a packet to the destination. This is due to the fact that with the use of end-to-end metrics, all nodes are aware of the entire range of network topology information. This leads to the selection of a better set of candidates compared to the local metrics, which only provide neighbors' information. However, obtaining all network information in the wireless environment is difficult and costly when links are unreliable.

We can also classify the OR metrics into three main categories:

- *Hop count*
- *Link delivery probability*
- *Geographic position*

The hop count, which is borrowed from traditional routing protocols, is the simplest metric in OR to be considered by the initial proposals (see [6] for more details). The link delivery probability between a node and its candidates, or all link probabilities on the path to the destination, is considered in the link delivery probability metric. Finally, some researchers employ the geographic positions of nodes to find and order the CS.

Candidate Selection

Candidate selection algorithms are involved in the selection of neighboring nodes that can help forward the packet toward the destination. Assuming that there is perfect coordination between candidates, selecting more candidates decreases the expected number of packet transmissions from the source to the destination. On the other hand, selecting many candidates results in high coordination overhead.

To select a CS for a node, a candidate selection algorithm needs some information about the network. In this scene, candidate selection algorithms are divided into two main categories:

Assuming that there is perfect coordination between candidates, selecting more candidates decreases the expected number of packet transmissions from the source to the destination. On the other hand, selecting many candidates results in high coordination overhead.

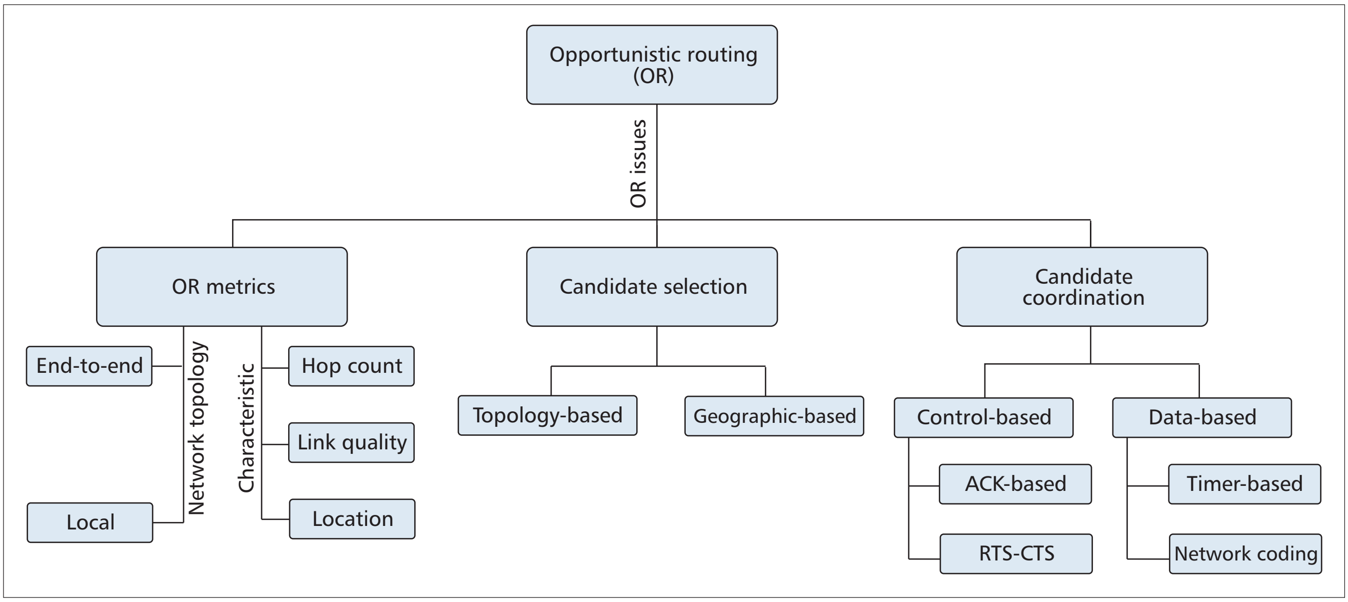


Figure 1. OR issues and classification.

geographic-based and topology-based selection. Usually, the taxonomy of candidate selection algorithms is relatively close to the metric classification they use. End-to-end metrics are usually used by topology-based candidate selection algorithms, while geographic-based algorithms use local metrics. As mentioned earlier, information from the entire network topology can lead to an optimum solution and selection of the best CS. Conversely, topology-based algorithms have a longer computational time and are more complex than geographic-based algorithms.

OR protocols are divided into two main categories: *sender-based* and *receiver-based*. In the sender-based approach, each node selects the candidates, and includes their IDs in the packet header and broadcasts the packet. The nodes that receive the packet contribute to the forwarding process if their ID is included in the packet header. In the receiver-based approach, there is no CS in the data packet header. When a source has a packet to send, it broadcasts the packet. Each neighbor that receives the packet decides whether or not to act as a potential candidate. In other words, in the receiver-based approach, the nodes that have received a packet decide whether or not to be a candidate, while in the sender-based approach, the CS is already determined by the previous hop.

The advantage of the sender-based method is that each node needs the topology information from their neighbors in order to select some as candidates. However, to have some topological information, the nodes need to exchange some control packets, which may increase the control overhead of the network. In the receiver-based approach, nodes do not need to have topological information, and each receiver can decide independently whether or not to continue forwarding the packet.

Candidate Coordination

Once a node finds its CS, it places it in the data packet header and broadcasts it. The candidates that have received the packet should collaborate with each other to decide which one must continue forwarding the packet, while the others should discard the packet. The coordination process requires signaling between candidates. For this purpose, candidate coordination is divided into two main categories: control-based and data-based coordination. Generally, control-based methods must transmit a control message between the candidates to inform each of them about the reception of the packet, and decide how to continue forwarding the received packet. In the data-based approaches there is no control packet, and coordination between candidates is achieved by using the data packet.

The different methods for control-based and data-based approaches are thoroughly explained in [6]. In this article, we explain the most commonly used approach in the literature, which is known as *timer-based* coordination. The forwarding and coordination processes in the timer-based mechanism are consolidated. A time slot is assigned to each candidate according to its priority in the CS. Each candidate must listen to the medium before its time slot arrives. When a time slot arrives for a candidate, it will forward the packet only if it does not hear the transmission of this packet from other candidates. As we can see, the timer-based mechanism does not rely on any control packet; the overhead in this candidate coordination is the wait time for each candidate before it decides to forward the packet.

One of the main drawbacks of timer-based coordination is duplicate transmissions. There is no guarantee that all candidates can overhear the transmission of the higher-priority candidate. Therefore, some candidates that do not hear the transmissions from higher-priority candidates may forward the packet, resulting in degradation of network performance. A longer wait time for low-priority candidates is another issue in timer-based coordination. The link quality between a node and its high-priority candidates is low, and the packet may not reach this candidate; therefore, lower-priority candidates must wait for a period of time before they forward the packet. This situation results in higher delays in coordinating and forwarding of the packet, which also causes high end-to-end delay.

For greater simplicity in understanding different issues in OR, we summarize each classification of different issues of OR in Fig. 1.

Opportunistic Routing in Underwater Sensor Networks

In the previous section we discuss issues related to OR. Combining weak links and creating a stronger virtual link is one of the best advantages of using OR. Some routing protocols in UWSNs use the advantages of OR in order to increase link reliability and path stability between sensor nodes and destinations (sink nodes). In this section, we provide an overview of the existing OR protocols in UWSNs by classifying them into two main categories:

- Geographic-based
- Pressure-based

The geographic-based category uses the geographic position of sensor nodes as information for selecting the candidates and forwarding decisions in OR. The latter category uses the depth (pressure) information of nodes for selecting candidates and forwarding packets. In both types of approaches, a well-known problem known as *communication void region* may occur. This problem results when no neighbor exists to forward the packet that is in closer proximity to the destination than the current forwarder. Note that the node which is located in a communication void region is called a *void node*. The routing algorithms should have some approaches to achieve recovery when the packet reaches a void node.

As mentioned earlier, candidate selection algorithms are divided into sender-based or receiver-based. We see in the following sections that most of the proposed OR protocols for UWSNs are classified in the receiver-based candidate selection approach. Furthermore, since in the control-based approaches for candidate coordination some extra control packets are needed, and considering the fact that UWSNs have low bandwidth and unreliable links, using the timer-based approach, which does not need extra control packet transmissions, results in bandwidth savings in UWSNs.

Geographic Opportunistic Routing Protocols in UWSN

The geographic positions of nodes are used in this category of OR protocols in UWSNs to select candidates and make the decision to forward the packets. By knowing the positions or the coordinates of their neighbors, the nodes are able to forward the packets.

Vector-Based Forwarding

Vector-based forwarding (VBF) [8] is an OR geographic-based protocol that does not require the state information of the sensor nodes. To make the protocol reliable against packet loss, VBF forwards the packets along redundant paths.

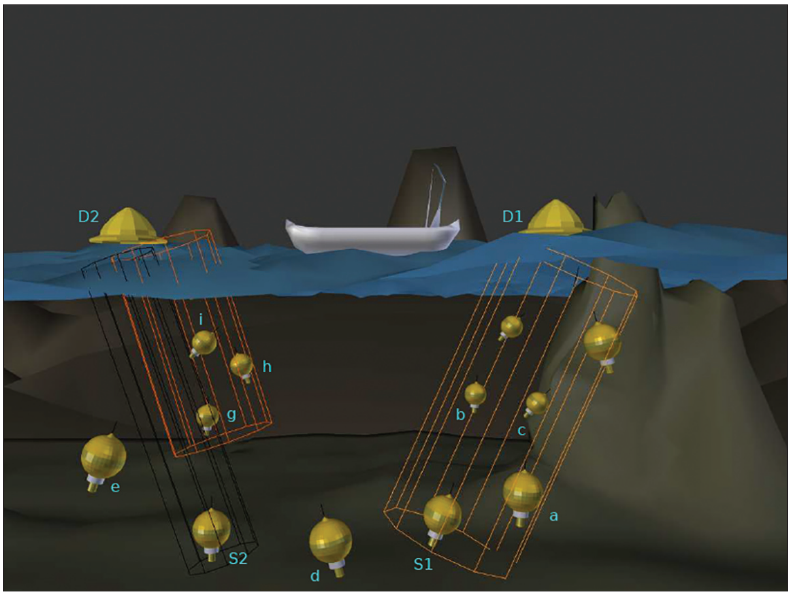


Figure 2. An illustration of VBF and HH-VBF.

The positions of the source, destination, and forwarder are included in the header of the data packet, which is transmitted using VBF. Using the positions of the source and destination, a virtual pipe (vector) between the two nodes is created through which packets can be forwarded. Consider the example shown in Fig. 2 where S_1 and D_1 are source and sink nodes, respectively. A vector of $\overrightarrow{S_1D_1}$ is established between source and destination. According to VBF, the nodes located in the virtual pipe are potential candidates for forwarding the packets. Therefore, compared to the flooding algorithm in which all nodes forward the packet, in VBF only a few nodes in the vector between source and destination forward the packet. Therefore, as we can see, VBF is a receiver-based and stateless routing protocol that needs only the destination position.

In VBF, when a node receives the packet it checks to see if it is close enough to the line between source and destination. If a node finds itself close enough to the routing vector, it includes its position in the packet header as the forwarder and transmits the packet. Note that the nodes which have received the packet but are not in the virtual pipe discard the packet. To define the pipe, a predefined distance threshold w for the closeness of nodes to the line between source and destination is considered. As shown in Fig. 2, if we assume that all neighbors of source nodes (a, b, c, and d) have received the packet, they are then considered as potential forwarding nodes. Node d, which is not located in the virtual pipe, discards the received packet. Therefore, many nodes may forward the packet, resulting in duplicate transmissions and a waste of energy in each node. In this way, all nodes in the pipe between source and destination are considered to be forwarding nodes. A self-adaptation algorithm to reduce the number of forwarding nodes and conserve energy is proposed in VBF.

When a node receives a packet, it first determines if it is in the pipe, and whether or not it can be considered as a potential forwarder. If the

Duplicate transmissions remain a problem because of the coordination method used in HH-VBF. That is, more than one candidates that are potential forwarders will forward the packet, resulting in duplicate transmissions. These duplicates transmission can cause packet collisions and waste in the network bandwidth, which is a critical parameter in UWSNs.

node is a potential candidate, it waits for a period of time determined by its *desirability factor*. The desirability factor shows the proximity of the node to the previous forwarder, and the vector between source and destination. The more desirable the node, the less time it must wait. During the wait time, the node listens to the medium to see how many nodes are forwarding the same packet as the current node. When a candidate's wait time expires, it forwards its packet if the minimum desirability factor of the other forwarders is less than a predefined value. The timer approach used in VBF is similar to timer-based coordination, with the following differences. The source node in VBF does not specify the CS, but instead specifies the virtual pipe between source and destination. In contrast to traditional OR, in which only one node should forward the packet, in VBF more than one candidate may forward the packet toward the destination.

VBF is a stateless routing protocol, in which only nodes in the pipe between the source and the destination will forward the packet. Therefore, the advantage of VBF is that it is scalable to the number of nodes in UWSNs. In addition, since more than one candidate forwards the packet over different paths, the robustness of the protocol against packet loss may be increased. On the other hand, in a network with low density or node movement, it is possible that the number of nodes located in the virtual pipe from source to destination is few or none, which will degrade the performance of the network. For example, Fig. 2 shows source S_2 sending packet to sink D_2 when there is only one node, *g*, in the virtual pipe between S_2 and D_2. There may be other nodes (*h* and *i*) outside of the pipe that can deliver the packet to the destination. VBF does not have any restriction on the number of candidates. In OR, as shown in [9], using many candidates does not increase the performance of the network compared to the duplicate transmissions that may occur due to imperfect coordination between candidates. Furthermore, the nodes located in the virtual pipe are used for many packet transmissions; they may lose their energy, while nodes around the pipe may be better candidates to which the packet may be forwarded. VBF does not have a recovery mechanism in the event that a packet reaches a void node. Therefore, on the rare occasion that a packet reaches a void node in sparse scenarios, it is discarded after a few trials.

Hop-by-Hop Vector-Based Forwarding

As mentioned earlier, considering one fixed virtual pipe between source and sink in VBF results in low packet delivery in sparse networks. Hop-by-hop VBF (HH-VBF) improves on VBF by creating a virtual pipe in each hop as packets progress toward the sink node. In contrast to VBF, which creates a virtual pipe between source and sink, HH-VBF considers a virtual pipe per hop between the current forwarder and the sink node. In another word, for each candidate there is a virtual path toward the destination, which opportunistically forwards the packet. The CS in HH-VBF is changed as packets progress toward the sink node, while the CS in the VBF is fixed and predefined by the source node. Therefore, in a sparse network there is a higher chance of delivering the packet to the destination through nodes in each virtual pipe, which is defined in each candidate.

Figure 2 shows an example of how HH-VBF works. When source S_2 sends a packet to sink D_2, the virtual pipe between S_2 and D_2 is created; as we can see, only node *g* is inside this virtual pipe. When *g* receives the packet from source S_2 and knows that it is a potential candidate to continue forwarding the packet, it creates its own virtual pipe with its position, and forwards the packet. In the new virtual pipe there are some nodes (*h* and *i*) considered to be potential candidates. As we can see, HH-VBF can function better than VBF in a sparse network by creating a virtual pipe in each hop.

Duplicate transmissions remain a problem because of the coordination method used in HH-VBF. That is, more than one candidate that are potential forwarders forward the packet, resulting in duplicate transmissions. These duplicate transmissions can cause packet collisions and waste network bandwidth, which is a critical parameter in UWSNs. VBF and HH-VBF share the problem of communication void regions. That is, if a packet is delivered to a node that can no longer forward it, the packet is discarded as there are no other nodes in closer proximity to the destination.

GEographic Opportunistic Routing Depth Adjustment-based Topology Control for Communication Recovery (GEDAR)

A sender-based OR for UWSNs which includes both geographic routing and OR was introduced in [10]. GEographic opportunistic routing Depth Adjustment-based topology control for communication Recovery (GEDAR) has two modes: opportunistic and recovery phases. In GEDAR, the CS is determined in each forwarder. In the opportunistic mode, each node uses greedy opportunistic forwarding to route the packets. In contrast to VBF and HH-VBF, which do not have recovery modes when the packet reaches a void node, GEDAR changes to recovery mode when the packet reaches a void node, moving it to a new depth in order to adjust the topology. Furthermore, GEDAR is designed such that it is able to use one or multiple sink nodes on the water's surface to gather the transmitted information.

In GEDAR, each node assumes that it has the geographic positions of its neighbors. The Expected Packet Progress (EDP) [11] metric is used to select the CS. Note that EDP is the expected packet progress using a set of nodes as the CS (see [6] for more details). The basic idea of the greedy heuristic method to select the CS in GEDAR is to find a set of nodes that maximizes the EDP metric. In the proposed algorithm, the forwarding node *S* calculates the Normalized Advance (NADV) for each of its neighbors *i*, defined as the normalized packet progress, d_{si}, by its associated cost p_{si}. That is $NADV_{sv} = d_{si} \times p_{si}$, where D_{si} and p_{si} represent the advancement of packets using neighbor *i* as the destination and the link delivery probability between nodes *S* and *i*, respectively. GEDAR adds the neighbor *j* with the highest *NADV* to the first set of f_1. It then looks for the neighbors of j which are in $R/2$ where R is the communication range of nodes to be included in f_1. The above process is repeated until there is no remaining neighbor which is not in any of the sets $f_1, f_2, \ldots, f_n$. The final CS is selected among the sets $f_1, f_2, \ldots, f_n$ which maximizes the EDP metric. For the coordination phases GEDAR uses the timer-based approach. Note that to reduce the number of duplicate transmissions and increase the probability that lower priority candidates hear higher priority transmissions, GEDAR selects the candidates such that they are in $R/2$ of communication range of each other. Note that GEDAR supports multiple-sink nodes. When a node wants to find the CS, it considers the closest sink as a destination.

Figure 3 shows an example of how GEDAR creates the CS. Assuming that the sensor s wants to find its candidates, it calculates the NADV for each neighbor and picks the one with the highest value. If node a has the best NADV, it is added to set f_1 and the algorithm adds the neighbors of a which are in half of its communication range. Therefore, nodes c, d and e which are in half of the communication range of sensor a will be added to f_1. The GEDAR candidate selection algorithm calculates the EDP of each set; the one with the highest value is chosen as the CS.

When a node has data to send to a sink node, it includes its CS in the data packet header and broadcasts it. Upon receiving the transmitted packet by a node, it verifies whether its ID is included in the CS. If it is a candidate to forward, it uses the timer-based coordination to decide whether or not it is the highest priority candidate to receive the packet. The above process will continue until the packet reaches one of the sink nodes on the surface of the water. If the packet reaches a void node, GEDAR stops and the OR switches to recovery phase. When a node changes from greedy phase to the recovery phase, it stops sending packets and calculates the new depth to move to. Note that, since this article focuses on OR in UWSNs, we are not going to explain the recovery process in GEDAR.

As discussed earlier, in order to select candidates, GEDAR does not consider energy level of the sensors. Therefore, after a period of time, GEDAR may not be able to select any nodes as a candidate if it has the energy to continue forwarding the packet. Furthermore, to select the candidates, each node in GEDAR, needs to know its neighbors' positions. This results in a large number of beaconing messages, and consequently a waste of network bandwidth. Although GEDAR tries to select candidates that are within the communication range of each others, due to fading effects, they may not hear one anothers' transmissions in the coordination phase. Thus, more than one candidate would forward the packet, resulting in duplicate transmissions.

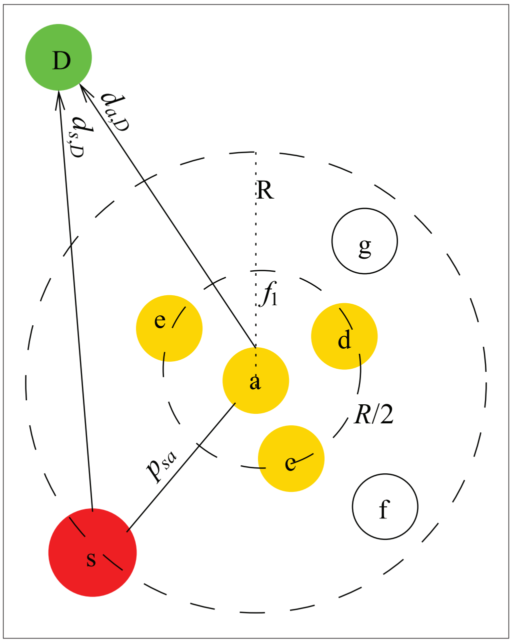


Figure 3. An illustration of GEDAR and HydroCast.

The geographic-based approaches require the locations of mobile and sink nodes; because of low data rates and limited bandwidth in UWSNs, this is a difficult challenge. Using depth sensor equipment in UWSNs is another solution that is used to forward data packets to the water's surface.

Pressure-Based Opportunistic Routing Protocols

The geographic-based approaches [6] require the locations of mobile and sink nodes; because of low data rates and limited bandwidth in UWSNs, this is a difficult challenge. Using depth sensor equipment in UWSNs is another solution that is used to forward data packets to the water's surface. In this section we review some protocols which use the water pressure at the depth of the node as an OR metric to forward the packets toward the sink node(s).

Depth-Based Routing (DBR)

An any-cast opportunistic routing protocol for UWSNs is proposed in [12]. Depth-Based Routing (DBR) uses the depth information of nodes as a metric to decide whether to forward the packets. It assumes that each node is equipped with a depth sensor to measure the depth of node. With the OR concept, the nodes closer to the water surface are potential candidates for receiving and forwarding the packet. Like VBF and HH-VBF, DBR is a receiver-based OR protocol, but uses the depth of nodes as an OR metric.

Multi-sink architecture is used in DBR, which helps increase packet delivery in the network. That is, the packet, which is generated from a sensor node, can be delivered to any of the sinks which are located on the surface of the water. The idea of using multi-sinks in DBR is similar to that used in *Plasma* [13], in which each packet can be delivered to any of the gateways.

When a sensor node has data to send in DBR, it measures its depth, puts it in the data packet header, and broadcasts it. Each node that has received the packet compares its depth with the one included in the data packet header. If a node's depth is lower (closer to the surface of the water) than the depth of the previous forwarder, the receiver is considered a potential candidate to forward the packet. DBR uses timer-based coordination to prevent duplicate transmissions. Clearly, the candidate that has less depth waits for a shorter period of time. When the wait time in a candidate expires, it will

DBR does not need a control packet to obtain depth information, therefore it is a scalable protocol. It does not have any mechanism for recovery when a packet reaches a void node. Therefore, like VBF and HH-VBF, it suffers from the local maximum problem in a sparse scenario.

forward the packet if it does not hear the transmission of the same packet. Having a large number of candidates increases link reliability and the chance for packet progress. In order to reduce the number of candidates and have fewer duplicate transmissions, DBR adds a *Depth Threshold* in the data packet header. A node is allowed to be considered a potential candidate if its depth difference to the previous hop's depth is greater than the threshold.

DBR does not need a control packet to obtain depth information, therefore it is a scalable protocol. It does not have any mechanism for recovery when a packet reaches a void node. Therefore, like VBF and HH-VBF, it suffers from the local maximum problem in a sparse scenario. Note that a larger depth threshold may result in fewer nodes in the CS. Therefore, this situation causes a lower packet delivery ratio. On the other hand, if the depth threshold is small, many nodes may be eligible for forwarding the packet, and duplicate transmissions in the network may occur.

In addition to the previously mentioned challenges for DBR, such as VBF, HH-VBF and GEDAR, DBR does not consider the energy level of sensor nodes as a metric for deciding which node should act as a member of CS. That is, a node with less depth but little remaining energy may choose to forward the packets, while another node with similar depth but greater energy will not have a chance to be in the CS.

Hydraulic Pressure Based Anycast Routing (HydroCast)

Hydraulic Pressure Based Anycast Routing (HydroCast) [15] is proposed to DBR's problem of local maximum issue. HydroCast is similar to DBR in the sense that it uses nodes' depth information to select candidates from the neighboring nodes. There are two main differences between HydroCast and DBR. The algorithm for selecting candidates in DBR is simple, while HydroCast proposes a candidate selection algorithm which considers the Expected Packet Progress (EDP) [11] as an OR metric. In addition to the new candidate selection algorithm in HydroCast, the authors proposed an approach for local maximum recovery when a packet is delivered to a void node.

Similarly to GEDAR, the candidate selection and coordination of HydroCast is based on EDP and the heuristic algorithm, creating different sets so that they can hear one another, while minimizing the EDP. For the candidate coordination, the timer-based algorithm is used in the same manner as the other existing approaches. However, HydroCast and GEDAR differ in their performance of the recovery phase. HydroCast has the same problem as the OR metric, which fails to consider the energy of sensors - a challenging issue in UWSNs. Furthermore, gathering the neighbors' information, including their depths and particularly their link delivery probabilities, is a concern due to acoustic communication in UWSNs.

Void-Aware Pressure Routing (VAPR)

Void-Aware Pressure Routing (VAPR) [16] is proposed in order to solve the void node problem. However, it uses the idea of OR for its packet forwarding. It has created two phases: beaconing and opportunistic data forwarding. In the beaconing phase, each sink starts sending the beacon message including depth, hop count and direction of the current node. When a node receives a beacon message, it uses the beacon to store information, updates packet information by depth and hop count, and re-broadcasts it. Note that if a node receives a message from another one with smaller depth, then the direction of the node is revised to upwards; otherwise, it is revised downwards. In the opportunistic forwarding section, using neighbors' directions, VAPR removes nodes from its search area; using them, packets will reach void nodes, and use the HydroCast candidate selection algorithm to find the CS. In a different manner that HydroCast, VAPR relies on no recovery phase in the event that the packets reach a void node. In fact, VAPR selects nodes according to the information in each node, such that the packet will not trapped in a void node.

Discussion

In this article, we have reviewed important key issues and research challenges in using OR in UWSNs. We have summarized all the explained protocols including their advantages and disadvantages in Table 1. By using this table, the comparison between the various OR protocols in UWSNs is simplified. The significance of each column is explained in the following text.

Protocol: This field identifies the name assigned by the authors to the proposed approach. The corresponding reference is also provided here.

OR Metric: This specifies the OR metric used by the algorithm to select the candidates. As we can see from Table 1, the proposed algorithms use the geographic positions of nodes, or their depths, as usual OR metric.

Cand Sel: We have used the *local* term if the candidate selection algorithm uses only local information of nodes for candidate selection. We have used *NA* if there is no special algorithm to select the candidates.

Cand Coor: This field shows the algorithm to coordinate between candidates. As we can see, most proposals use the timer-based coordination, which is a reasonable approach for the UWSNs with limit bandwidth, as using fewer control packets is more desirable. VBF does not use a coordination method, and each of the candidates which has received the packet can forward the packet. Therefore, for VBF we have used *NA*.

Void Recovery: As we mentioned earlier, geographic routing faces void region communication problem. This filed shows whether or not the protocol has a void recovery approach. Note that VAPR does not have a void recovery mechanism, since the selection of forwarding nodes is such that this problem will not occur.

Receiver/Sender-based: OR protocols are divided into two main categories, based on whether the sender selects the candidates, or the receiving nodes decide to continue forwarding. The benefit of a sender-based protocol is that the sender selects the candidates, therefore it

Protocol	OR metric	Cand sel.	Cand coor.	Void/ recovery	Receiver/ sender-based	Multi-sink	Pros	Cons
VBF [8]	Geographic	NA	Timer-based	×	Receiver-based	×	Scalable, robustness	Duplicate Tx, void region
HH-VBF [14]	Geographic	NA	Timer-based	×	Receiver-based	×	Scalable, robustness	Duplicate Tx, void region
GEDAR [10]	Geographic	Local-based	Timer-based	✓	Sender-based	✓	Maximize EDP, recovery mechanism	Complex algorithm, control pkt
DBR [12]	Depth	NA	Timer-based	×	Receiver-based	✓	Simple algorithm	Duplicate Tx, void region
HydroCast [15]	Depth	Local-based	Timer-based	✓	Sender-based	✓	Recovery mechanism	Complex algorithm
VAPR [16]	Depth	Local-based	Timer-based	×	Sender-based	✓	No void region	Complex algorithm

Table 1. Summary of opportunistic routing protocol in UWSNs.

has control on the number of candidate to reduce packet duplication. On the other hand, the sender-based approaches must know the topology information of neighboring nodes that need beaconing messages. At the same time the receiver-based algorithms do not need control packets and each node, based on some criteria, can act as a candidate. The duplicate transmissions happen more in receiver-based approaches, since the sender does not have full control of the CS.

Multi-Sink: this field in Table 1 shows whether the protocol supports only one sink on the surface or if it can handle multiple sinks. Having multiple sinks on the water surface can result in more reliable packet delivery; however, the underwater sensor nodes are expensive. In the multi-sink supports the packets can be delivered to any sink on the surface, while obviously in the single sink the sensors must deliver their packets to only one existing sink. As we can see in Table 1, most protocols support the multi-sinks solution, since delivering the information gathered from the bottom of the sea is a costly task which is very important for UWSNs that face long delay and limited bandwidth.

Pros, Cons: There two fields in Table 1 show the advantages and disadvantages of each protocol under study. As we can see most of the protocols have the disadvantage of duplicate transmissions; and those which do not have this problem are a complex protocols (HydroCast and VAPR). Note that HydroCast and GEDAR are two approaches which have the recovery mechanism in case that the packet is reached to a void node. VAPR does not have the recovery mechanism because its algorithms is such that the packet never reaches a void node. We should consider this fact that none of the proposed approaches in the literature consider the energy efficiency of nodes in their proposals. Since it is a common disadvantage among all of the protocols, we do not put it in Table 1.

Future Research Directions for OR in UWSNs

There is no doubt of the benefit of using OR in UWSNs that have limited bandwidth and lossy links. However, there are many challenges that deserve greater attention in order to achieve a better performance in UWSNs with the use of OR protocols. In this section, we discuss some directions of future research in this area to be investigated.

The first direction of research could be investigating the different parameters of the use of OR in UWSNs. It is clear that a greater number of candidates leads to higher chances of delivery packets progressing toward the destination. On the other hand, more candidates need precise candidate coordination, in order to prevent duplicate transmissions; these will use extra bandwidth, an important concern in UWSNs. The effect of the number of candidates used in CS on the performance of protocols is another potential direction for future research.

In OR protocols used in the mesh network or terrestrial environment, the highest priority is assigned to the candidate that is closest to the destination. In the coordination phase, the lower priority candidates must wait for a certain period of time to allow the highest priority candidate to forward the packet. In the UWSNs the priority assignment could be done another way, and not necessarily by the proximity of nodes to the surface or destination.

None of the proposed protocols have considered the acoustic communication in UWSNs which has an significant impact on the link delivery probability. This will affect the candidate selection algorithm which needs to be considered in the future OR protocols.

The sensor nodes in the UWSNs have mobility. Therefore, a new metric and candidate selection algorithm which consider the mobility of nodes are necessary in order to deal with the mobile scenario. In addition, selecting the candidates according to some Quality of Service

(QoS) parameter could be another future research direction.

All the proposals in the literature used the simulation to study the performance of the proposed protocols. It is worthwhile to look for a mathematical model to investigate the performance of OR in UWSNs.

Finally, none of the existing works consider the energy level of each sensor node in candidate selection. Some sensors may always be selected as candidates, and after a period of time they die as a result of consuming their energy reserves. Therefore, in a network in which energy is a concern, considering the energy consumption as a metric to select candidates is an important issue that should be considered in the future.

Conclusion

OR has been proposed as an approach for exploiting the broadcast nature of wireless multihop networks by selecting multiple nodes as candidates for packet forwarding. In OR, based on network conditions multiple candidates are used to forward the packet toward the destination. The main idea of OR is to combine weak links and create a strong virtual link in order to have one reliable link. UWSNs use acoustic channels, which have lower bandwidth and longer propagation delays compared to radio channels. These kind of networks face an unreliable environment compared to WSNs. These properties of UWSNs make them suitable candidates for using the OR concept, resulting in more reliable links and fewer transmissions.

UWSNs use acoustic channels, which have lower bandwidth and longer propagation delays compared to radio channels. These kind of networks face an unreliable environment compared to WSNs. These properties of UWSNs make them suitable candidates for using the OR concept, resulting in more reliable links and fewer transmissions.

In this article, we have presented the OR protocols proposed for UWSNs. We have classified the OR protocols in different categories, from different points of view: Geographic-based vs depth-based, sender-based vs receiver-based. From each category, we have reviewed the most important and representative proposals in the literature. Finally, we have addressed some future directions of research that need to be investigated further in order to develop an efficient OR protocol for UWSNs.

References

[1] I. F. Akyildiz, D. Pompili, and T. Melodia, "Underwater Acoustic Sensor Networks: Research Challenges," *Ad Hoc Networks*, vol. 3, no. 3, 2005, pp. 257–79.

[2] M. Ayaz *et al.*, "A Survey on Routing Techniques in Underwater Wireless Sensor Networks," *J. Network and Computer Applications*, vol. 34, no. 6, 2011, pp. 1908–27.

[3] D. Zeng *et al.*, "A Survey on Sensor Deployment in Underwater Sensor Networks," *Advances in Wireless Sensor Networks*, ser. Communications in Computer and Information Science, L. Sun, H. Ma, and F. Hong, Eds. Springer Berlin Heidelberg, 2014, vol. 418, pp. 133–43.

[4] I. F. Akyildiz, D. Pompili, and T. Melodia, "Challenges for Efficient Communication in Underwater Acoustic Sensor Networks," *ACM Sigbed Rev.*, vol. 1, no. 2, 2004, pp. 3–8.

[5] S. Biswas and R. Morris, "ExOR: Opportunistic Multi-Hop Routing for Wireless Networks," *ACM SIGCOMM Comp. Commun. Rev.*, vol. 35, no. 4, 2005, pp. 133–44.

[6] A. Boukerche and A. Darehshoorzadeh, "Opportunistic Routing in Wireless Networks: Models, Algorithms, and Classifications," *ACM Comp. Surv.*, vol. 47, no. 2, Nov. 2014, pp. 22:1–22:36.

[7] A. Boukerche *et al.*, "Routing Protocols in Ad Hoc Networks: A Survey," *Computer Networks*, vol. 55, no. 13, Sept. 2011, pp. 3032–80.

[8] P. Xie, J.-H. Cui, and L. Lao, "VBF: Vector-based Forwarding Protocol for Underwater Sensor Networks," *Networking 2006*, ser. Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2006, vol. 3976, pp. 1216–21.

[9] A. Darehshoorzadeh *et al.*, "On the Number of Candidates in Opportunistic Routing for Multi-Hop Wireless Networks," *Proc. 11th ACM Int'l. Symp. Mobility Management and Wireless Access (MobiWac)*, 2013, pp. 9–16.

[10] R. W. L. Coutinho *et al.*, "GEDAR: Geographic and Opportunistic Routing Protocol with Depth Adjustment for Mobile Underwater Sensor Networks," *IEEE ICC 2014 — Ad-Hoc and Sensor Networking Symp. (ICC'14 AHSN)*, June 2014, pp. 251–56.

[11] K. Zeng, W. Lou, J. Yang, and D. Brown, "On Geographic Collaborative Forwarding in Wireless Ad Hoc And Sensor Networks," *Int'l. Conf. Wireless Algorithms, Systems and Applications*, 2007, WASA 2007, Aug. 2007, pp. 11–18.

[12] H. Yan, Z. Shi, and J.-H. Cui, "Dbr: Depth-based Routing for Underwater Sensor Networks," *Proc. IFIP Networking*, 5 2008, pp. 1–13.

[13] R. Laufer *et al.*, "Plasma: A New Routing Paradigm for Wireless Multihop Networks," *2012 Proc. IEEE INFOCOM*, Mar. 2012, pp. 2706–10.

[14] N. Nicolaou *et al.*, "Improving the Robustness of Location-based Routing for Underwater Sensor Networks," *OCEANS 2007 — Europe*, June 2007, pp. 1–6.

[15] U. Lee *et al.*, "Pressure Routing for Underwater Sensor Networks," *2010 Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.

[16] Y. Noh *et al.*, "Vapr: Void-Aware Pressure Routing for Underwater Sensor Networks," *IEEE Trans. Mobile Computing*, vol. 12, no. 5, 2013, pp. 895–908.

Biographies

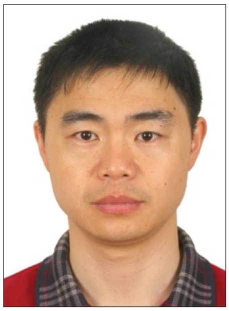
Amir Darehshoorzadeh (adarehsh@uottawa.ca) Research Associate at the the School of Electrical Engineering and Computer Science, University of Ottawa, Canada. He received his Ph.D. degree in 2012 with Cum Laude from the Technical University of Catalonia (UPC), Barcelona, Spain. He received his M.Sc. Degree from Iran University of Science and Technology (IUST), Tehran, Iran in 2006. His main research areas are Opportunistic Networks, Modeling and Network optimization, Wireless Networks including VANETs, MANETs, WSNs, Multicast protocols and QoS provision. He has extensively published research papers in international conferences and journals and presented several seminars in mentioned areas.

Azzedine Boukerche [FIEEE, FEiC, FCAE, FAAAS] is a full professor and holds a Canada Research Chair position at the University of Ottawa (Ottawa). He is the founding director of the PARADISE Research Laboratory, School of Information Technology and Engineering (SITE), Ottawa. Prior to this, he held a faculty position at the University of North Texas, and he was a senior scientist at the Simulation Sciences Division, Metron Corp., San Diego. He was also employed as a faculty member in the School of Computer Science, McGill University, and taught at the Polytechnic of Montreal. He spent a year at the JPL/NASA-California Institute of Technology, where he contributed to a project centered about the specification and verification of the software used to control interplanetary spacecraft operated by JPL/NASA Laboratory. His current research interests include wireless ad hoc, vehicular, and sensor networks, mobile and pervasive computing, wireless multimedia, QoS service provisioning, performance evaluation and modeling of largescale distributed systems, distributed computing, large-scale distributed interactive simulation, and parallel discrete-event simulation. He has published several research papers in these areas. He served as a guest editor for the Journal of Parallel and Distributed Computing (special issue for routing for mobile ad hoc, special issue for wireless communication and mobile computing, and special issue for mobile ad hoc networking and computing), ACM/Kluwer Wireless Networks, ACM/Kluwer Mobile Networks Applications, and Journal of Wireless Communication and Mobile Computing. He has been serving as an Associate Editor of ACM Computing Surveys, IEEE Transactions on Parallel and Distributed systems, IEEE Transactions on Vehicular Technology, Elsevier Ad Hoc Networks, Wiley International Journal of Wireless Communication and Mobile Computing, Wiley's Security and Communication Network Journal, Elsevier Pervasive and Mobile Computing

Journal, IEEE Wireless Communication Magazine, Elsevier's Journal of Parallel and Distributed Computing, and SCS Transactions on Simulation. He was the recipient of the Best Research Paper Award at IEEE/ACM PADS 1997, ACM MobiWac 2006, ICC 2008, ICC 2009 and IWCMC 2009, and the recipient of the Third National Award for Telecommunication Software in 1999 for his work on a distributed security systems on mobile phone operations. He has been nominated for the Best Paper Award at the IEEE/ACM PADS 1999 and ACM MSWiM 2001. He is a recipient of an Ontario Early Research Excellence Award (previously known as Premier of Ontario Research Excellence Award), Ontario Distinguished Researcher Award, Glinski Research Excellence Award, IEEE CS Golden Core Award, IEEE Canada Gotlieb Medal Award, IEEE ComSoc Expectional Leadership Award, IEEE TCPP Exceptional Leadership Award. He is a co-founder of the QShine International Conference on Quality of Service for Wireless/Wired Heterogeneous Networks (QShine 2004). He served as the general chair for the Eighth ACM/IEEE Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems, and the Ninth ACM/IEEE Symposium on Distributed Simulation and Real-Time Application (DS–RT), the program chair for the ACM Workshop on QoS and Security for Wireless and Mobile Networks, ACM/IFIPS Europar 2002 Conference, IEEE/SCS Annual Simulation Symposium (ANNS 2002), ACM WWW 2002, IEEE MWCN 2002, IEEE/ACM MASCOTS 2002, IEEE Wireless Local Networks WLN 03–04; IEEE WMAN 04–05, and ACM MSWiM 98–99, and a TPC member of numerous IEEE and ACM sponsored conferences. He served as the vice general chair for the Third IEEE Distributed Computing for Sensor Networks (DCOSS) Conference in 2007, as the program cochair for GLOBECOM 2007–2008 Symposium on Wireless Ad Hoc and Sensor Networks, and for the 14th IEEE ISCC 2009 Symposium on Computer and Communication Symposium, and as the finance chair for ACM Multimedia 2008. He also serves as a Steering Committee chair for the ACM Modeling, Analysis and Simulation for Wireless and Mobile Systems Conference, the ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks, and IEEE/ACM DS–RT.

GUEST EDITORIAL

SOFTWARE DEFINED WIRELESS NETWORKS (SDWN): PART 1


Honglin Hu

Hsiao-Hwa Chen

Peter Mueller

Rose Qingyang Hu

Yun Rui

The growing popularity of smart phones, tablet computers and mobile cloud services places an increasing demand for dynamic services from wireless networks. This demand creates new requirements for the network architecture, such as flexibility in management and configuration, adaptability and vendor-independence. To meet these requirements, software defined wireless network (SDWN) has been proposed as a cost-effective solution. SDWN decouples the data plane from the control plane, enabling direct programmability of network controls and an abstraction of the underlying infrastructure for wireless applications. With SDWN, we can create a service delivery platform that is adaptable to the users' varying demands. However, issues such as supporting a large number of subscribers, frequent mobility, fine-grained measurement and control, and real-time adaptation need to be addressed by future SDWN architectures. In this Feature Topic, we provide an overview of the latest major developments and progresses in SDWN.

The first article, "Network Virtualization and Resource Description in Software-defined Wireless Networks" by Zhou *et al.*, focuses on the state-of-the-art SDWN architecture, including control plan virtualization strategies as well as semantic ontology for network resource description. The key technology to implement network resource description, semantic web technology, has been introduced in detail along with its three key elements including metadata, ontology and Resource Description Framework (RDF).

Inspired by an information centric view, in the second article, "When ICN Meets C-RAN for HetNets: An SDN Approach," Yang *et al.* realize that the rapid increase of network traffic and the change of the communication mode in the HetNet require a new wireless architecture for easier, flexible and reconfigurable infrastructure deployment and network management. As such, the authors propose an information-centric software defined networking (SDN) and Cloud Radio Access Network (C-RAN) architecture consisting of three planes, named application, control and forwarding, respectively. The vision of the proposed system is demonstrated, followed by an analysis of advantages and challenges. The authors also use a large-scale wireless heterogeneous campus network as an example to demonstrate the outstanding performance on network throughput and traffic offloading of the proposed architecture.

The software oriented design in mobile networks will be fundamentally different from SDN for Internet. Mobile networks predominantly have to care on the wireless access problem in complex radio environments, while the Internet needs to handle the packet forwarding problem. With this key differentiation in sight, the third article, "Software Defined Mobile Networks: Concept, Survey and Research Directions" by Chen *et al.*, presents the needs and requirements of SDWN, with the focus on the software-defined design for radio access networks (RAN). The article analyzes the fundamental problems in RAN which require novel SDN design approaches. They identify several areas for SDN on RAN which largely remain as open research topics.

The flexibility introduced with the SDWN architecture provides many opportunities for novel resource management concepts and methodologies. The fourth article, "Service Providers Competition and Cooperation in Cloud-based Software Defined Wireless Networks" by Ding *et al.*, is dedicated to the resource management problem in cloud-based SDWN. It explores the benefits and disadvantages of cooperation and competition between cloud service providers (CSPs). Through analyses and benchmarks the authors advocate cooperation and resource sharing among CSPs which has great significance in efficiently utilizing constraint resources at a high level of Quality-of-Service.

The fifth article, "An Intelligent SDN Framework for 5G Heterogeneous Networks" by Sun *et al.*, proposes an SDN based intelligent model to efficiently manage the heterogeneous infrastructure and resources under dynamic demands. The paper first reviews the latest SDN standard progresses and then discusses possible extensions. It proposes a number of SDN based schemes for different application scenarios to improve traffic control, subscriber management and resource allocation. The authors also present performance analysis of the proposed schemes.

The five articles included in this issue cover a wide variety of SDWN-related topics from candidate architectures to technological enablers for future SDWN. We believe that these articles will give an overall direction for those interested in this topic.

The Guest Editors would like to thank the previous Editor-in-Chief (Sean Moore) and the current Editor-in-Chief (Osman Gebizlioglu) for the guidance, feedback, and encouragement along the way. We also would like to thank the large number of people who significantly contributed to this Feature Topic, including the authors, reviewers, and the IEEE Communications Magazine Publications staff.

BIOGRAPHIES

HONGLIN HU [SM] received his Ph.D. in 2004 from the University of Science and Technology of China (USTC), China. Then, he was with Siemens AG Communications in Munich. Now he is the vice director of Shanghai Research Center for Wireless Communications (WiCO) and also serves as a Professor at the Shanghai Institute of Microsystem and Information Technology (SIMIT), Chinese Academy of Sciences (CAS). He is a Finland Distinguished Professor (FiDiPro).

HSIAO-HWA CHEN [F] is currently a Distinguished Professor in the Department of Engineering Science, National Cheng Kung University, Taiwan. He obtained his BSc and MSc degrees from Zhejiang University, China, and a Ph.D. degree from the University of Oulu, Finland, in 1982, 1985 and 1991, respectively. He served as the Editor-in-Chief for IEEE Wireless Communications from 2012 to 2015. He is a Fellow of IET and an elected Member at Large of IEEE ComSoc.

PETER MUELLER [M] joined IBM Research as a Research Staff Member in 1988. His research expertise covers broad areas of human-machine interface, systems architecture, microwave technology, device physics, nano science and modeling. His current field of research is in the areas of data center storage, security and quantum technologies. As an active member of IEEE, he authored and co-authored more than 100 papers, 2 books, granted 10 patents and served as guest editor for many publications.

ROSE QINGYANG HU is an Associate Professor of Electrical and Computer Engineering Department at Utah State University. She has more than 10 years of R&D experience with Nortel, Blackberry and Intel as a senior researcher. Her current research interests include next-generation wireless communications, wireless system design and optimization, multimedia QoS/QoE, communication and information security. She is an IEEE ComSoc Distinguished Lecturer 2015–2016 and received Best Paper Awards from IEEE Globecom 2012 and ICC 2015.

YUN RUI [SM] received the B.S. degree from Southeast University (SEU) in 2004 and the Ph.D. degree from Chinese Academy of Sciences (CAS) in 2009, all in telecommunications engineering. From Feb. 2011 to Aug. 2011, he was a visiting fellow at the Department of Electronic Engineering, City University of Hong Kong. Since Sept. 2011, he has been an Associate Professor in the Chinese Academy of Sciences.

Network Virtualization and Resource Description in Software-Defined Wireless Networks

Qianru Zhou, Cheng-Xiang Wang, Stephen McLaughlin, and Xiaotian Zhou

Qianru Zhou, Cheng-Xiang Wang, and Stephen McLaughlin are with Heriot-Watt University.

Xiaotian Zhou is with Shandong University.

The authors acknowledge the support from the EPSRC TOUCAN project (No. EP/L020009/1), EU H2020 ITN 5G Wireless project (No. 641985), EU FP7 QUICK project (No. PIRSES-GA-2013-612652), MOST 863 project in 5G wireless networking (No. 2014AA01A701), and International S&T Cooperation Program of Chin a (No. 2014DFA11640).

ABSTRACT

Future networks will be defined by software. In contrast to a wired network, the software defined wireless network (SDWN) experiences more challenges due to the fast-changing wireless channel environment. This article focuses on the state-of-the-art of SDWN architecture, including control plane virtualization strategies and semantic ontology for network resource description. In addition, a novel SDWN architecture with resource description function is proposed, along with two ontologies for the resource description of the latest wireless network. Future research directions for SDWN, control strategy design, and resource description are also addressed.

INTRODUCTION

It is widely accepted that future networks will be defined by software. With an ever-increasing demand for broadband communications, new challenges for networks keep coming up, such as intelligent and ubiquitous connectivity, efficient and flexible allocation of resources, etc. To meet these challenges, future networks must support convergence over traditionally separated network domains and offer greater granularity and flexibility in control and in data throughput. With the core idea of separating the control and data planes, software defined networking (SDN) has been considered as one promising approach to meet those challenges in the future.

SDN naturally virtualizes the network architecture and isolates the data/control traffic. The logically centralized control plane, with the global knowledge of the network state, is able to obtain, update, and even predict the global information. Thus, SDN can guide end users to select the best accessing network, or even provide them with services from multiple networks. SDN can be treated as one paradigm, rather than an ossified architecture, where one central software program, the controller, is employed to optimize and dictate the overall network behavior [1].

SDN can naturally be extended to versatile scenarios, such as optical networks, mobile wireless networks, data center networks, and cloud computing. The design of wireless network architecture is much more challenging as it must deal with various physical restrictions caused by the fast-changing nature of wireless channels [2, 3]. In the fifth generation (5G) wireless communications, by implementing SDN in eNodeB, distributing control is proven to have higher efficiency [3]. Server virtualization of wireless networks is also more challenging than that of wired networks, as the former has to satisfy the requirements of both coherency and hardware isolation [1].

Furthermore, to implement multiple control strategies at real networks, a universal agreed description of network resources is needed. However, to the best of the authors' knowledge, no such effort has been reported for software defined wireless networking (SDWN), which will be introduced in Section II. One possible reason might be that the existing test-beds for SDWN are relatively small and simple, thus there is no need to develop specialized technologies for network resource description. Nevertheless, considering the heterogeneous networks that exist in the real world, it is important to reach an agreement on the format and schema to uniquely represent the network resources for all layers. Hence, the massive resources can be possibly manipulated simultaneously with high efficiency.

Semantic Web technologies are believed to be a promising tool for unique resource description. The semantic web is created by the World Wide Web Consortium (W3C), with the intention "to create a universal medium as the exchange of data can be shared and processed by automated tools as well as people" [4]. The goal of the Semantic Web is to define a universally recognized model, which the machines can directly adopt to process and relate the information, as if they can think. With semantic technologies, one can treat the structure of complex hybrid networks as a distributed database, where an SQL-ish language is ready to search, locate, and edit a node or link of the network. Moreover, with a network description language based on a RDF, one can describe the network topolo-

0163-6804/15/$25.00 © 2015 IEEE

gy as a flexible and extendable graph to the computer (and also to us).

Semantic technologies have been implemented in optical networks [5]. However, the situation is much more complicated for wireless networks, not only because of the fast-changing channel environment, but also because of the varied service requirements from different customers. Thus, it is a difficult task to describe wireless network resources through semantic technologies. In this article, we focus on this topic and propose an SDWN architecture with the function of resource description, as well as two ontologies for the description of LTE wireless network resources.

The rest of this article is structured as follows. We give an overview of existing SDWN architectures. Current control strategies are introduced and discussed in detail. We demonstrate network description and semantic web technologies. Several current semantic description languages, as well as the software implementations and performance evaluation, are discussed. In this section, we also propose a novel SDWN architecture with resource description module and two ontologies for 5G wireless networks. Finally, conclusions and future challenges are addressed.

SDWN Virtualization Architecture

SDWN and OpenFlow

SDWN, by its nature, is about making decisions on how a connection or a flow is transmitted across the whole network. As mentioned above, the core concept of SDWN is to split the data and control planes. The control plane is responsible for the network resource assignment and decision-making. Once a decision is set, it will communicate with the data plane through a particular protocol to finish the transmission. The most widely adopted protocol is OpenFlow [6]. OpenFlow[1] is the first SDN standard, and hence the most frequently used SDN language. Its inventors deem it as the enabler of SDN. Not only does it configure the network elements, it also provides an open protocol to program the flow-table in different switches and routers. SDWN could solve the problem of network ossification efficiently and make both the control plane and data plane programmable. It also helps new technologies to be integrated and tested in networks considerably simpler, and therefore accelerates network evolution.

Architecture Designs of SDWN

Current SDWN research mainly focuses on network architectures. Existing designs often focus on different positions in the network. For example, RouteFlow [1] focuses on the IP routing services, while FlowVisor [7] and FlowN [8] concentrate on slicing the network physical infrastructure, by placing a slicing layer between the data plane and the control plane. OpenRoads [1] was proposed with the intention to replace current WiFi networks. The details of RouteFlow, FlowVisor and FlowN will now be introduced in this subsection.

RouteFlow consists of two parts, an OpenFlow controller and a RouteFlow server running as an application on top of it. There is a virtual network topology with virtual machines, where the routing engine is installed on each of them [1]. The virtual network topology mirrors the physical network topology. The virtualized topology may not necessarily reflect the exact topology of the real physical infrastructure, due to the nature character of SDN. That is to say, according the state of the real-time communication demand, the controller will mirror the actual physical network in different ways. There are three main mapping schemas: 1 : 1, 1 : n, and m : 1 or m : n, where m and n are both positive integers, denoting the numbers of nodes in physical network topology and logical network topology, respectively. These three mapping schemas represent logical split, multiplexing, and aggregation [9].

Another SDWN controller that is built on OpenFlow, called FlowVisor, has already been practically implemented. Since May 2009, it has been deployed into a small network at Stanford University. In this network, FlowVisor acts as a transparent proxy between OpenFlow switches and other OpenFlow controllers, such as NOX [7]. Within FlowVisor, network resources are sliced in various dimensions, by bandwidth, topology or device CPU, etc. It provides much stronger separation between the control and data planes, by slicing the network resources into many orthogonal and independent slices through a transparent slicing layer [7]. It enforces block and rewrite control messages as they pass through FlowVisor [7]. A demo of network slicing in FlowVisor is shown in Fig. 1. Four isolated slices over the same physical network are demonstrated in Fig. 1. In these slices two of them are wireless and two are wired [10].

FlowN inherits similar strategies from FlowVisor, except it adopts a database to help process the mapping between the control plane and physical infrastructure [8]. As a consequence, it takes more time than FlowVisor to process the control message in small virtual networks. However, when more virtual networks are involved, say, more than 100 virtual networks, FlowN turns out to have slightly smaller latency than that of FlowVisor [8]. Note that FlowN with memcached is still slower than FlowVisor. Adopting database technology in virtual network mapping is a promising trend, but a lot of work is still required.

Control Strategy

The control strategies could be applied in different layers and different parts of the network architecture. For example, SoftRAN and OpenRAN [1] concentrated on providing software defined centralized control on radio access network; while RouteFlow [1] tries to execute remote centralized IP routing on computer networks. Odin [1] is a SDN framework based on WLAN and can achieve a virtual access point abstraction on physical switches. FlowN [8] is an advanced version of FlowVisor while OpenRoads is a SDWN application based on FlowVisor. Both FlowN and OpenRoads apply software defined control over physical switches [1]. Details

The core concept of SDWN is to split the data and control planes. The control plane is responsible for the network resource assignment and decision-making. Once a decision is set, it will communicate with the data plane through a particular protocol to finish the transmission.

[1] Although OpenFlow and SDN are often easily confused, they are two totally different concepts [6].

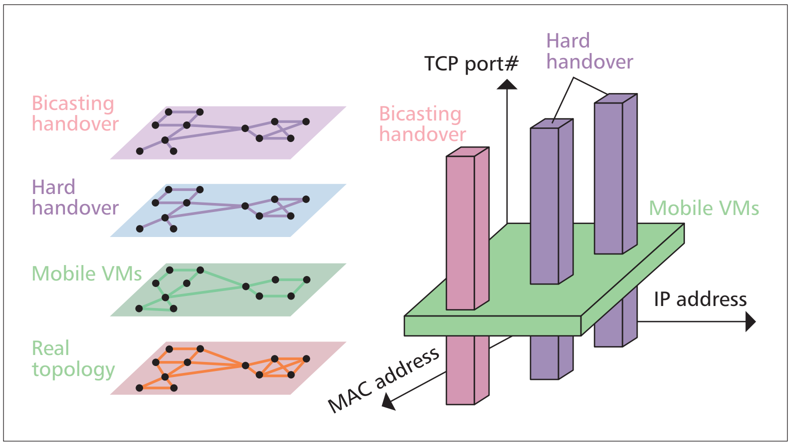


Figure 1. Demonstration of network slicing in FlowVisor [10].

of those control strategies are summarized in Table 1.

Just like FlowVisor, FlowN acts transparently to both control and data planes. Unlike FlowVisor, however, it uses databases to do the mappings between virtual and physical resources. The advantages of FlowN over FlowVisor fall into three categories: Firstly, FlowVisor does not supply a full virtualization over the physical resources as FlowN does. It simply slices them. Secondly, the mapping algorithm FlowVisor adopts is not efficient. For example, it has to iterate every node when doing a mapping. In FlowN, mapping can be done by a simple SQL query. It has been proven to be more scalable in a large network that may contain more than 100 virtual networks [8]. Thirdly, FlowN can supply securer network virtualization. Its physical mapping topology is invisible to tenants, whilst in FlowVisor mapping is exposed to tenants [8].

Network Description Based on RDF

Introduction of Semantic Technology in SDWN

In SDWN architecture design, a fundamental element is the information model, which describes all the resources of the network. This information model is the foundation of network virtualization. It describes both the physical layer infrastructure and its visualization. The ideal information model should be “technology independent, reusable, easily extensible and linkable to other existing models” [11]. Thus, the information models should be built based on an RDF syntax.

In the semantic web all of the information and services can be understood and used both by humans and computers. The semantic web is proposed with the intent of defining a universal model for resource or information description. Rather than trying to teach the machine how to understand human languages, semantic web defines a set of regulations that the machine can use to process the information of resources. The Semantic web is composed of three elements: metadata, RDF, and ontology.

Metadata is the “data about data” [4], which contains the information about the property of data. The RDF is a standard about making statements about the resources. It describes resources in the form of triples. A triple is a form of subject-predicate-object expressions. A set of triples is a graph. More details can be found in [4].

The topology information of a network can be collected and updated in real time from the network. It can be stored in an SQL-like database, in the form of a table with three columns, namely subject column, predicate column, and object column [11].

In semantic web, the ontology is also known as the vocabulary. Although ontologies do not have a universal definition to date, in the field of information science, an ontology defines a set of classes and the links between these classes. More specifically, in network description, the ontology describes a set of nodes and the links between them. It is a body of knowledge describing the network resource domain. The resource domain here is defined by physical entities in networks, such as switches, routers, devices for computation and storage, and the links between them [5].

An ontology describes the network resource with its own layered structure. There are upper level ontology and lower level ontology [15]. The Upper level ontology is the fundamental ontology which can be inherited by all sub-ontologies. It is technology independent, while the lower level ontology is technology specific.

Semantics provide a well-designed structure to describe the information for the resource domains and requirements from the clients. The main goal of network description language is to make sure that all the applications involved in a certain network have the same understanding of the network architecture topology and network resources [11].

Network Semantic Ontology Applications

Until now, the semantic ontology languages proposed to describe the network structure are numberless. These languages have different grammars, different parameters, and different specificities of application. Some of the most well-known ontology proposals are introduced below and their details are given in Table 2. However, a universally accepted language that is able to properly describe the resource of SDWNs has not been proposed yet to the best knowledge of the authors. Two typical ontologies are introduced below.

The Semantic Resource Description Language (SRDL) is the first ontology to describe both network and non-network resources in optical networks. However, the relationship between physical resources and virtualized resources are not provided by this ontology.

Resource Specification (RSpec) is an ontology language provided by GENI to describe SDN resources. In the new GENI v3 schema, RSpec is expanded to describe OpenFlow resources, with the support of FOAM. FOAM is an aggregate manager for OpenFlow resources in the GENI infrastructure.

Due to the complicated and variable wireless channel environment and the emerging new technology, building ontology for wireless networks is a long-term, arduous task. Many

SDWN Applications	Control Strategy	Description
FlowVisor [7]	FlowVisor can apply software control over any physical hardwares, such as routers, switches, by placing a slice layer between the control and data plane. It is already running on a production network in Stanford University.	FlowVisor positioned between an OpenFlow switch and several OpenFlow controllers. FlowVisor can slice any network resources in any control and data plane communication, for example, topology, bandwidth, device, CPU, or forwarding tables.
OpenRoads[1]	A wireless version of OpenFlow testbed, adopting FlowVisor to virtualize the physical WiFi switches, access points, and WiMax base-stations. It is currently running on a small network in Stanford, and is aimed to replace the present WiFi network.	OpenRoads consists of three layers: The flow layer, the slicing layer, and the controller layer. The slicing layer adopts FlowVisor to slice the physical resources. The NOX, an open source OpenFlow controller, is adopted in the controller layer.
FlowN [8]	FlowN is a SQL-based NOX controller, which could provide a full virtualization over a network of physical switches.	With the implement of database, FlowN could map the physical resource to virtual network topology by submitting SQL queries.
SoftRAN [1]	SoftRAN provides an SDN control plane in radio access network, by abstracting the base stations network.	SoftRAN achieves virtualization by abstract the base stations network into a virtual big base station, from which control decisions are sent to all the radio elements.
OpenRAN [1]	OpenRAN provides "match-action" control strategy	The architecture of OpenRAN consists of three parts: wireless spectrum resource pool, cloud computing resource pool, and SDN controller. It provides abstraction on the resources in both data and control plane.

Table 1. SDWN control strategies.

researchers of network semantic ontologies have put the development of wireless ontology on their agenda. Wireless ontology will be the new function of the future version of RSpec.

Implementation for Semantic Ontologies

There are some open-source tools to implement semantic web framework, such as Jena, Sesame [12]. Jena is an open-source semantic framework for JAVA. It supplies an API, which is able to extract data and compose RDF graphs, and an SPARQL engine, which will be introduced below, and a TDB for RDF storage and query [12].

As stated before, with RDF, the network resources can be organized as an SQL relational database with a table of three columns, namely, the subject column, the predicate column and the object column. The resource information can be considered as terms in the database. With ever increasing data being stored in this database, an SQL-like tool is required to search and locate the target information in the ocean of RDF data with higher efficiency. Many SQL-ish query languages have been proposed so far, such as SPARQL, rdfDB, RDQL, and SeRQL. SPARQL is one of the most widely used. It is a powerful query language for RDF, developed and recommended by W3C. Jena supplies a SPARQL engine by an application API — ARQ. It supports operations such as SELECT, CONSTRUCT, DESCRIBE, and ASK queries, etc. [12] Sesame is another powerful JAVA framework for RDF, which is very similar with Jena, except that Jena supports Web Ontology Language (OWL) and Sesame does not [13]. However, Sesame can also supply an easy API for RDF as Jena does. In addition, it can support two query languages, SPARQL and SeRQL, as well as Alibaba, an API that can map Java classes to ontology and generate Java source files from ontologies [13].

Performance Evaluation of Ontology

The ontology evaluation is the process to determine which resources the ontology defines correctly/incorrectly and those it does not define. It is a technical judgment for the content of the ontology. Evaluation of an ontology is not only necessary when it is implemented and published, but should be supported during the whole lifecycle of the ontology [14].

The criteria for performance evaluation are consistency, completeness, conciseness, expandability, and sensitiveness [14].

Consistency means every definition in the ontology is consistent and no contradictory conclusion can be deduced from other definitions and axioms. Ontology is consistent if and only if every one of its definitions is consistent.

Completeness is the basic requirement of ontologies, thus, incompleteness is a fundamen-

Semantic Ontologies	Description Technology	Description	Group
Network Description Language (NDL) [11]	Resource Description Framework (RDF)	To describe physical networks	Universiteit van Amsterdam, Netherlands
CineGrid Description Language (CDL) [5]	Web Ontology Language (OWL)	To describe media resources and services in CineGrid exchange	Universiteit van Amsterdam, Netherlands
Semantic Resource Description Language (SRDL) [5]	Web Services Modeling Ontology (WSMO)	To describe IT and network resources of a Service Oriented Optical Internet	University of Essex, UK
Media Application Description Language (MADL) [11]	Resource Description Framework (RDF)	Its purpose is similar with CDL SPARQL is adopted in MADL to supply query on the information models generated.	University of Essex, UK
Common Information Model (CIM) [11]	Unified Modeling Language (UML)	It provides a description of management information for networks of enterprises	DMTF, an industry standards organization
Virtual private execution infrastructure Description Language (VXDL) [11]	XML syntax	VXDL is a language for virtual network resources, it enables users to describe the virtual network topology, including virtual routers and timeline.	Universidade Federal de Santa Maria
Resource Specification (RSpec) [11]	XML syntax	It is a common language for describing resources, resource requests, and reservations. It is currently supported by PlanetLab, ProtoGENI, GENI v3, Omni and Flack, and Orca.	GENI: Global Environment for Network Innovations, a federated testbed. It provides a virtual laboratory for networking and distributed systems research and education.
A Service-Oriented Ontology for Wireless Sensor Networks proposed in [11]	Web Ontology Language Description Language (OWLDL)	It proposes a service-oriented sensor ontology for wireless sensor network environment	Cheju National University, Korea

Table 2. Semantic ontologies for network description.

tal problem in ontologies [14]. Completeness of an ontology is difficult to prove, however, we can prove the incompleteness of the ontology by proving the incompleteness of an individual definition.

Conciseness can be obtained for an ontology if and only if the following requirements are met:

- The ontology does not contain any unnecessary definitions or axioms;
- There are no redundancies between definitions;
- No redundancies can be inferred from other definitions and axioms.

Expandability means that new definitions can be added to ontology in the future without changing the already defined properties.

Sensitiveness is defined by the smallest change in a definition that can affect the already defined properties of the ontology [14].

Broadly speaking, evaluation can be divided into two parts: technical evaluation, which is carried out by developers, and users evaluation. From another perspective, the ontology evaluation approaches contain two aspects: ontology verification and ontology validation [14]. Ontology verification means that the ontology should be built correctly, which means its definition, in the natural language from the real world, matches the ontology requirements and competency questions of the target resource domain precisely. Ontology validation means that the ontology model matches the resource source in the real world correctly. Besides, ontology assessment is to judge the understanding, usability, usefulness, quality and portability of the definitions, taking the stand as the users. Various users and applications need various approaches to assess ontology [14].

The target context for ontology evaluation comprises the following aspects [14]:

- Each individual definition and axiom;
- All the groups of definitions and axioms that are stated explicitly in the ontology;
- Definitions imported from other ontologies;
- Definitions and axioms that can be inferred from other definitions.

In a nutshell, many reports have been published introducing ontology evaluation. However, very few offer details about exactly how the ontologies are evaluated or how the evaluation tools are built for different ontologies.

It is worth mentioning that, just like software testing, ontology evaluation should be performed as early as possible during the developing process of the ontology and should be carried out throughout the entire life circle of this ontology [14].

SDWN Architecture with Resource Description and Ontologies

In this section, we propose a novel SDWN architecture with the resource description module, as well as two ontologies, namely, resource ontology

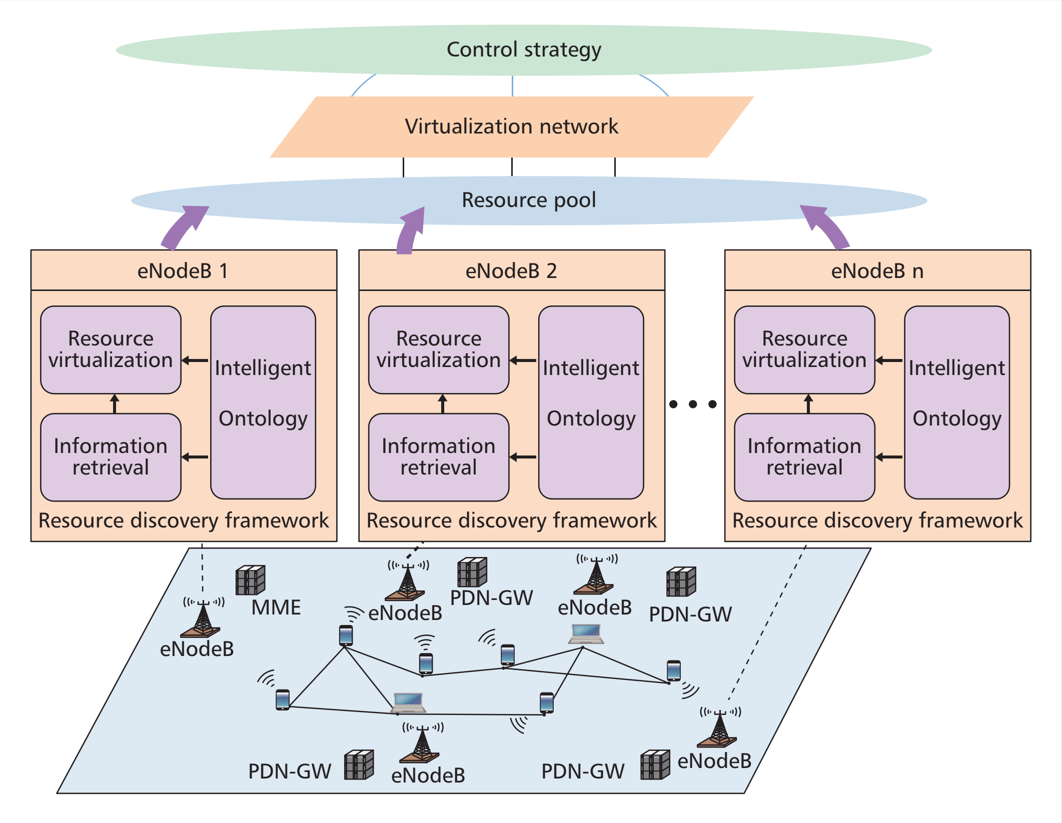


Figure 2. SDWN architecture with resource description function module.

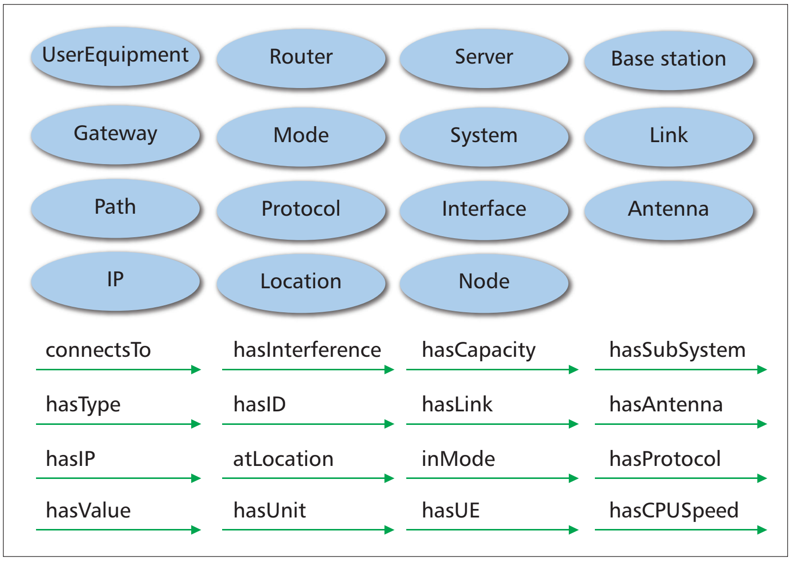


Figure 3. Resource ontology for SDWN.

and QoS ontology. Resource ontology is the main ontology for the resource domain in the network, while the QoS ontology describes the requirement for the communication services and the expectation from customers. These are designed for project "Towards Ultimate Convergence of All Networks (TOUCAN)" funded by EPSRC (No. EP/L020009/1). TOUCAN is proposing to develop full network convergence for any elements, domains, devices, and applications. It covers all technical domains, such as the mobile network, WiFi, optical network, and LiFi. In this article, we focus on SDWN. In our design, the control plane is embedded in eNodeB, along with the resource discovery module. This module acts as middleware, where applications can be built upon it.

Such architecture is shown in Fig. 2. The core module is the Resource Discovery Framework, where resource description function is performed. The framework consists of two parts: Information Retrieval and Resource Virtualization. An intelligent ontology translator based on RDF is adopted to syntactically analyze the network resources and customer requests, and translate them into machine-understandable syntax. Using the same ontology throughout the network, the controller can manipulate resources more efficiently.

The proposed resource and QoS ontologies are shown in Fig. 3 and Fig. 4, respectively. Resource ontology in Fig. 3 is the core ontology, which describes the main concepts of SDWN and can be inherited by other sub-ontologies. The number of classes and properties in the resource ontology is limited, because it cannot be technology-specific and it has to be general. Consequently, the ontology does not need to be revised every time a new technique or technology emerges. The terms in the ellipse are classes. They are subjects and/or objects, while the terms with underlines are predicates. With this ontology we can express the resources in RDF triples like this: "BaseStation A" "hasAntenna" "Antenna 1," where "BaseStation A" is an instance of class "BaseStation," "Antenna 1" is an instance of class "Antenna," and "hasAtenna" is a property of class "BaseStation."

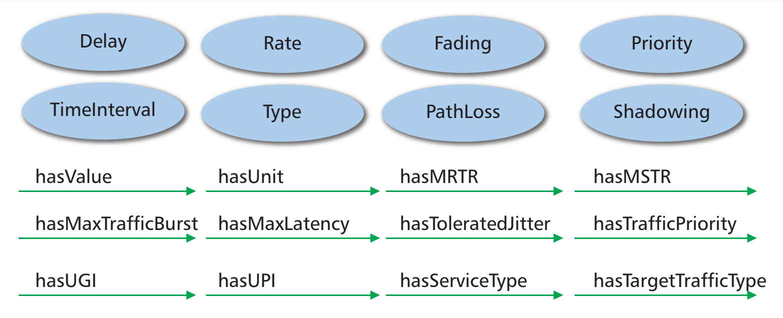


Figure 4. QoS ontology for SDWN.

One of the major goals of network resource description is to precisely depict the network resources and leave the details stay confidential to upper layers. On this ground, we divide the resources into technology-specific and technology-independent resources. Note that only the technology-independent resources are exposed to the upper layers. While resource ontology is technology-independent, the sub-ontologies are technology-specific. The QoS ontology in Fig. 4 is a sub-ontology applied to QoS domain. Just like its variable channel environment, the QoS requirements of SDWN are diverse and flexible. There are audio, video, audio/video streaming, HTTP, online gaming, and social network services, and each one of those has different QoS requirements.

The QoS ontology depicts the key QoS attributes in the IEEE 802.16e standard. The Maximum Sustained Traffic Rate (MSTR) stands for the capping rate level of service flow; Minimum Reserved Traffic Rate (MRTR) means the minimum rate of a service flow; Unsolicited grant interval (UGI) is the time interval between data grant opportunities for an downlink service flow; and Unsolicited polling interval (UPI) is the maximum time interval between polling grant opportunities for an uplink service flow. As stated above, the core ontology can be inherited by sub-ontologies. Thus, we can describe the QoS resources in RDF triples like "Node A" "hasMSTR" "10Mb/s." The subject "Node A" is an instance of class "Node" inherited from resource ontology. Object "10Mb/s" here is a value, rather than an instance of class. In nature language, this triple can be stated as "Node A has a MSTR of 10Mb/s" or "MSTR of Node A is 10Mb/s," but when we describe it in triples, it can be understood by computers.

Conclusions and Future Research Challenges

In this article, we have presented a comprehensive study on a SDWN virtualization strategy and resource description technology. The details of the control strategies, including the network virtualization design and the existing SDWN testbeds architectures are presented. The key technology to implement network resource description, semantic web technology, has been introduced in detail along with its three key elements including metadata, ontology, and RDF. We also have presented the performance evaluation methodology for SDWN at current stage. We have proposed a novel SDWN architecture adopting the semantic technology for the resource description. The ontologies for wireless network, including resource ontology and QoS ontology, have also been reported.

The ontologies proposed in this article are just sketches. We will continue to refine the ontologies by evaluating them on real-life applications. Eventually we will evaluate these ontologies using our own wireless testbed. Technical issues about how to extract resources from current network management system also need to be addressed.

References

[1] M. Yang *et al.*, "Software-Defined and Virtualized Future Mobile and Wireless Networks: A Survey," *Mobile Netw. Appl.*, 2014, pp. 1-15.

[2] C.-X. Wang *et al.*, "Cellular Architecture and Key Technologies for 5G wireless Communication Networks," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 122–30.

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

[4] E. H. Aart, *True Visions: The Emergence of Ambient Intelligence*, Springer-Verlag, 2006.

[5] O. Ntofon, D. K. Hunter, and D. Simeonidou, "Towards Semantic Modeling Framework for Future Service Oriented Networked Media Infrastructures," *Proc. 4th. IEEE Computer Science and Electronic Engineering Conf.*, Sept. 2012, pp. 200–05.

[6] N. McKeown *et al.*, "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Comp. Commun. Rev.*, vol. 38, no. 2, Apr. 2008, pp. 69–74.

[7] R. Sherwood *et al.*, "Can the Production Network Be the Testbed?," *Proc. Operating Systems Design and Implementation OSDI*, Oct. 2010, vol. 10, pp. 1–6.

[8] D. Drutskoy, E. Keller, and J. Rexford, "Scalable Network Virtualization in Software-Defined Networks," *IEEE Int'l. Comp.*, vol. 17, no. 2, Feb. 2013, pp. 20–27.

[9] X. Wei *et al.*, "Topology-aware Partial Virtual Cluster Mapping Algorithm on Shared Distributed Infrastructures," *IEEE Trans. Parallel and Distributed Syst.*, vol. 25, no. 10, Oct. 2014, pp. 2721–30.

[10] R. Sherwood *et al.*, "Carving Research Slices Out of Your Production Networks with OpenFlow," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 1, pp. 129-130, Jan. 2010.

[11] M. Ghijsen *et al.*, "A Semantic-Web Approach for Modeling Computing Infrastructures," *J .Comput. & Elect. Eng.*, vol. 39, no. 8, Aug. 2013, pp. 2553–65.

[12] J. J. Carroll *et al.*, "Jena: Implementing the Semantic Web Recommendations" *Proc. ACM 13th Int'l. World Wide Web Conf. Alternate Track Papers & Posters*, May 2004, pp. 74–83.

[13] S. Tomic, A. Fensel, and T. Pellegrini, "SESAME Demonstrator: Ontologies, Services and Policies for Energy Efficiency," *Proc. ACM 6th Int'l. Conf. Semantic Systems*, Sept. 2010, pp. 24.

[14] A. Gómez-Pérez, "Ontology Evaluation," *Handbook on ontologies*, S. Staab and R. Studer, Eds. Springer Berlin Heidelberg, 2004, pp. 251–73.

[15] C. Villalonga *et al.*, (2009, January), "Mobile Ontology: Towards A Standardized Semantic Model for the Mobile Domain," *Proc. ICSOC 2007*, Springer Berlin Heidelberg, Jan. 2009, pp. 248–57.

Biographies

Qianru Zhou (qz1@hw.ac.uk) received her M.Sc. degree in optical engineering from Beijing University of Posts and Telecommunications, China, in 2013. Currently, she pursues studies toward a Ph.D. in electrical engineering at Heriot-Watt University, U.K. Her research interests include software-defined wireless network, converged wireless and optical networks, semantic network description and ontology engineering for networks.

Cheng-Xiang Wang (cheng-xiang.wang@hw.ac.uk) received his Ph.D. degree from Aalborg University, Denmark, in 2004. He has been with Heriot-Watt University since 2005 and became a professor in 2011. His research interests include wireless channel modelling and 5G wireless communication networks. He has served or is serving as an Editor or Guest Editor for 12 international journals, including *IEEE Transactions on Vehicular Technology* (2011–), *IEEE Transactions on Communications* (2015–), *IEEE Transactions on Wireless Communications* (2007–2009), and *IEEE Journal on Selected Areas in Communications*. He has published one book and over 220 papers in journals and conferences. He is a Fellow of the IET.

Stephen McLaughlin (S.McLaughlin@hw.ac.uk) received the B.Sc. degree in Electronics and Electrical Engineering from the University of Glasgow in 1981 and the Ph.D. degree from the University of Edinburgh in 1989. He held a number of academic positions within the University of Edinburgh including that of Director of Research and Deputy Head of the School of Engineering. He has been the Head of School of Engineering and Physical Sciences at Heriot-Watt University since 2011. He is a Fellow of the Royal Academy of Engineering, Royal Society of Edinburgh, IET, and IEEE.

Xiaotian Zhou (xtzhou@sdu.edu.cn) received his B.S. degree in Electrical Information Engineering and the Ph.D. degree in Communication and Information Systems, both from Shandong University, China, in 2007 and 2013, respectively. He has been with the School of Information Science and Engineering at Shandong University as a lecturer since 2013. His current research interests include multiple access schemes, multiple antenna technologies, and 5G wireless communication networks.

When ICN Meets C-RAN for HetNets: An SDN Approach

Chenchen Yang, Zhiyong Chen, Bin Xia, and Jiangzhou Wang

ABSTRACT

With ever growing mobile Internet and the explosion of its applications, users are experiencing abundant services from different content providers via different network service providers in the heterogeneous network. The information-centric network (ICN) advocates getting rid of the current host-centric network protocol because information dissemination rather than end-to-end communication contributes to the majority of today's network traffic. Furthermore, it is better for network entities to converge as a whole in order to take advantage of open, scalable, and smart traffic transmission. The cloud radio access network is one of the emerging architecture evolutions on the wireless side for easier infrastructure deployment and network management. Therefore, when the information-centric network meets the C-RAN in the HetNet, it is worthwhile and consequential to integrate ICN protocol with C-RAN architecture to achieve more efficient communication and information management. Moreover, SDN has been recognized as another promising architecture evolution to achieve the flexibility and reconfigurability of dense HetNets; its inherent advantage lies in global uniform control of the wired network. Thus, any of ICN, C-RAN and SDN is the complementary to the others. In this article, we contribute to a fresh proposal for and elaboration of the integration of the ICN, C-RAN, and SDN for the HetNet to achieve a win-win situation. The vision of the proposed system is demonstrated, followed by the advantages and challenges. We further present a hybrid system in a large-scale wireless heterogeneous campus network.

BACKGROUND AND MOTIVATIONS

To meet the 1000-fold growth of mobile data traffic, a promising way is embedding mass small cells into the cellular network, that is, the heterogeneous network (HetNet) [1]. Heterogeneity will be an essential feature of the future radio access network (RAN) consisting of different kinds of devices. For highly flexible interfaces among different devices, software defined networking (SDN) has been recognized as a basic enabler to achieve the flexibility and reconfigurability of dense HetNets [2]. Furthermore, under the pressure of inter-/intra-cell coordination and tremendous deployment cost in the large-scale HetNet, cloud RAN (C-RAN) has emerged [3]. C-RAN is widely regarded as a green radio architecture to cope with the increasing network data traffic without incurring significant additional capital and operational expenses. In addition to the evolution of the network architecture, human behavior and the service property are being widely considered to unleash the ultimate potential of networks, for example, the cache-enabled information-centric network (ICN) [4]. The state of the art is elaborated in the following from the perspective of network architecture and network protocol, respectively.

FROM THE PERSPECTIVE OF NETWORK ARCHITECTURE

The HetNet is an expansion of the existing network in terms of size and type: Compared to the fourth generation/Long Term Evolution (4G/LTE) network, the fifth generation (5G) network is expected to attain a 1000-fold capacity increase measured in bits per second per Hertz per square meter, 5 times reduced latency, and 10 times longer battery life [5]. Motivated by the ever-increasing pressure to enhance network capacity, cellular networks are overlaid by a wide variety of proliferating infrastructure layers including outdoor pico/indoor femto base stations (BSs), relays, distributed antennas [6], WiFi, and device-to-device (D2D) nodes [7]. Heterogeneity will be a key attribute of future networks.

HetNets bring the cell site closer to users and shorten the radio transmission distance, yielding significant cost reduction and capacity enhancement. Furthermore, small cells are deployed as complements of the macrocellular networks for seamless coverage and for intentional offloading of traffic of the macrocell. Besides, advanced technologies such as high-order spatial multiplexing multiple-input multiple-output (MIMO) come along with higher spectral efficiency (SE) in the HetNet. However, there are still problems:

- Resource management and load balancing are two important issues for the successful deployment of HetNets. The inter-/intra-

Chenchen Yang, Zhiyong Chen, and Bin Xia (corresponding author) are with Shanghai Jiao Tong University.

Jiangzhou Wang is with the University of Kent.

0163-6804/15/$25.00 © 2015 IEEE

cell coordination and convergence are nontrivial operations in this distributed and large-scale environment.
- HetNets enable more flexible and economical deployments of new infrastructure instead of tower-mounted macro systems, but are not sufficient to achieve the flexibility and reconfigurability of networks due to nonuniform and curable device interfaces.
- The increasing demands on the backhaul network need to be evaluated when deploying more small cell sites.

Trivial expansion of the size and type of the existing network can temporarily relieve the pressure, but disruptive evolutions of the network architecture are urgent for further capacity extension.

C-RAN is the architecture evolution on the wireless network side: Unlike the existing HetNet, where each transport node (e.g., macro/micro BS, relay, etc.) posesses computing resources for baseband processing in its local cell site, C-RAN aims to separate radio access units (RAUs) and baseband processing units (BBUs). C-RAN consists of three fundamental components:
- Distributed RAUs equipped with remote radio heads (RRHs) at the cell site
- The centralized BBU pool formed in a data center cloud
- The intermediate entity unit providing the high-bandwidth low-latency links to connect RRHs and the BBU pool

C-RAN gathers a large number of BBUs into a logically centralized BBU pool. The centralized and coordinated baseband processing/scheduling in the BBUs allows for soft and dynamic cell reconfiguration. The lightweight RRHs are implicitly decoupled from the BBU pool in the C-RAN, yielding easier deployment of different types of cells in the HetNet and reducing the energy consumption of each site. RRHs cooperate flexibly and seamlessly in the HetNet so that the SE and the capacity can be improved significantly.

However, advantages come with challenges: full-scale coordination leads to high computational overhead in the BBU pool, especially for a large-scale network; real-time virtualization and high-bandwidth links are urgent to achieve reliable connection and mapping between the BBU and the RRH. C-RAN only focuses on the radio air interface and cannot solve the problems emerging in the core network (CN) or other upper layers. But C-RAN can be regarded as a specific sub-unit of SDN, which is elaborated in the following.

SDN is the architecture evolution on both the wired and wireless sides: Tremendous numbers of infrastructure equipments have been patched into the network through the present. Traditionally, each device is an ossified unit of hardware and software developed by unique manufacturers. Every new feature or capability expansion requires a professional manual update and reconfiguration of the software stack with proprietary languages. SDN decouples the control plane from the data plane via *controllers* and allows software to be designed independently from the hardware, simplifying network access, design, and operation [8]. It fosters network virtualization and makes it possible to evolve the consolidation of different types of network equipment, which reduces the network complexity for sophisticated heterogeneous systems.

SDN provides an open interface with centralized controllers remotely controlling the data forwarding tables in the network entities such as switches, routers, and access points. The network entities are virtualized and solely use forwarding strategies defined by external programmable controllers. Logically centralized controllers possess the flexibility to dynamically and automatically redirect or optimize the traffic load and resource management with a global view of the system.

However, SDN was originally designed for wired networks. There are inherent weaknesses to achieve the SDN ideas in wireless networks due to various challenges. Software defined wireless networks (SDWNs) need to define slices, which requires wireless channels to be isolated so as to provide non-interfering networks to different coordinators. Handoff situations should be considered for wireless HetNets with smaller cells and richer access technologies. The status and locations of all network entities should be reported to SDN in real time only based on which controllers make decisions efficiently. These challenges are not trivial operations for the wireless medium. Much work still needs to be done for the wired and wireless sides to be seamlessly integrated as a whole in the SDN system.

The status and locations of all network entities should be reported to SDN in real time only based on which controllers make decisions efficiently. These challenges are not trivial operations for the wireless medium. Much work still needs to be done for the wired and wireless sides to be seamlessly integrated as a whole in the SDN system.

From the Perspective of Network Protocol

ICN provides a new basis on how information can be labeled and distributed across networks: Despite the tremendous amount of data traffic in the network, only a few contents are frequently accessed by users [9]. A small portion of popular contents contribute to the majority of traffic during a period of time [10]. Furthermore, today's network is increasingly occupied by information dissemination rather than end-to-end communications. Thus, caching the most popular contents in the RAN or the evolved packet core (EPC) can reduce redundant access and duplicated transmission [11]. The pivotal role caching technology can play for the 5G wireless network is elaborated in [11, 12]. ICN emerges as a promising candidate for full use of in-network caching and multicast mechanisms [13].

Different from the exiting host-centric networks where users have to transmit/receive information to/from a particular computing entity (host or server), they are increasingly interested in transmitting/receiving information wherever they may be located. The ICN decouples information from its location and sources by defining named data objects (NDOs). Widespread caching and broadcasting allow users to get information from the optimal node based on the name and/or location of the information. Providers and requesters are no longer connected in the traditional pair-wise and time-synchronized mode; they are decoupled in terms of time and space. There is no need for them to

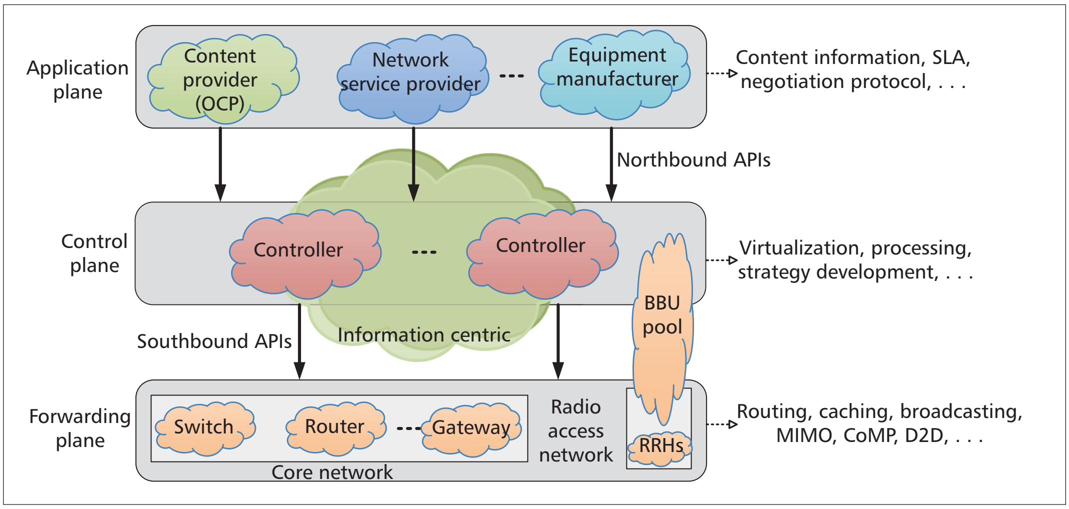


Figure 1. The information-centric SDN with C-RAN.

We propose to achieve the information-centric approach on the SDN architecture to unleash the ultimate potential of the network, and C-RAN undertakes the role and task of SDN on the wireless side. Few existing studies has considered the coexisting network and we will give an integrate sight of them.

know each other's location or to be online at the same time.

Contribution

As mentioned above, the widespread in-network caching of ICN brings great opportunities of cooperative communication among entities in the high-density wireless HetNet, while C-RAN is emerging for easier infrastructure deployment and system management. Then integrating ICN protocol into the C-RAN architecture can achieve better communication with efficient distribution of information via ubiquitous cache-enabled devices. At the same time, the burden of the upper layer (e.g., the CN) will be significantly reduced when user requests can be responded to immediately by the entity (e.g., the BBU) that has cached a copy of the requested information. However, C-RAN solely focus on the radio air interface regardless of the states of upper layers. Coincidentally, SDN was born with the original purpose of dealing with wired networks, but will meet more difficulties when introducing its concepts to wireless networks. Moreover, the ICN protocol is available to the SDN as well. Therefore, appropriate combination of ICN, CRAN, and SDN is a promising way to achieve mutual and complementary benefit.

We organize the remainder of this article as follows. First, we describe our vision of the coexisting system of the SDN, ICN, and H-CRAN. Then the advantages and main challenges are elaborated. A large-scale exemplary campus network is presented. Finally, we give a brief summary of this article.

The Vision of the Future Network System

The information-centric concept is better suited to today's use, which mainly consists of information distribution rather than host-centric communications. For the long tail style of the user request statistical property, in-network caching and multiparty communication can be fully leveraged via replication and cooperative models. We propose to achieve the information-centric approach in the SDN architecture to unleash the ultimate potential of the network, and C-RAN undertakes the role and task of SDN on the wireless side. Few existing studies have considered the coexisting network, and we give an integrated view of them.

Information-Centric Software Defined Networking

A high-level view of information-centric SDN architecture is shown in Fig. 1, where the application, control, and forwarding planes are included. The forwarding and control plane components can evolve independently via defining standard application programming interfaces (APIs) between them.

The application plane consists of the application and service provided by the following three entities: the content provider (CP), which contains the traditional CP and the emerging over-the-top (OTT) content provider (OCP) such as Google, Amazon, and Netflix; the network service provider (NSP); and the equipment manufacturer (EM). The control plane consists of a set of distributed but logical centralized controllers. A controller can control a quantity of network entities, and a network entity can be controlled by different but logically centralized controllers. However, there is no need to provide a strictly consistent centralized view to each controller, which would cause processing overload and additional cost when the network expands. A control functionality can even be completed by different controllers. Therefore, selecting an appropriate consistency level (e.g., the weakest possible consistency level) of controllers is an important design consideration in SDN to preserve scalability [14]. The CN and RAN are included in the forwarding plane. The infrastructure of the CN can be virtualized and controlled by the control plane. The centralized BBU pool and distributed RRHs of C-RAN are deployed to achieve efficient collaboration among different cells in HetNets. Therefore, independent networks can be reconfigured flexibly and automatically on the same physical infrastructure under the "soft" strategy change of controllers. The role that each plane plays in the network is discussed below.

- Application plane: The CP (e.g., OCP) distributes the NDO in its server to end users with the help of NSPs. The objective/content information is decoupled from its location and sources by the naming approach of ICN; thus, it is more suitable for the CP to understand and forecast user behaviors (e.g., the request frequency and content population distribution). Given the service-level agreement (SLA) provided by different NSPs, the CP chooses the optimal NSPs via negotiation to distribute the objective. Decoupling the software and hardware via SDN may help CPs and NSPs get rid of the shackles of the EM. It is not what the EM wants, and negotiation protocol should also be developed among them. The application requirements are negotiated based on the protocol running in the underlying control plane through open APIs.
- Control plane: The control plane running in a network operating system (NOS) is the core part of the architecture, and logically centralized controllers complete the infrastructure virtualization, programming abstractions, and even the content naming, addressing, and matching procedures for the ICN. The controller exploits complete knowledge of the system and gets consolidated control functions to facilitate network reconfiguration and management via the NOS. For example, wired backhual and wireless bandwidth owned by the NSPs can be dynamically allocated to the CPs and users, respectively, based on the negotiation protocol running in the controller.
- Forwarding plane: Since the control function has been extracted and integrated into the control plane, the forwarding plane thus consists of simplified and virtualized network devices that solely provide information switching and forwarding. However, a typical HetNet may have tens of thousands of devices, and the sheer number of control events generated at that scale is enough to overload any controller. Pushing all the control functionality to centralized controllers makes them the potential bottleneck for network operation when the HetNet scales.

To reduce the burden of the controllers as well as to get rid of the aforementioned nontrivial challenges of SDWNs, we propose to deploy the wireless side of SDN in the concept of C-RAN where the BBU pool has both control and data forwarding functions. The logically centralized BBU pool has a network-wide view of the RAN and CN, yielding the seamless integration of the wired and wireless sides of SDN. Collaboratively controlled by the controllers and the BBU pool, the NDO of ICN can be flexibly and optimally distributed and stored in diverse devices of the CN and RAN via caching and broadcasting.

The Deployment of C-RAN in the Information-Centric SDN

Embedding C-RAN into the information-centric SDN is a promising way to reduce the burden of controllers, and integrate the wired and wireless sides of SDN seamlessly. It is more efficient and easier to set up better broadcasting/multicasting and in-network caching mechanisms (e.g., information placement and replacement strategy), which are cornerstones of ICN.

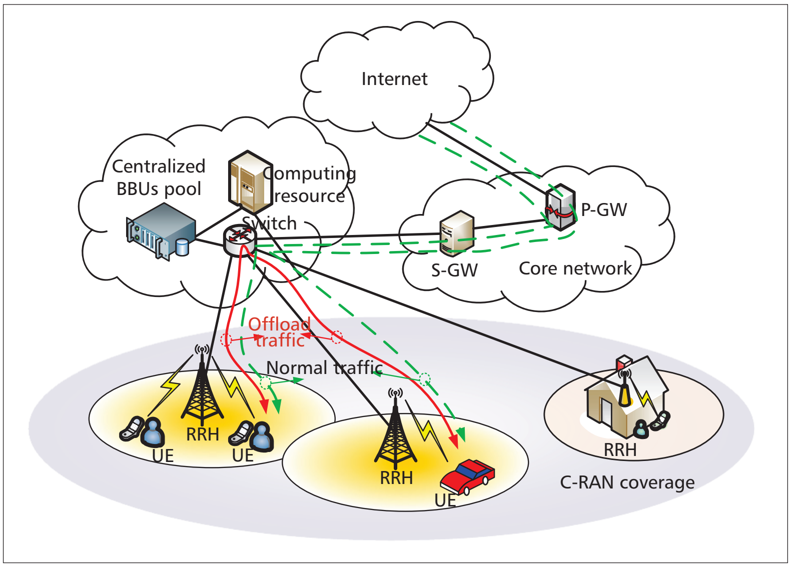


Figure 2. An overview of offload traffic.

Without doubt, the explosive traffic demand increases the delivery cost of mobile Internet content on the operator. With the help of cloud resources, C-RAN involves all the computing, storage, and content-aware elements required to functionally and efficiently deliver the content, which enables new smart traffic of ICN, such as offload traffic and cache traffic as depicted in Figs. 2 and 3, to meet the challenge.

Offload Traffic: A promising solution that offloads the data traffic bypassing the CN of the operator and Internet. As illustrated in Fig. 2, when the user equipment (UE) communicates, normal traffic is required to traverse the RAN, CN, and Internet before reaching other UE. For the offload traffic, the centralized BBU pool enables a Cloud-identification (ID)-enabled UE, connected via RRH, to directly connect to other UE configured with the identical Cloud-ID without passing the CN and Internet. Note that such offload traffic is established at the RAN, which is different from the traditional offload approach enabling data traffic offload via a wireless local area network (WLAN). Through enabling the data traffic closer to the edge of network (i.e., RAN), the operator can relieve the pressure on the CN elements such as serving gateways (S-GWs) and packet data gateways (P-GWs), yielding cost savings in gateway upgrades. Moreover, the offloaded traffic allows direct communications with UEs within one C-RAN network, reducing the content latency and efficiently improving the quality of service (QoS) and user experience. Besides, two UEs can even communicate directly without traversing the BBU pool in a mature ICN, which will bring more benefit.

Cache Traffic: Figure 3 demonstrates a positive solution that selectively pushes and stores mobile contents at the centralized BBU pool, supporting direct access to the contents without going through content servers in the Internet via

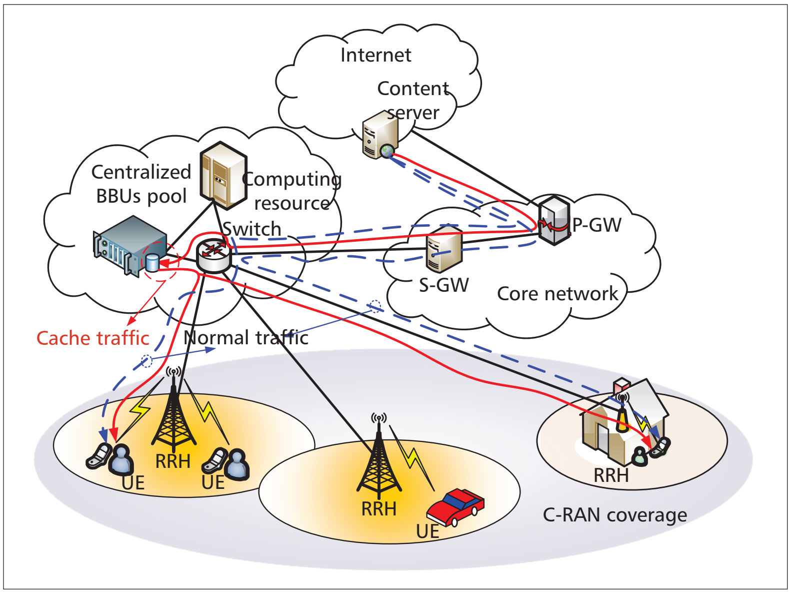


Figure 3. An overview of cache traffic.

the CN. Meanwhile, following the uncannily accurate Moore's law, the storage and processing capability of the intelligent devices becomes stronger and stronger. When the network is at off-peak traffic load (e.g., at night) the most frequently accessed content can be broadcast and then cached at the BBU pool. The user can obtain the requested content immediately from the BBU pool when the cache hit event occurs. As a result, the deployment of the C-RAN infrastructure in the information-centric SDN leverages the storage at the network edge for caching to provide more and richer services, whereby the cache is independent of the application and can be applied to various sources, including user-generated content, which is an attractive feature of ICN.

In addition to caching in the BBU pool, popular contents can also be pushed and cached in the RAN or EPC when the network is idle. It offloads the corresponding redundant traffic and duplicated access in the CN and RAN, which can reduce pressure on the backhaul link and the processing overhead in the BBU pool and controllers. For example, NSPs can proactively push and cache popular contents published by CPs to cache-enabled intermediate proxy devices or some of the UEs in advance; then the UE can get the content from the optimal node of ICN via D2D or a proxy-UE link when time is available.

Two Dimensions and the Main Challenges

In order to get rid of the obstacles to the development of SDWN and make use of the efficient transmission of ICN, we propose a new ICN, SDN, and H-CRAN integrated system, as described in Fig. 4, which encompasses the complete system platform. The novelty of the proposed concept can be elaborated from the following two dimensions.

The Architecture Dimension: Infrastructures are deployed with the devices the software of which is decoupled from the hardware and centralized to the control entity (i.e., BBU pool and controllers). Different CPs provide abundant services to subscribers with the help of different NSPs based on the negotiation protocol, which is well known to the control plane. Controllers are the intermediate entities between the application plane and the forwarding plane. The necessary state message of the control and forwarding planes should be reported to controllers via APIs. Appropriate forwarding strategies are developed by the control plane with a global view of the network (the BBU pool can cooperate with controllers). The control plane distributes the incoming traffic provided by the servers of the application plane to the forwarding plane. Entities such as routers only take the strategy developed by the controllers and forward traffic to C-RAN. Non-uniformly distributed traffic and requests can be balanced in the BBU pool and then transmitted to UEs.

The Information Dimension: The main abstraction of information in ICN is the NDO for identifying information independent of its location or publisher. The role the NDO plays in the ICN is as important as that IP plays in the current host-centric Internet. When information labeled with unique NDOs has been published by the server in the application plane or the UE in the RAN, its transcript can be held by ubiquitous cache-enabled nodes (e.g., nodes in EPC or RAN) afterward. For example, the transcript can be cached in the BBU pool, proxies, or UEs based on the caching strategy and the content access protocol developed by the BBU pool, offloading the processing burden of controllers and the traffic of the CN. Obviously, the design of the information dimension can affect the strategy and procedure in the architecture dimension.

However, there are still some open problems. The main challenges and future research directions can be summarized as follows.

Processing Overload: The huge performance gain of the proposed system mainly comes from the centralized and coordinated signal processing at the cloud BBU pool and controllers. However, full-scale coordination in large-scale HetNets requires the processing of very large network information such as the channel matrices. Processing overload could be a major problem for such a centralized environment as the network size grows, even though the BBU pool and controllers are able to cooperate with each other. One way to tackle this problem is to decrease the redundant and duplicated flow before they enter the control plan with appropriate strategy (e.g., optimal caching in ICN). Or, as [14] proposes, proactively install rules on the devices to eliminate some control requests, while yielding some loss of control precision and reactivity.

Backhual and Fronthaul: The common assumption that the data can be routed to the RANs without backhaul and fronthaul limitation is not valid for the future high-density HetNet, where a large number of nodes need to access information. High-capacity wired backhual and fronthaul are needed for the connections of the application, control, and forwarding planes (e.g.,

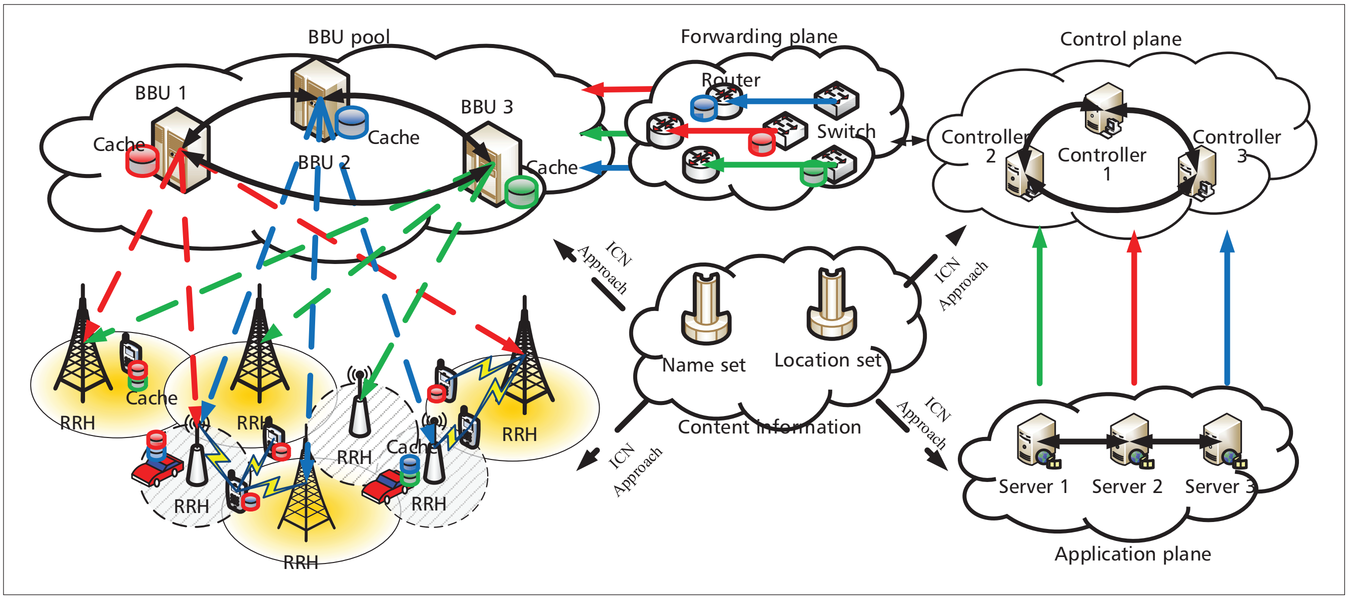


Figure 4. The coexisting system of the SDN, ICN, and H-CRAN.

the controllers and CPs; BBUs and RRHs). Few existing studies jointly consider the wired backhaul and fronthaul and the radio resource management in HetNets when optimizing system performance.

Universal Caching: Universal in-network caching is a salient feature of the proposed system, for which the caching strategy and content replacement are important issues. The caching strategy decides what, where, when, and how the information should be cached to achieve optimal system performance. Furthermore, the cached information should be consistent with that in the server and publisher so that invalid information can be replaced with an appropriate mechanism in real time. User behaviors (e.g., the access frequency follows the long tail distribution) are not yet fully investigated, especially in dense mobile HetNets, but it can significantly influence the performance of the caching strategy.

Access Protocol and Data Routing: The proposed system should apply to any access protocol in a sophisticated HetNet, not just a specific protocol (e.g., HTTP, LTE). It should thus provide a uniform content distribution paradigm underlying all access protocols. Moreover, flexible and convenient information-aware mechanisms should be developed for data routing based on the location-independent name of the information. Then the subscriber can be responded to by the optimal node that has cached the information rather than only by the original publisher.

There are some other challenges such as mobility and security management. With the RRH deployment more and more concentrated in the future HetNet, it becomes more frequent for a user to hand off between different RANs. On the other hand, similar to traditional security techniques such as transport layer security (TLS), equivalent security measures should be developed for naming objects, caching, and communication in the proposed ICN. The key issue is that requesters can get content from any ubiquitous cache-enabled entity that has capabilities other than those of the relatively uniform host servers, so the security measures should be based on the content itself (e.g., naming) rather than the communication channel or path.

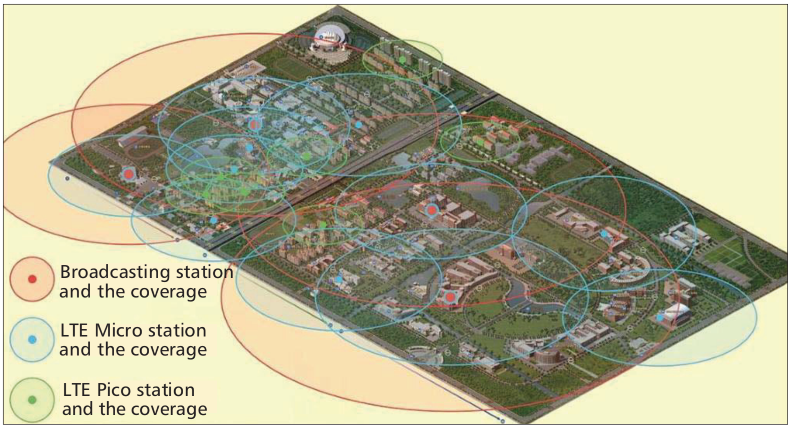

Broadcasting station and the coverage
LTE Micro station and the coverage
LTE Pico station and the coverage

Figure 5. A large-scale wireless heterogeneous campus network.

Implementation of an Exemplary Network

We have established a large-scale wireless innovation campus network with 3 km^2 coverage in Shanghai Jiao Tong University. The heterogeneous campus network, consisting of a digital broadcasting system, an LTE cellular system, and a WiFi system, is demonstrated in Fig. 5. The important features of the wireless heterogeneous campus network are as follows:

- There are more than 80 LET micro and pico stations with blanket coverage of the campus. More than 2500 WiFi access points are patched across the entire campus. Three

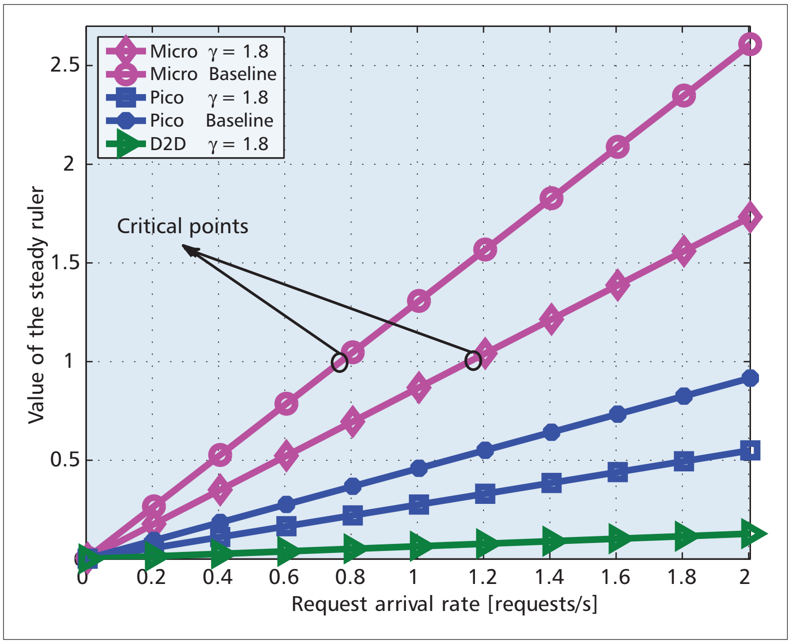


Figure 6. The throughput gain for the cache-enabled network compared to that of the baseline.

additional stations constitute a single-frequency digital broadcasting network.

- All cellular RF transceivers are connected to a computing center through over 66 km of fiber, yielding a true C-RAN.

Especially for the cooperative caching communication of SDN, each of the pico stations and some UEs are cache-enabled. Based on the strategy developed by the controllers with the global view of the network, popular contents can be broadcast to and then cached at the cache-enabled nodes. The cached content can be reused for frequent access. Besides cellular communication, there is a D2D link (i.e., from cache-enabled users to requesting users) for the content sharing, yielding a three-tier HetNet (i.e., micro BSs–users, pico BSs–users, D2D transmitters–users). The UE in overlapping coverage associates with the optimal node according to the instruction of the controllers, and the UE can obtain the requested content immediately from its local caching disk if the content has been cached.

We verify the throughput gain of the coexisting system in [15] compared to the baseline where there is no in-networking cache. The content access is triggered according to the well-known Zipf distribution with parameter γ. Larger γ implies that fewer contents account for the majority of the requests. In Fig. 6 we circle out the critical point deciding the maximum throughput of the network based on whether the steady ruler is larger than 1. We observe that when γ = 1.8 the throughput gain is 53.9 percent that of the baseline. Moreover, the pico and D2D tiers are far from the fully loaded state when the micro tier comes to the critical steady state. Therefore, more appealing performance improvements can further be realized with appropriate resource scheduling and load balancing mechanisms.

Conclusions

In this article, to cope with the rapid increase of network traffic and the change of communication mode in the HetNet, we have presented a coexisting system where the novel ICN, C-RAN, and SDN are integrated to reap mutual and complementary benefits. First, advantages and drawbacks of each novel kind of network are summarized, based on which we have proposed an information-centric SDN architecture consisting of three planes. Network entities in each plane have been explained and their corresponding roles clarified. Then we have elaborated the benefit the C-RAN can bring to the information-centric system via examples of offload traffic and cache traffic. Furthermore, the vision of the network platform for the proposed system is described from architecture evolution and information dissemination points of view, at the same time challenges are enumerated. Finally, a demo large-scale wireless heterogeneous campus network is presented.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China (Grant No. 61301115) and the National High Technology Research and Development Program of China under 863 5G Grant 2014AA01A702.

References

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

[2] 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.

[3] C. Liu *et al.*, "The Case for Re-configurable Backhaul in Cloud-RAN Based Small Cell Networks," *Proc. IEEE INFOCOM*, Apr. 2013, pp. 1124–32.

[4] N. Golrezaei *et al.*, "Femtocaching and Device-to-Device Collaboration: A New Architecture for Wireless Video Distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, 2013, pp. 142–49.

[5] Metis, "Scenarios, Requirements and KPIs for 5G Mobile and Wireless System," May 2013.

[6] J. Wang, H. Zhu, and N. Gomes, "Distributed Antenna Systems for Mobile Communications in High Speed Trains," *IEEE JSAC*, vol. 30, May 2012, pp. 675–83.

[7] Q. Li *et al.*, "5G Network Capacity: Key Elements and Technologies," *IEEE Vehic. Tech. Mag.*, vol. 9, no. 1, Mar. 2014, pp. 71–78.

[8] A. Gelberger, N. Yemini, and R. Giladi, "Performance Analysis of Software-Defined Networking (SDN)," *IEEE 21st Int'l. Symp. MASCOTS*, Aug. 2013.

[9] M. Cha *et al.*, "I Tube, You Tube, Everybody Tubes: Analyzing The World's Largest User Generated Content Video System," *Proc. ACM SIGCOMM Internet Measurement*, Oct. 2007.

[10] K. Wang, Z. Chen, and H. Liu, "Push-Based Wireless Converged Networks for Massive Multimedia Content Delivery," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, May 2014, pp. 2894–2905.

[11] X. Wang *et al.*, "Cache in The Air: Exploiting Content Caching and Delivery Techniques for 5G Systems," *IEEE Trans. Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 131–39.

[12] J. Andrews *et al.*, "What Will 5G Be?," *IEEE JSAC*, vol. PP, no. 99, June 2014.

[13] G. Xylomenos *et al.*, "A Survey of Information-Centric Networking Research," *IEEE Commun. Surveys Tutorials*, vol. 16, no. 2, July 2014, pp. 1024–49.

[14] S. Yeganeh, A. Tootoonchian, and Y. Ganjali, "On Scalability of Software-defined Networking," *IEEE Commun. Mag.*, vol. 51, no. 2, Feb. 2013, pp. 136–41.

[15] C. Yang *et al.*, "Analysis on Cache-Enabled Wireless Heterogeneous Networks," http://arxiv.org/abs/1508.02797, Aug. 2015.

Biographies

CHENCHEN YANG (zhanchifeixiang@sjtu.edu.cn) received his B.Eng. degree in the School of Electronics and Information in 2013 from Northwestern Polytechnical University (NPU), Xi'an, China. He is currently a Ph.D student in the Institute of Wireless Communications Technology (IWCT), Department of Electronic Engineering, Shanghai Jiao Tong University (SJTU), China. His research expertise and interests include heterogeneous networks, cooperative communications, and computing communications.

ZHIYONG CHEN (zhiyongchen@sjtu.edu.cn) received his Ph.D degree from the School of Information and Communication Engineering in 2011 from Beijing University of Posts and Telecommunications. From 2009 to 2011, he was a visiting Ph.D student in the Department of Electronic Engineering at the University of Washington, Seattle. He is currently an assistant professor in the Department of Electronic Engineering, SJTU. He served as Publicity Chair for IEEE ICCC 2014 and a TPC member for major international conferences.

BIN XIA (bxia@sjtu.edu.cn) received his Ph.D. degree in electrical engineering from the University of Hong Kong in 2004. From 2005 to 2012, he was a senior research scientist and project manager at Huawei Technologies Co. Ltd. Since 2012, he has been a professor with the Department of Electronic Engineering at Shanghai Jiao Tong University. His research interests include MIMO, CDMA, OFDM, cross-layer design, radio resource management, and radio network architecture.

JIANGZHOU WANG (J.Z.Wang@kent.ac.uk) is currently the chair of telecommunications and head of the Communications Research Group with the School of Engineering and Digital Arts, University of Kent, United Kingdom. He has published over 200 papers in international journals and conferences in the areas of wireless mobile communications and has written/edited three books. He received the Best Paper Award from IEEE GLOBECOM 2012. He is a Fellow of the IET and was an IEEE Distinguished Lecturer from January 2013 to December 2014. He serves/has served as an Editor or Guest Editor for a number of international journals, such as *IEEE Transactions on Communications* and *IEEE Journal on Selected Areas in Communications*. He was Technical Program Chair of IEEE WCNC 2013 in Shanghai and Executive Chair of IEEE ICC 2015 in London.

Software Defined Mobile Networks: Concept, Survey, and Research Directions

Tao Chen, Marja Matinmikko, Xianfu Chen, Xuan Zhou, and Petri Ahokangas

Tao Chen, Marja Matinmikko, and Xianfu Chen are with VTT Technical Research Centre of Finland.

Xuan Zhou is with Zhejiang University.

Petri Ahokangas is with Oulu Business School.

ABSTRACT

This article provides a brief overview on the current development of software-defined mobile networks (SDMNs). Software defined networking is seen as a promising technology to manage the complexity in communication networks. The need for SDMN comes from the complexity of network management in 5G mobile networks and beyond, driven by increasing mobile traffic demand, heterogeneous wireless environments, and diverse service requirements. The need is strong to introduce new radio network architecture by taking advantage of software oriented design, the separation of the data and control planes, and network virtualization to manage complexity and offer flexibility in 5G networks. Clearly, software oriented design in mobile networks will be fundamentally different from SDN for the Internet, because mobile networks deal with the wireless access problem in complex radio environments, while the Internet mainly addresses the packet forwarding problem. Specific requirements in mobile networks shape the development of SDMN. In this article we present the needs and requirements of SDMN, with particular focus on the software-defined design for radio access networks. We analyze the fundamental problems in radio access networks that call for SDN design and present an SDMN concept. We give a brief overview on current solutions for SDMN and standardization activities. We argue that although SDN design is currently focusing on mobile core networks, extending SDN to radio access networks would naturally be the next step. We identify several research directions on SDN for radio access networks and expect more fundamental studies to release the full potential of software-defined 5G networks.

INTRODUCTION

After three decades of evolution, mobile networks are moving into the fifth generation (5G) [1]. In Europe, the 5G infrastructure public private partnership (PPP) defined the following ambitious performance goals for 5G networks: 10 to 100 times higher typical user data rate, 10 to 100 times more connected devices, 10 times lower network energy consumption, less than 1 ms end-to-end latency, and 1000 times higher mobile data traffic per geographical area. To satisfy these new requirements, we will witness more disruptive changes in mobile networks.

One prominent feature would be the full embrace of software-defined networking (SDN) design in mobile networks. Indeed, the software-defined design of mobile networks could effectively tackle the most difficult problems in current cellular and other wireless access networks, to manage heterogeneity, complexity, and consistency in the network and further catalyze fundamental changes in the mobile ecosystem.

While the definition of software-defined mobile networks (SDMNs) remains open, SDN for the Internet is widely used as the reference model for SDMN design. The essential ideas from SDN for the Internet are the decoupling of the data and control planes, and the use of logical centralized control to manage the forwarding problem in large scale networks. Clearly SDMNs will not be a simple extension of the SDN concept for the Internet, because the radio access in mobile networks is different from the routing in the Internet. Software-defined features in SDMNs shall satisfy specific needs of mobile networks.

The evolution of computer systems may give some hints on the design of SDMN. Nowadays computer systems have advanced to such a level that the performance of a smartphone easily surpasses that of the supercomputer decades ago. This evolution is firmly backed by advances in the development of the operation system (OS) and programming languages. The OS successfully decouples high-layer programs from low-layer hardware implementation. The function abstraction and modular design of the computer system, along with the paradigm shift toward object-oriented programming, establish design principles to master the complexity in computer systems. Computer science was born to build the theoretic foundation that further guarantees the innovation and continuous evolution of computer systems.

The same trend fits the development of

0163-6804/15/$25.00 © 2015 IEEE

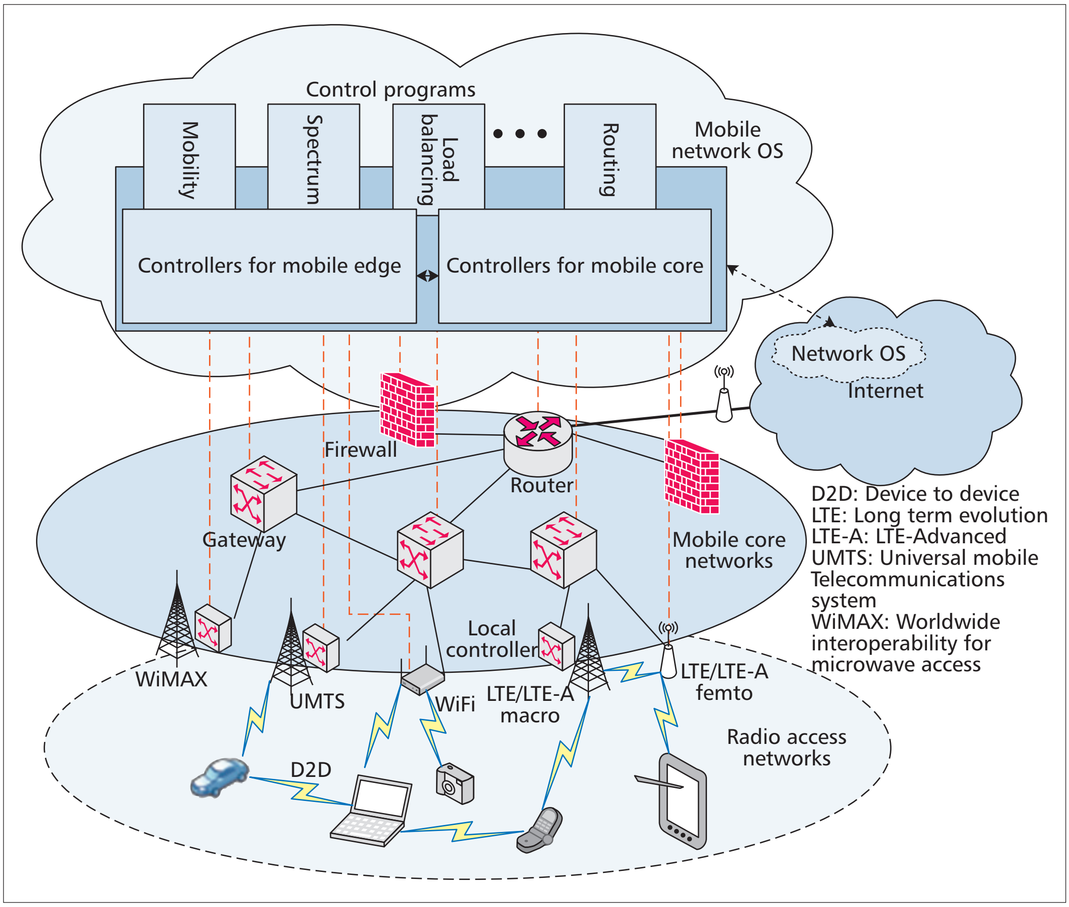


Figure 1. Illustration of software-defined mobile network.

The software-defined approach allows spectrum to be managed more efficiently, since the logical centralized control can be aware of the spectrum usage in the network, and allow proper spectrum mobility and effective implementation of spectrum sharing strategies in SDMN.

mobile communications. Interactions and complexity in current heterogeneous mobile networks (HMNs) are very similar to the early stages in the history of computer systems. We need to rethink the design of mobile networks. Referring to SDN design for the Internet, a simplified SDMN architecture is illustrated in Fig. 1.

This article provides an overview on the current development of SDMN. As for mobile networks there is a clear division between radio access networks (RANs) and mobile core networks (CNs), we focus the discussion on the RAN side. We start by listing the driving forces and enabling technologies for SDMNs. Following that, we examine fundamental problems in SDMNs, propose an SDMN concept, and briefly analyze the business impact of SDMNs. The current SDMN research is briefly surveyed. Finally, we identify several important research directions in SDMNs.

Driving Forces toward SDMN

The development of 5G networks and the new trend in spectrum regulation are the strong driving forces to make mobile networks software-oriented.

Requirements of 5G Networks

5G networks aim to provide native support for a variety of services with major differences in quality of service (QoS). In addition to applying advanced physical layer technologies and using new spectrum, 5G networks need an orchestrated service platform to effectively and efficiently coordinate network resources. The increasing complexity of 5G networks calls for new network design for flexibility and cost efficiency. It requests similar design principles to those driving the evolution of computer systems.

Flexible Spectrum Management

In 5G spectrum availability is one of the key challenges to fulfil the enormous mobile traffic demand. The access to new bands and flexible spectrum sharing become very necessary in 5G networks. The most promising approaches for sharing are the spectrum access system (SAS) and licensed shared access (LSA), where the licensed shared spectrum is made available to mobile operators. To improve spectrum reuse, mobile networks need to be aware of the spectrum usage, traffic load, and network conditions. The software-defined approach allows spectrum to be managed more efficiently, since the logical centralized control can be aware of the spectrum usage in the network, and allow proper spectrum mobility and effective implementation of spectrum sharing strategies in SDMNs.

Key Enablers of SDMNs

Technical advances in SDN, network functions virtualization (NFV), cloud computing, and fog computing provide technical enablers for SDMN.

The origin of the SDN concept can be traced back to the 1990s. However, the SDN concept received global attention after the introduction of the OpenFlow concept in 2006. SDN technologies are promoted and standardized by the Open Network Foundation (ONF). So far it has more than 150 member companies.

SDN

The origin of the SDN concept can be traced back to the 1990s. However, the SDN concept received global attention after the introduction of the OpenFlow concept in 2006. SDN technologies are promoted and standardized by the Open Network Foundation (ONF). So far it has more than 150 member companies, and SDN enabled network devices are commercially available.

The success of SDN comes from the systematic abstraction of complex networking problems in the Internet, which turns previous distributed networking problems into a logical centralized problem, where the rich theories and optimization tools well developed by computer science can be applied. The separation of data and control planes, open control interfaces for network devices of different vendors, and programmable control make a disruptive paradigm shift in the networking business.

The same level of complexity exists in HMNs, but has not been systematically studied. SDN gives rise to fundamental new thinking on the design of mobile networks. The key question is how to extract the simplicity from complex radio access problems and build principles to guide the mobile network design.

NFV

NFV is the recent initiative from the telecom industry to achieve more flexible and cost-efficient network architecture [2]. The key idea of NFV is to virtualize network functions (e.g., NAT, firewall, and load balancers) and implement them in industry standard high volume servers instead of proprietary hardware. A virtualized network function (VNF) can be run across different software and processes through virtualization techniques. The focus of NFV is currently on infrastructure networks. It will be an important technology to redesign the cellular CN. The combination of NFV and SDN bring new architecture design to mobile networks.

Cloud Computing and Fog Computing

The development of SDN is tightly connected to cloud computing, since cloud computing makes large-scale logical centralized control solutions feasible. Cloud computing allows centralized data storage and processing, and online access to computer resources through remotely deployed server farms and software networks [3]. It aims to maximize the effectiveness of resource sharing. Cloud computing is one enabler of NFV. The resource sharing nature of cloud computing is suitable for joint signal processing and control among RANs. Indeed, the cloud RAN (C-RAN) concept promoted by China Mobile and other major telecom companies is one concrete example of applying cloud computing in mobile networks.

However, the traditional cloud computing architecture may have a problem in meeting the strict latency requirements for fine timescale control functions in SDMNs. It is reasonable to move the logical centralized control close to the edge in mobile networks. Fog computing could fill this gap for better architecture design of SDMNs. Fog computing is a variant of the cloud computing concept [4] that uses the computer resources and storage at the edge of a network for a substantial amount of communication, storage, control, and configuration. In mobile networks, fog computing can be utilized for the control and joint signal processing at the RAN level to serve densely deployed cells, while cloud computing can be used for control in CNs for packet processing and forwarding. The integration of fog computing and cloud computing may lead to an end-to-end (E2E) SDN solution for mobile networks.

The SDMN Concept and Business Impact

Technique Aspects

The design of SDMNs for RANs needs to address three fundamental problems. In this section we first examine these problems and explain why software-defined approaches will provide a good solution. Then we present the SDMN concept and briefly analyze the business impact.

The first problem concerns distributed network states in HMNs, in which each network and even each base station (BS) makes its own resource allocation decision with limited state information from others. The spectrum reuse in mobile networks calls for optimization and control across cell borders. However, current mobile networks have limited support for network-level coordination. It is beneficial to use the network view, as in SDN for the Internet, for optimal control and coordination in mobile networks. Considering that even in a single BS hundreds of parameters need to be tuned, the information presented in the high-level network view needs to be simplified. The design principles from OSs may provide the answer to this problem: the abstraction of system functions and behaviors to shield the details of the low-layer implementation. Network views at the high control layer are built on the proper abstraction of lower layers through defined open control interfaces and primitives. It turns distributed control problems in mobile networks to a centralized coordination problem, allowing more fine-grained optimization. Figure 2 shows a preliminary simulation study on the performance of the abstracted network view applied on small cell networks. It illustrates the potential of the logical centralized control in small cell networks.

The second problem addresses the network configuration among multiple network entities. In future mobile networks, the performance of network entities is more likely to be coupled due to spectrum reuse, mobility, and traffic offloading. There is an increasing demand to configure network entities in RANs coherently, similar to those in SDN for the Internet. Network configuration needs to be done among coupled network entities by control algorithms based on the logical centralized control framework. To make it scalable, this high-level configuration should not go into detail on the lowwer-layer implementation. This means the network configuration only defines the preferred behaviors of lower layers. The behaviors are mapped to the configuration of lower layers through middleware. The new

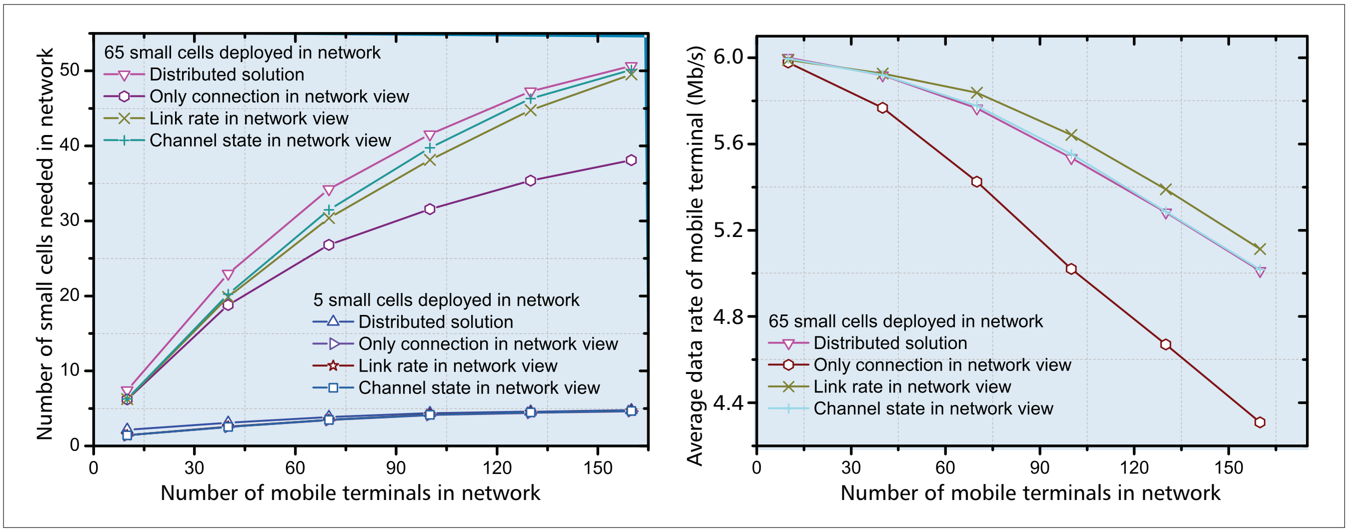


Figure 2. Performance of the network view for small cell network energy saving. In the network small cell BS and mobile terminals (MTs) are deployed ad hoc. MT has the date rate requirement. The objective is to find the smallest number of BSs to support MTs while other BSs go to sleep. In the distributed solution an MT connects to the nearest BS. In the network view approach three network views are evaluated: in the connection only case the network view only knows if an MT can connect to a BS; in the link tate case the interference-free link rate is known in the network view; in the channel state case the average channel state of links is known in the network view.

network configuration approach needs deep understanding of network behaviors, especially cooperative behaviors, such as in coordinated multipoint (CoMP) transmission and inter-cell interference coordination (ICIC).

The third problem addresses fine-grained cooperation among different entities in the network, for instance, CoMP and enhanced ICIC in Long Term Evolution (LTE) networks. ICIC may require joint resource allocation and signal processing among involved entities. Currently this kind of cooperation is mainly addressed by self-organizing network (SON) features implemented by a bottom-up approach targeting the specific problem. As fine-grained cooperation becomes a common feature in mobile networks, a top-down approach is needed to incorporate the cooperation in the native system design. With this in mind, we should define open control interfaces, programmable SON features, and proper network abstraction to implement and control different kinds of network cooperation in a software manner. It will provide flexibility and reduce the cost to implement new network features in mobile networks.

We believe software-defined design will expand in mobile networks through two dimensions, as illustrated in Fig. 3. The vertical dimension handles coarse network coordination among cells and networks. The horizontal dimension targets fine-grained network cooperation among network entities. In the vertical direction, the common control requirements and functions need to be extracted from different mobile networks. The systematic abstraction and modularity of network functions will enable hierarchical control architecture, in which the high control layer controls lower layers through defining behaviors without the need to know their specific implementation. It will allow programmable control to coordinate HMNs. In the horizontal direction, cooperative behaviors among network entities will be abstracted. Following that, common control protocols and open interfaces will be developed to support different cooperative behaviors under same software-defined architecture. It will facilitate the implementation of cooperative functions and enable the programmable SON for fine-grained low-layer cooperation in the network.

Different from the software-defined design in CNs, software-defined features at the RAN side are focused on joint resource allocation, spectrum management, mobility, and cooperative functions among HMNs. The benefits of SDMN are highlighted in Table 1.

Business Aspects

SDMNs will enable an open network architecture that allows vertical and horizontal control flexibility in mobile networks. The concept breaks the boundary of a single RAN, and provides the extendibility and programmability for control and coordination in HMNs. The development of SDMNs will have a profound business impact on the value chain of mobile industry.

Operators will be able to reduce capital expenditure (CAPEX) and time to deploy services, because new open control interfaces and software-defined control will reduce time and cost to reconfigure and optimize RANs, and to introduce new network features. By the software-defined control architecture, it will allow more efficient use of spectrum, energy resource, as well as the network infrastructure so as to reduce operational expenditure (OPEX).

For network equipment vendors, because of open control interfaces, they will have more flexibility to implement network functions, making their equipment easily integrated into operators' networks. It will reduce the time to market of their product and allow open innovation by embracing competition.

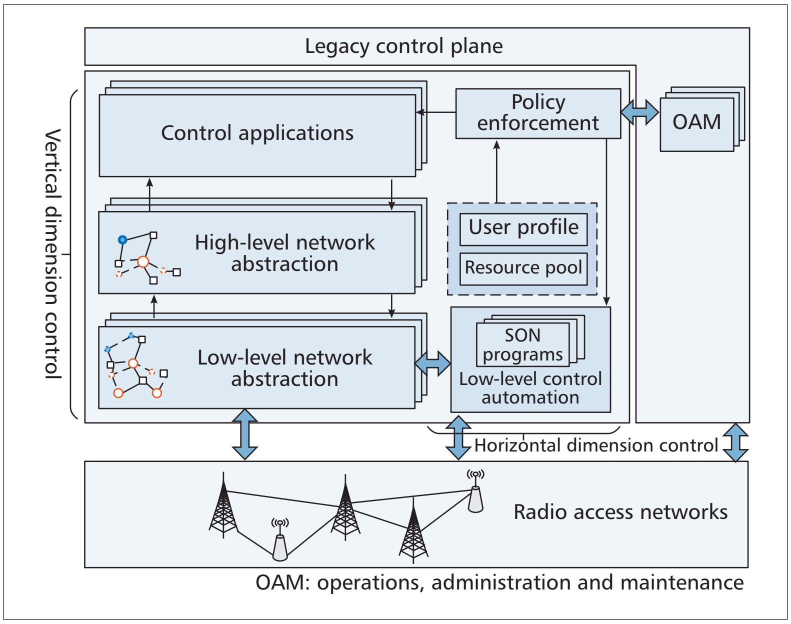

Legacy control plane
Control applications
Policy enforcement
OAM
Vertical dimension control
High-level network abstraction
User profile
Resource pool
Low-level network abstraction
SON programs
Low-level control automation
Horizontal dimension control
Radio access networks
OAM: operations, administration and maintenance

Figure 3. Example of software-defined control architecture for mobile networks.

For content providers, SDMNs could provide interfaces to allow over-the-top (OTT) services to be better provided. RANs can be tuned for OTT services according to software-defined control. It provides content providers and mobile network operators (MNOs) a cooperation framework to benefit their business.

For end users, the improved coordination among different mobile networks will provide smooth network experience. An SDMN is able to provide customized control to satisfy certain subscriber groups, and to deploy new services in shorter times. It will bring operators more value-added services for business growth.

The benefits of SDMNs for different business players is summarized in Table 2.

Current Research

The ongoing research on SDMNs is briefly summarized in this section. The survey is by no means complete, but aims to identify important research directions. We divide the research into three main directions: ideas derived from SDN for the Internet, centralized solutions similar to C-RAN, and approaches applied at the mobile edge. Note that a solution can be a combination of these three categories.

SDN-Oriented Approaches

The majority of SDMN research derives from the original SDN concept. The common features are the decoupling of the control and data planes, and the use of logical centralized control.

OpenRoad is a very early study on this topic [5]. It is a mobile version of SDN that uses OpenFlow for control, FlowVisor for network slicing, and NOX as the network operation system to support programmable control in WiFi and WiMAX networks. OpenRoad allows different control algorithms to concurrently run in one network, and thus realizes network slicing, one of the key features in SDN. Network slicing is extended to cellular networks in [6], where the network virtualization substrate and CellSlice are proposed to virtualize wireless resources and allow virtual MNOs to coexist in a single physical network.

Softcell is the first effort to extend the SDN concept to the mobile CN [7]. It applies SDN principles to redesign the control plane of the CN. The centralized controller and the flow concept allow the previously centralized packet processing in CNs to be distributed among separated packet processing middleboxes, and thus improve scalability and flexibility. Pentikousis *et al.* proposed another flow-based forward model, named MobileFlow, to facilitate the deployment of new services and network features in the mobile CN [8]. An OpenFlow controller is introduced in [9], which allows the separation of control and data plane in CNs of LTE networks and moves core control functions in CN to the cloud for reliability and scalability. A similar idea was proposed in [10], where the mobile network SDN controller governs not only LTE, but also Universal Mobile Telecommunications System (UMTS), WiFi, and other wireless networks.

Furthermore, an SDN-based plastic architecture was introduced for 5G networks [11], with the aim to support a heterogeneous set of services with flexibility. It introduces a clean-slate data plane design, and the SDN controllers at three levels: device, mobile edge, and CN, respectively. This design avoids the use of tunnel protocols for mobility and allows backward compatibility with 4G networks.

C-RAN-Oriented Approaches

C-RAN-oriented approaches centralize not only the control but also part of the radio signal processing in the network. Note that C-RAN, although it does not need to be implemented by the SDN approach, will definitely benefit from SDN design.

SoftRAN is one of the early proposals under this approach [12]. It virtualizes the RAN into a single virtual BS, performing resource allocation, mobility, load balancing, and other control functions in a single place. The centralized control plane of the network takes advantage of full network knowledge for global optimization. To solve the latency problem, time-critical controls remain at the local BS.

A recent design proposed by Arslan *et al.* combines the centralized signal processing in C-RAN and the programmable feature at the fronthaul [13]. Software-defined fronthauls (SDFs) form a fronthaul network, where jointly processed radio signals are forwarded to fronthauls by the centralized control. The control architecture is similar to SDN for the Internet. The programmability in the fronthaul network allows practical fine timescale physical layer cooperation like CoMP. It provides potential for fine-grained RAN optimization in extremely dense wireless networks.

Mobile-Edge-Oriented Approaches

Mobile-edge-oriented approaches apply SDN design at the RAN. The need for this approach comes from the adaptation of the air interface as

	Software-defined design at RAN	Software-defined design at CN
E2E services	Network awareness to improve QoS, support services with network reality	Traffic steering, QoS support
Heterogeneous network integration	Open control interfaces, network awareness, joint network configuration	Traffic steering to improve network resource utilization
Spectrum management	High level spectrum provision, network awareness, facilitate SAS and LSA	Traffic load awareness for spectrum allocation
Mobility	Network awareness, resource reservation, mobility prediction	Logic centralized control, reduce mobility overhead
RAN cooperation	Programmable SON, open cooperation interface	Traffic steering to better support CoMP and other cooperative techniques

Table 1. Benefits of SDMN for service support and network function implementation.

Role	Cost structure related benefits	Revenue structure related benefits.
MNOs	Decreased CAPEX through easier RAN configuration and optimization; decreased OPEX through more efficient use of spectrum, energy and infrastructure	New connectivity and content services; context information (big data, user profile) services; Business to Business (B2B) commerce services related to sharing
Equipment vendors	Easier and faster integration of technologies and services	Flexibility to add new functionality/services
Content providers	OTT cooperation framework	Opportunities for providing network as a service locally
B2B end users	Customized control features	Easier and faster adoption and integration of new services
Business to consumer (B2C) end users	Seamless and smooth network experience, new services	Easier and faster adoption and integration of new services

Table 2. Business benefits of SDMN.

well as the fine-grained radio function coordination in dense wireless networks. To adapt air interface behaviors to network conditions, Bianchi proposed the MAClet concept, which allows the central controller to dynamically change the MAC process in air interfaces (e.g., from contention-based medium access to time-division multiple access) [14]. The SDF proposed in [13] is also a mobile edge solution that brings programmability to the radio fronthaul. While more research is expected in this direction, we believe SDN principles will be widely applied in the radio architecture design of future wireless networks. This trend has been observed in ongoing major European 5G research projects.

In Europe the world's largest 5G research initiative, known as the Horizon 2020 5G-PPP Initiative, was launched in July 2015. The purpose of this initiative is to lay the foundation for 5G mobile communication networks. Among 19 funded projects in the first phase of this initiative, the METIS-II project will focus on overall 5G RAN design, the 5G-NORMA project will be dedicated to a novel radio adaptive network architecture, the COHERENT project will concentrate on a uniform control platform for heterogeneous RANs, and the XHual and 5G-XHual projects aim to develop adaptive and sharable 5G transport network solutions. These projects have special interests to investigate and implement software-defined mobile control in 5G RANs. We will see the SDMN design and development from them over the next three years.

Standardization Activities

While the standardization of SDMN has yet to come, the efforts from the Open Networking Foundation (ONF), European Telecommunications Standards Institute (ETSI), and Third Generation Partnership Project (3GPP) are paving the way for the realization of SDMNs.

ONF is the organization promoting and standardizing SDN and OpenFlow technologies. In 2014, ONF formed the Wireless and Mobile Working Group (WMWG) to extend ONF based SDN technologies to the wireless and mobile domain. The current tasks in WMWG include defining the use cases, and architectural and protocol requirements for the ONF extension to wireless backhaul networks, cellular CNs, and other wireless access technologies. Three major use cases identified by WMWG are wireless transport networks, cellular access networks, and enterprise networks.

In 2012, ETSI formed the NFV Industry Specification Group (ISG) to promote the IT virtualization technologies in the telecom industry. We have mentioned NFV as an enabler for

SDMN aims to provide programmable and unified control solutions for 5G networks. We believe network abstraction will be essential for the SDMN architecture design. For this, we need to build the theoretic foundation for network abstraction and derive control principles for SDMN.

SDMNs earlier. Currently, the NFV ISG has four working groups: infrastructure architecture, management and orchestration, software architecture, and reliability and availability; and two expert groups: security and performance, and portability. NFV ISG is not a standards development organization, but will be the main driving force for the standardization of NFV technologies.

The first discussion on 5G was started in 3GPP in November 2014. The potential 5G study items were discussed in 3GPP Service and System Aspects (SA) Technique Specifications Group (TSG), in which the user perceived performance, business enabling capabilities, cost, operation, and energy efficiency were highlighted. Given the key requirements identified by 5G promotion groups and the research trend in the mobile industry, SDN is expected to receive the main attention in the radio network architecture design.

More information on standardization of SDMNs can be found in [10, 15].

Challenges and Research Directions

Architecture Design

The SDMN aims to provide programmable and unified control solutions for 5G networks. We believe network abstraction will be essential for the SDMN architecture design. For this, we need to build the theoretical foundation for network abstraction and derive control principles for SDMNs. First, spectrum sharing behaviors in mobile networks need to be abstracted for high-level coordination. Second, the abstraction should consider control operations with different timescales. This leads to the question of implementing certain control functions at local or central points. Third, in the ideal case the SDN-based control plane for SDMNs should behave like an OS, where low-layer implementations are encapsulated by abstraction and seen by the high-level control plane through application programming interfaces (API). This creates a general programmable control framework across different physical entities. Programmable control combined with network virtualization will enable very flexible control functions for different groups of network entities or end users. It allows fast deployment of new control algorithms and services. To encapsulate the low-layer implementation, the programmable SON will play an important role in handling the automation.

Advanced Radio Resource Management

The decoupling of the control and data planes in SDMNs needs to consider different timescales in radio resource management. For instance, the frame scheduling in a BS will occur in milliseconds, while the spectrum assignment among small cells may change hourly according to busy hours. In small cell networks the control and data plane separation will certainly benefit the radio resource utilization in the network. However, radio resource management needs to carefully be split locally or remotely in order to match new network architecture. The first step is to model behaviors of the physical and link layers so that we can build an open but accurate control framework for different RANs. For instance, for dynamic CoMP in SDMNs, the high-level control plane only needs to know which BSs are selected.

RAN sharing of mobile network operators adds a huge requirement on radio resource management, because the network slicing to support RAN sharing does not simply mean spectrum slicing, but also the logical isolation of wireless resources. Spectrum sharing implies that the spectrum access of different virtual mobile networks is coupled. The challenge is how to abstract spectrum sharing properly so that the high level of the control framework is able to share radio resources among virtual networks for guaranteed services.

E2E SDN Solution

SDN solutions for RANs, mobile CNs, and the Internet have different control targets. For RANs the main objective is to coordinate and control the radio resource; for mobile CNs it is to orchestrate the packet processing for mobility, billing, and service provisioning; for the Internet the main target is for effective and efficient packet forwarding. Because of different control requirements, the integration of SDN in different network segments toward an E2E solution is extremely challenging. However, the demand is high as different services, especially time-critical services from vehicles or the Internet of Things, will be provisioned in the same network infrastructure. The key to enabling E2E SDN solutions lies in software oriented design. It requires systematic software-oriented thinking to integrate different SDN solutions.

Network Intelligence

Along with SONs, network intelligence is important to support automation in SDMNs. It will be embodied in multiple layers of SDMNs. In the physical and link layer, cognitive radio may apply to improve spectrum sharing of SDMNs. At the high level of the control framework, network intelligence, particularly the deep learning and artificial intelligence, will find the place on predictive modeling, traffic prediction, and dynamic configuration of network resources according to learning from the environment, traffic patterns from previous traffic data statistics, and even users' behaviors in network access. The control framework should provide open interfaces to support network intelligence at different layers. It is also important to thoroughly evaluate the developed network intelligent methods to avoid over-control of networks.

Conclusion

5G development has been put on the schedule of the ICT industry around the world. Following the development history of previous mobile systems, in 10 years we will see the deployment of 5G networks. The industry has achieved consensus to redesign the radio network architecture for 5G. Software-defined design has been identified as an important evolution path for 5G networks. In this article, we summarize the key research problems in SDMNs. The current research shows open problems and the diversity of solutions. We believe SDMNs will be the next big thing for the mobile industry. New thinking and more fundamental research are expected to consolidate SDMN design and development.

References

[1] J. G. Andrews *et al.*, "What Will 5G Be?," *IEEE JSAC*, vol. 32, no. 6, 2014, pp. 1065–82.

[2] B. Han *et al.*, "Network Function Virtualization: Challenges and Opportunities for Innovations," *IEEE Commun. Mag.*, vol. 53, no. 2, 2015, pp. 90–97.

[3] S. Azodolmolky, P. Wieder, and R. Yahyapour, "Cloud Computing Networking: Challenges and Opportunities for Innovations," *IEEE Commun. Mag.*, vol. 51, no. 7, 2013, pp. 54–62.

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

[5] K.-K. Yap *et al.*, "OpenRoads: Empowering Research in Mobile Networks," *ACM SIGCOMM Comp. Commun. Rev.*, vol. 40, no. 1, pp. 125–126, 2010.

[6] R. Kokku *et al.*, "CellSlice: Cellular Wireless Resource Slicing for Active RAN Sharing," *Commun. Systems and Networks 2013*, 2013, pp. 1–10.

[7] X. Jin *et al.*, "SoftCell: Scalable and Flexible Cellular Core Network Architecture," *9th ACM Conf. Emerging Networking Experiments and Technologies*, 2013, pp. 163–74.

[8] K. Pentikousis, Y. Wang, and W. Hu, "MobileFlow: Toward Software-Defined Mobile Networks," *IEEE Commun. Mag.*, vol. 51, no. 7, 2013, pp. 44–53.

[9] J. Kempf *et al.*, "Moving the Mobile Evolved Packet Core to the Cloud," *Wireless and Mobile Computing, Networking and Commun.a 2012*, 2012, pp. 784–91.

[10] C. Bernardos *et al.*, "An Architecture for Software Defined Wireless Networking," *IEEE Wireless Commun.*, vol. 21, no. 3, 2014, pp. 52–61.

[11] R. Trivisonno *et al.*, "SDN-Based 5G Mobile Networks: Architecture, Functions, Procedures and Backward Compatibility," *Trans. Emerging Telecommun. Technologies*, vol. 26, no. 1, 2015, pp. 82–92.

[12] A. Gudipati *et al.*, "SoftRAN: Software Defined Radio Access Network," *2nd ACM SIGCOMM Wksp. Hot Topics in Software Defined Networking*, 2013, pp. 25–30.

[13] M. Arslan, K. Sundaresan, and S. Rangarajan, "Software-Defined Networking in Cellular Radio Access Networks: Potential and Challenges,"*IEEE Commun. Mag.*, vol. 53, no. 1, 2015, pp. 150–56

[14] G. Bianchi *et al.*, "MAClets: Active MAC Protocols over Hard-Coded Devices," *Proc. 8th Int'l. Conf. Emerging Networking Experiments and Technologies*, 2012, pp. 229–40.

[15] F. Granelli *et al.*, "Software Defined and Virtualized Wireless Access in Future Wireless Networks: Scenarios and Standards," *IEEE Commun. Mag.*, vol. 53, no. 6, 2015, pp. 26–34.

Biographies

Tao Chen [S'05, M'10, SM'13] (tao.chen@vtt.fi) received his B.E. degree from Beijing University of Posts and Telecommunications, China, in 1996, and his Ph.D. degree from the University of Trento, Italy, in 2007, both in telecommunications engineering. He is currently a senior researcher at VTT Technical Research Centre of Finland. He is the project coordinator of the EU H2020 5G-PPP COHERENT project. His current research interests include software defined networking for 5G mobile networks, dynamic spectrum access, energy efficiency and resource management in heterogeneous wireless networks, and social-aware mobile networks.

Marja Matinmikko (marja.matinmikko@vtt.fi) is a senior scientist at VTT Technical Research Centre of Finland. She received her M.Sc. degree in industrial engineering and management, and her Dr.Sc. degree in telecommunication engineering from the University of Oulu. She is the coordinator of the Finnish project consortium on Cognitive Radio Trial Environments (CORE). Her current research interests include technical, trialing, regulatory, and business aspects of spectrum sharing for mobile communications.

Xianfu Chen received his Ph.D. degree from Zhejiang University, Hangzhou, China, in 2012. He is currently a senior scientist at VTT Technical Research Centre of Finland Ltd., Oulu. His research interests cover various aspects of wireless communications and networking, with emphasis on software-defined radio access networks, green communications, centralized and decentralized resource allocation, and the application of artificial intelligence to wireless communications.

Xuan Zhou received his Ph.D. in communication and information systems from Zhejiang University. From 2009 to 2014, he worked as a system engineer in China Mobile Zhejiang Company. Since 2014, he has been a solution architect in the Service Provider Operation Lab (SPO Lab) of Huawei Technologies. His research efforts focus on innovative service scenarios in 5G and NFV/SDN.

Petri Ahokangas (petri.ahokangas@oulu.fi) received his M.Sc. (1992) and D.Sc. (1998) degrees from the University Vaasa, Finland. He is currently adjunct professor and senior research fellow at Martti Ahtisaari Institute, Oulu Business School at the University of Oulu. His research interests are in how innovation and technological change affect international business creation, transformation, and strategies in highly technology-intensive or software-intensive business domains. He has 80+ publications in scientific journals, books, conference proceedings, and other publications. He is actively working in several ICT-focused research consortia leading business-related research streams.

Following the development history of previous mobile systems, in 10 years we will see the deployment of 5G networks. The industry has achieved the consensus to redesign the radio network architecture for 5G. The software-defined design has been identified as the important evolution path for 5G networks.

Service Provider Competition and Cooperation in Cloud-Based Software Defined Wireless Networks

Jiefei Ding, Rong Yu, Yan Zhang, Stein Gjessing, and Danny H. K. Tsang

ABSTRACT

Software-defined wireless networking (SDWN) is an emerging paradigm in the era of the Internet of Things. In cloud-based SDWNs, resource management is seperated from the geo-distributed cloud, forming a virtual network topology in the control plane. Thus, a centralized software program is able to control and program the behavior of the entire network. In this article, we focus on resource management in cloud-based SDWNs, and discuss the competition and cooperation between cloud service providers. We present a Nash bargaining game approach to process the resource trading activity among cloud service providers in cloud-based SDWNs. Utility functions have been specifically considered to incorporate operation cost and resource utilization. Illustrative results indicate that cooperation is able to generate more benefits than competition. Moreover, resource sharing among cloud service providers has great significance in efficiently utilizing limited resources and improving quality of service.

INTRODUCTION

Software-defined wireless networking (SDWN) is a promising programmatically controlled platform [1, 2]. Beyond the traditional cloud computing network infrastructures, SDWN makes a breakthrough in several aspects. The control logic (the control plane) is separated from the underlying routers and switches (the data plane). In addition, by breaking the network control problem into tractable tasks, SDWN makes control, data transmission, and network switches simple and manageable. SDWN also introduces new abstractions that facilitate software-defined networking (SDN) evolution and innovation. By employing the SDWN framework, we are able to generate and execute control programs conveniently.

The concept of SDN has developed rapidly from an academic exercise to commercial success. For example, Google has employed SDN to interconnect its data centers across the globe in order to improve operational efficiency and reduce costs [3]. VMware network virtualization platform (NSX) is another successful employment of SDN. Users can build their own data center networks with large transport capacity and high security [4]. The study in [5] extended the SDN framework into radio access networks for centralized control and global optimization. SDWN extends the traditional SDN architecture for OpenFlow into mobile cloud computing networks. Cloud-based SDWN combines the SDWN multi-layer framework with cloud computing. The fundamental idea is to utilize data centers for complex network management. Data centers are geographically distributed, and connected through Ethernet networks, LANs, and WANs. The capabilities of virtualization, traceable traffic flows, and centralized management make SDWN effective in optimizing the utilization of distributed cloud resources.

In cloud-based SDWNs, cloud service providers (CSPs) are the agencies of cloud resources, and they are responsible for providing the application services to users. Therefore, users are able to access cloud storage through wireless network access (e.g., wireless mesh networks) [6]. Applications that are not easily used on a resource-constrained mobile device due to computation resources shortage can be implemented in the cloud [7] As various applications have different resource requirements, efficient resource management and sharing schemes become very important in cloud computing [8]. Computation resources are virtualized and allowed to be shared among multiple CSPs. Traditional studies propose a joint radio and computation resources optimization scheme in a single data center to minimize the operation power [9]. The benefits from CSPs sharing resources and forming a resource pool are discussed in [10]. However, resource sharing should be studied and evaluated in a large area in the future cloud market. Based on the market mechanism, CSPs either compete with each other or form cooperation to achieve mutual benefits.

In this article, we propose cloud-based SDWNs by SDN architecture and cloud computing in wireless networks. We focus on resource allocation and sharing schemes for CSPs in SDWN. The cloud resources in data centers are geographically distributed over a WAN. Therefore, the resource configuration among CSPs are further classified into local resource configuration and

Jiefei Ding and Rong Yu are with Guangdong University of Technology. Rong Yu is the corresponding author for this article.

Yan Zhang is with Simula Research Laboratory and the University of Oslo.

Stein Gjessing is with the University of Oslo.

Danny H. K. Tsang is with Hong Kong University of Science and Technology.

0163-6804/15/$25.00 © 2015 IEEE

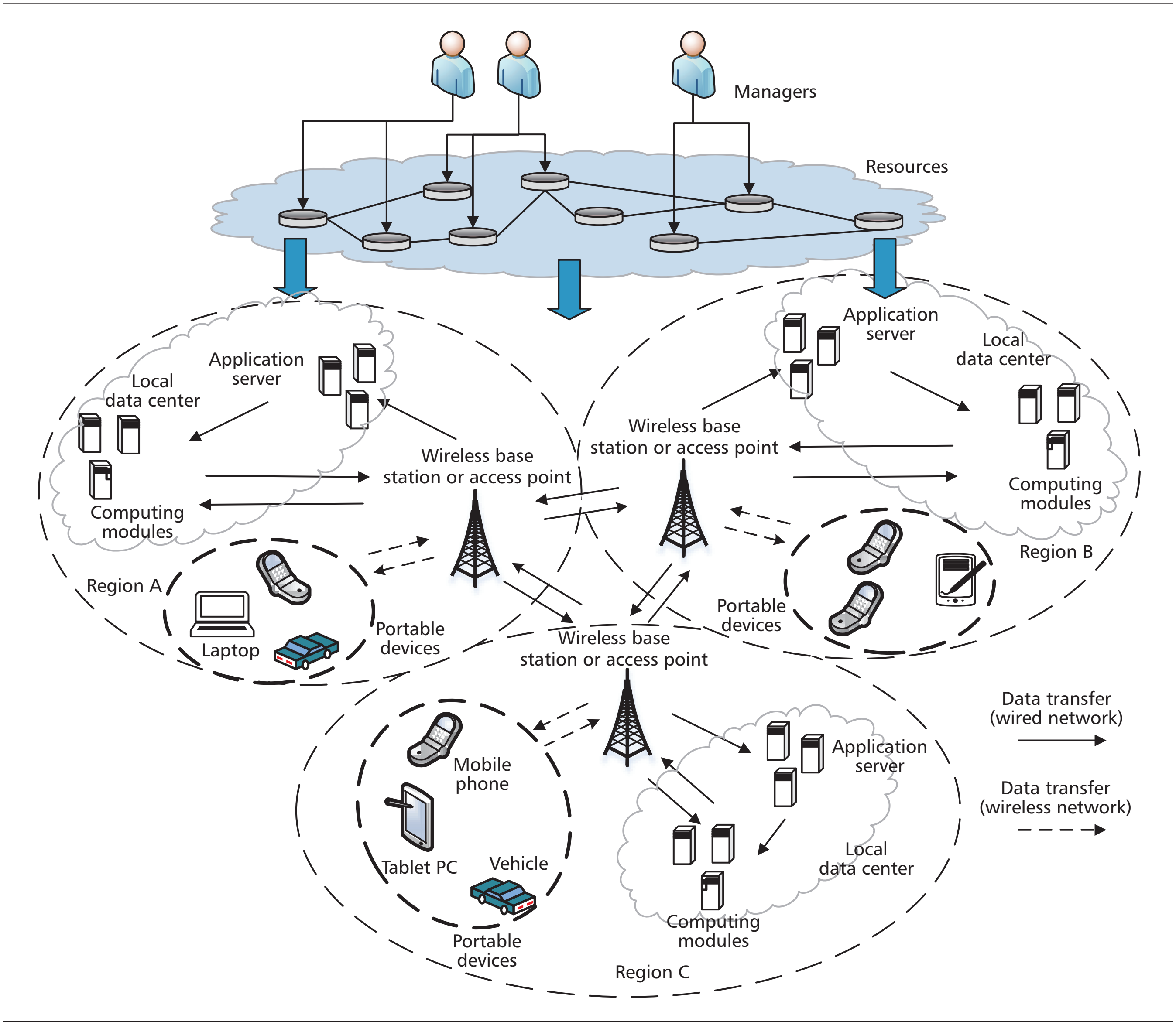


Figure 1. Our proposed cloud-based software-defined wireless networks.

remote resource configuration. In local resource configuration, CSPs can either act as competitor or cooperator. We formulate the resource configuration among CSPs in a bargaining game under the policies of competition and cooperation, respectively. Then we leverage price and user demand to stimulate resource cooperation. Eventually, we have two main observations. An optimal resource configuration among CSPs is able to improve resource utilization and revenue. In the method of resource configuration, the performance of the cooperation policy is better than that of the cooperation policy. Illustrative results demonstrate that resource cooperation among CSPs is able to significantly increase the performance of mobile cloud computing networks.

The rest of this article is organized as follows. We describe our proposed SDWN architecture. We explain the scenarios of cloud resource configuration. We discuss our competition policy and our cooperation policy for resource sharing among CSPs. Illustrative results are given. We then conclude the article.

Cloud-Based Software Defined Wireless Network

In this section, we first describe an application scenario of cloud-based software-defined wireless networks (SDWNs). Then we introduce the three-layer vertical integration architecture.

Why Cloud-Based SDWNs?

Figure 1 shows our proposed cloud-based SDWNs that cover a large area with a number of distributed data centers. Each data center is located in a region and mainly responsible for local services. The manager is in the virtualized plane and responsible for decision making. The main motivation of cloud-based SDWNs is to enhance computing capability and improve management efficiency.

Resources from the cloud can enhance the computing capability of devices. In traditional communication networks, a portable device runs applications on its own hardware and software.

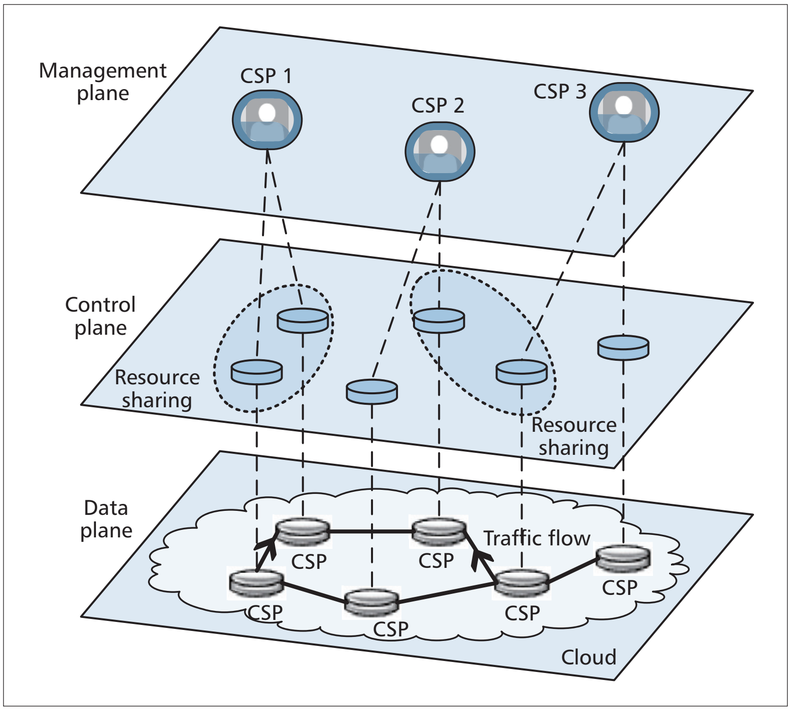


Figure 2. Three-layer vertical integration architecture for cloud-based SDWNs.

Therefore, limited battery supply and computing capability become the bottleneck of application services on a portable device. Cloud-based SDWNs can tremendously enhance the computing capability of portable devices. The cloud is a powerful virtualized supercomputer and can be accessed through the Internet [11]. By offloading application requirements to service providers in data centers, portable devices can execute a task on the associated cloud and obtain the computing results. Therefore, with alternative resources from cloud data centers, portable devices can greatly improve capacity.

SDN allows flexible administration of networked computing systems. A large number of users (e.g., laptops, mobile phones, and vehicles) employ cloud-based SDWNs for service, which requires cloud resources to be managed efficiently. When portable devices access a network, they require resources from data centers to run applications. As the manager, a CSP rents long-term resources from distributed data centers and resells them to users [12]. Geo-distributed networks can decrease transmission cost and delay. However, this increases the difficulty of global resource management unless there is centralized control. Therefore, we combine the SDN framework with cloud computing for optimal control. A control decision is required to separate from the forwarding facilities in the data plane. The resource configuration policy is first calculated in the top management layer; then the policy is executed by each CSP in the bottom layer.

Vertical Integration Architecture for Cloud-Based SDWNs

Cloud computing networks can be divided into three planes of functionality: the data plane, the control plane, and the management plane, as shown in Fig. 2. The data plane is the bottom layer and managed by data center devices. A CSP can rent resources from different data centers to build an independent resource network. The second layer is the control plane, which is responsible for data forwarding strategies and resource sharing strategies. This layer is related to the network topology and heavily influences the resource configuration result. Moreover, the second layer delivers information between the manager and each CSP in the bottom layer. Therefore, CSPs in the management plane can make a global decision on optimal resource allocation. In order to improve hierarchical organization and virtualization, the control function is moved from the data center to an external entity. The advantages mainly include:

- **Global optimum**: Working as managers in the top plane, CSPs can obtain global network information quickly (e.g., traffic flows, network utilization, and price information). Then they find a global optimal solution through competition or cooperation with other CSPs. Thus, the control policy can be consistent and optimal.
- **Optimality of resource configuration**: Innovative control strategies can easily be devised and conducted in this vertically integrated framework. The control layer is a virtualization platform, which specializes the network topology and collects CSPs' information. Therefore, we can optimize the traffic flows in networks for better communications and perform CSP cooperation for better resource sharing.
- **Flexibility of administration**: A well defined programming interface between data centers and the SDWN controller plays a very important role. This approach effectively presents control data from the control plane to the data plane, and makes it easy for CSPs to program [13].

In the top layer, CSPs are managers and take charge of resource allocation in the management plane. When a CSP receives application requests from users, it evaluates the resource utilization in distributed data centers and allocates adequate resources to users. Since application demands in each region are different, CSPs may have different resource utilization in different data centers. Even in the same data center, resource utilization may be different between different CSPs. Therefore, optimal resource allocation is significant to improve service quality and increase revenue.

Cloud Resource Configuration in SDWNs

Cloud Resource Configuration Scenarios

In cloud-based SDWNs, CSPs are allowed to share resources for capability enhancement. Figure 3 shows two resource configuration approaches. In this example, $CSP1$ rents resources from data center A and data center B. $CSP1$ in data center A has enough resources to run applications for users. Thus, after receiving an application request from $PD1$, $CSP1$ can directly allocate local resources to $PD1$. On the contrary, $CSP1$ in data center B will potentially refuse new applications when it has fewer resources. Thus, *CSP*1 rents resources from either remote or local CSPs.

Remote Resource Sharing: Users are allowed to access remote CSPs and run applications on remote devices [14]. In remote resource sharing, CSPs consider communications cost, latency constraints, and resource utilization. Let us take *PD*2 as an example. Its applications run on *SP*1 devices in data center *A*, since *SP*1 in data center *B* has fewer resources to support applications from *PD*2. The remote cooperator evaluates revenue before it agrees to establish remote resource sharing cooperation.

Local Resource Sharing: The local resource sharing strategy is complicated by the economic agreements between different CSPs. CSPs will be divided into two parts. CSPs that have sufficient resources put their resources into a resource pool and obtain profit from leasing their resources to other CSPs. On the other hand, CSPs with resources that are already highly utilized prefer to have more available resources through dealing in the local cloud resource market. Dealing is generally divided into two forms: competition (CSPs compete with each other and act selfish) and cooperation (CSPs cooperate with each other and maximize the global benefits). Local resource sharing is to run *PD*3 applications on *CSP*3 devices.

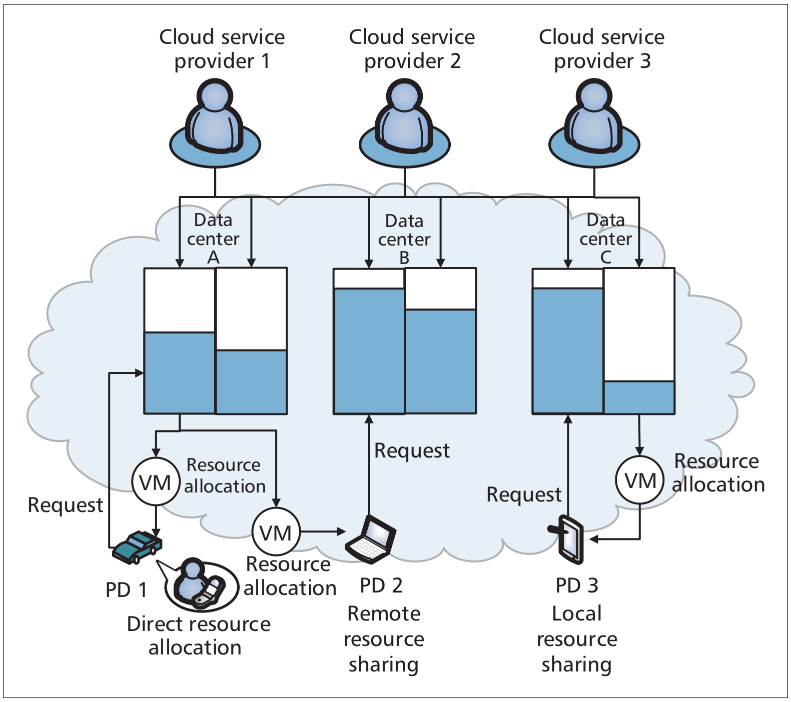


Figure 3. Cloud resource configuration in SDWNs.

Cloud Resource Configuration from a Game Perspective

We consider M geographically distributed clouds in different regions. The resources in each cloud can be rented by N different CSPs. One CSP can rent reserved resources from more than one region for a long term. CSP k in region l is denoted by S_k^l ($k = 1, 2, \ldots, N$, $l = 1, 2, \ldots, M$). The long-term rent resources are a fixed asset of S_k^l and denoted by the maximum capacity of each kind of resource: the CPU resource max C_k^l, the memory resource M_k^l, and the bandwidth resource B_k^l. Resource allocation schemes take the maximum capability into consideration. The resource requirement from user j can be denoted by $R_j = (k, l, C_j, M_j, B_j, T_j)$, which includes information such as service provider k in region l, the amount of CPU resource C_j, the amount of memory resource M_j, required bandwidth B_j, and the maximum latency T_j. Every application has a specific maximum latency T_j which should not be exceeded to provide a satisfactory quality of service (QoS) to a user.

A CSP could enhance its capabilities by allocating resources to remote CSPs or renting resources from the local cloud resource pool. Resource configuration in the local market follows the market rule (e.g., a buyer can bid for more resources at a higher price). We propose that all S_k^l as the player of a Nash bargaining game could obtain resources from the resource pool by competition policy. In order to obtain more resources, CSPs pay a high price to maximize their utility.

The payment from users to S_k^l is based on resource requests: the CPU resource C_k, the memory resource M_k, and the bandwidth resource B_k.

$$x_k^l = c\frac{C_k}{\max C_k^l} + a\frac{M_k}{\max M_k^l} + b\frac{B_k}{\max B_k^l}, \qquad (1)$$

where x_k^l is the weighted sum of resources. c, a, and b are fixed coefficients of three kinds of resources (CPU resource, memory resource, bandwidth resource) and satisfy the normalization condition $c + a + b = 1$. γ_k denotes the revenue, which includes the revenue of S_k^l from running applications, leasing resource to the local CSPs, and renting resources from the resource pool. Resources consumed by S_k^l are x_k^l. The user will pay for the cloud service on time.

The utility of each CSP involves two items: payment $\gamma_k x_k^l$ and penalty for overutilization $f(x_k^l)$. Overutilization indicates that little is left of that resource when the resource is already highly utilized (e.g., the cloud resource utilization is 99 percent), and little is left for newcomers. In this case, a CSP will refuse the application requests from newcomers until it has enough resources. Thus, the penalty for overutilization is related to increased utilization and decreased QoS. A monotonically increasing function can be employed to represent $f(x_k^l)$ according to the relationship between the penalty and cloud resource utilization.

Nash Bargaining Game among CSPs

In this section, we discuss the resource configuration when CSPs compete or cooperate with each other. CSPs in cloud-based SDWNs intend to maximize their own revenue while keeping high QoS. Revenue can be obtained from service provision. Thus, CSPs have higher resource utilization and obtain more revenue if they provide more services. However, CSPs with that resource already highly utilized will potentially obtain a lower QoS since fewer resources are available for new users. In this case, we believe that CSPs can extend their capability by resource sharing in

Application index	CPU resource	Memory resource	Bandwidth resource
1	3	3	2
2	2	2	2
3	2	1	1

Table 1. Resource requirements of different applications.

a short term. CSP buyers can rent resources from the CSP sellers. CSPs aim to maximize their utility, which is used to evaluate the performance of CSPs (e.g., revenue and QoS). Therefore, we formulate the resource sharing behavior as a Nash bargaining game. CSPs in the game will negotiate price for better utility.

Competition among Cloud Service Providers

In the cloud resource market, CSPs are divided into two sides: buyers and sellers. Sellers form a resource pool and sell resources with an agreed price. Buyers compete for limited resources. Sellers are the CSPs with low resource utilization who put unoccupied resources in the resource pool for short-term leasing. Buyers are the CSPs who can rent the resources from the resource pool. In the competition policy, CSPs are independent and compete with each other. n denotes the number of CSPs that rent resources from data center l. In the resource pool, seller S_k^l obtains the increased utility $\Delta(\Theta_k^l, x_k^l, \gamma_k)$ through leasing x_k^l resources. Increased utility can be calculated by the difference of utility function after and before leasing x_k^l resources. Θ_k^l is the original value of cloud resource utilization. If $\Delta(.) > 0$, S_k^l will obtain profit from leasing resources. The maximum utility can be obtained by

$$\frac{\partial \Delta(.)}{\partial x_k^l} = 0 \text{ when } \frac{\partial^2 \Delta(.)}{\partial x_k^{l^2}} < 0. \tag{2}$$

Therefore, S_k^l will rent out resources x_k^{l*} to maximize its utility $\Delta(.)$.

All buyers, such as S_k^l, compete for the limited resources. The increased utility of S_k^l can also be calculated by the difference of utility function after and before renting x_k^l resources from other CSPs. The increased utility is the objective function of each competitor. Thus, the objective function of S_k^l is

$$\begin{aligned} O(x_k^l) &= \Delta(\Theta_k^l, x_k^l, \gamma_k) \\ &= -\gamma_k x_k^l - (f(\Theta_k^l - x_k^l) - f(\Theta_k^l)). \end{aligned} \tag{3}$$

If $\Delta(.) > 0$, S_k^l can make profit from renting resources. Each CSP will adjust the payment γ_k for exchanging the required amount of resources. The payment from S_k^l is affected by two factors: the price given by other CSPs and the amount of resources from the resource pool. CSPs with higher prices are allocated more resources. First, S_k^l has a share of resources x_k^l according to the relative price given and the price given by other competitors. The relative price can be obtained by Cardano's formula. Then the share of resources x_k^l is derived by maximizing increased utility

$$\frac{\partial O(.)}{\partial x_k^l} = 0.$$

Therefore, the competition result x_k^{l*} can be derived based on this two conditions. After several iterations, the global optimal result of all CSPs can be stable and converge.

Cooperation among Cloud Service Providers

Cooperation among buyers is another market policy. In this policy, the buyers form a group and collaborate with each other. The resources from the resource pool are divided among group members. For example, CSPs with higher utility enjoy more resources. Therefore, the objective function is the total increased utility of all buyers. By calculating the optimal result of the objective function, we obtain the amount of resource x_k^{l*} that rents from resource pool at price γ_k^* given by all CSPs to maximize the global utility.

All CSPs share the same unit price in the cooperation group, while the CSP with already highly utilized resources obtains the larger shares of resources. n CSPs are in the group of buyers and need to rent resources from the resource pool. They bid for the amount of resources in the resource pool π with unit price γ_k. Thus, the objective function is the sum of increased utility of n CSPs.

$$\begin{aligned} O(x_k^l) &= \sum_{k=1}^{n} \Delta(\Theta_k^l, x_k^l, \gamma_k) \\ &= -\sum_{k=1}^{n} \gamma_k x_k^l - \sum_{k=1}^{n} (f(\Theta_k^l - x_k^l) - f(\Theta_k^l)), \end{aligned} \tag{3}$$

where Θ_k^l is original value of cloud resource utilization of a CSP in the resource pool. The deal price γ_k^* and the amount of resources x_k^{l*} are obtained when buyers and sellers are in the state of maximum utility. In the cooperation group, CSPs are renting resources by the same price γ_k^*. The amount of resources,

$$\pi^* = \sum_{k=1}^{n} x_k^{l*},$$

will be derived on the demand of each CSP. For example, if the demands of three CSPs are 3, 2, 1, and $\pi^* = 3$, the resource allocation result is 2, 1, and 0. The CSPs with higher demands are allocated more resources. Moreover, the demand from each CSP is according to the original utilization Θ_k^l.

Competition or Cooperation: Which is Better?

In this section, we will show illustrative examples to demonstrate the performance of competition and cooperation among service cloud providers. We consider six cloud service providers. The capability of each data center is the same, and shared by each CSP equally. We suppose the applications from users are random from three sets of applications (Table 1). The resources in this table are transformed into the units of each kind of resource (e.g., one unit of CPU means 4000 MIPS, one unit of memory means 4000 MB). We focus on the mobile applications in our daily lives. Let us take an application in a vehicle as an example. From the vehicle travel data of downtown San Francisco, California household transportation survey, the application stream during 11:00 to 20:00 is assumed to arrive at a Poisson process. The required resource stream of each service provider in data centers is random and follows a uniform distribution.

CSP sellers (e.g., *CSP*1, *CSP*2) divide revenue according to the resources that they lease to the resource pool. In the competition policy, CSP buyers such as *CSP*3, *CSP*4, *CSP*5, and *CSP*6 try to maximize Eq. 2 independently. The Nash equilibrium is obtained when no CSP buyers adjust their resource demand x_k^l. In the cooperation policy, CSP buyers form a group to bargain with CSP sellers. Therefore, the objective function is to maximize group utility (Eq. 3), which is the sum of Eq. 2. The Nash equilibrium is obtained when the negotiation price is unchanged. In the cooperation policy, the price given by *CSP*3, *CSP*4, *CSP*5, and *CSP*6 are 233.22, 288.29, 302.13, and 264.11, respectively. The amount of resources rented from the resource pool is 0.29. In the cooperation policy, the prices are 0, 280, 280, and 280, respectively for the amount of resources 0.33. To *CSP*3, the price 280 is beyond the price it expected, such as 233.22, so it will not rent any resources from the resource pool at the current price. On the contrary, *CSP*4, *CSP*5, and *CSP*6 can rent more resources at a lower price than competition policy. Thus, we can conclude that CSP buyers can obtain resources at a reasonable price in the cooperation policy.

Resource configuration brings many benefits to CSPs. Revenue can be increased significantly, as shown in Fig. 4. If CSPs form a resource sharing group, the buyers should pay the sellers for each application running, and the trade price can be decided through the bargaining game. Except for *CSP*3 and *CSP*6, the bars in light blue or blue are longer than dark blue. This result indicates that resource configuration can increase resource utilization and revenue. CSPs in the resource pool, such as *CSP*1 and *CSP*2, sell their resources for revenue. Other CSPs renting resources from the resource pool can expand capabilities in order to accept many more application requests. For example, the utilization of *CSP*4 and *CSP*5 is between 0.95 and 1. Resource sharing allows them to accept many more applications to improve revenue. However, for *CSP*3 and *CSP*6, utilization from 0.8 to 0.9 at best only slightly increases their utility. Resource sharing can improve utilization and QoS. When CSPs work together, *CSP*4 and *CSP*5 operate at the maximum utilization, as shown in Fig. 5. Resource sharing can enhance the capability of *CSP*4 and *CSP*5, and lease them more resources for more applications. The application drop rate of *CSP*4 decreases from 13.79 to 3.16 percent. Thus, the utilization of *CSP*4 decreases as QoS increases.

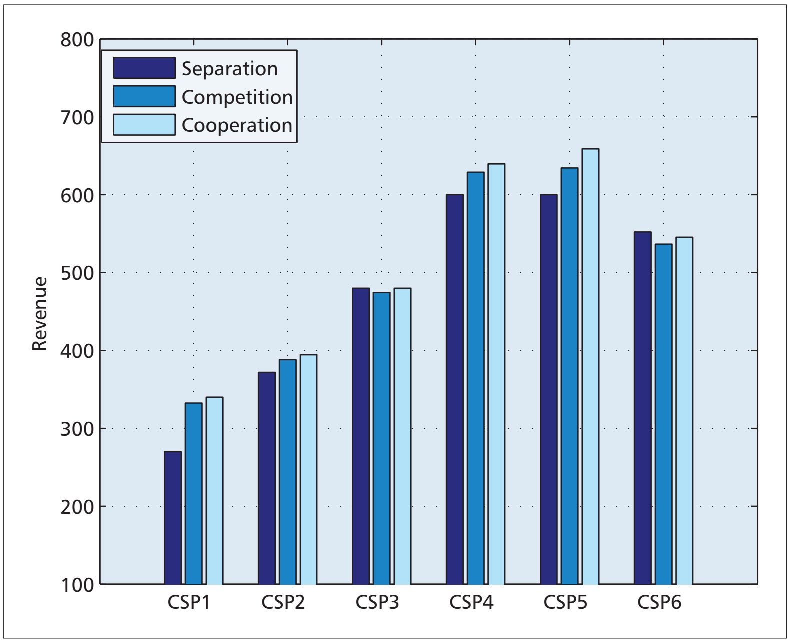


Figure 4. Revenues of six CSPs.

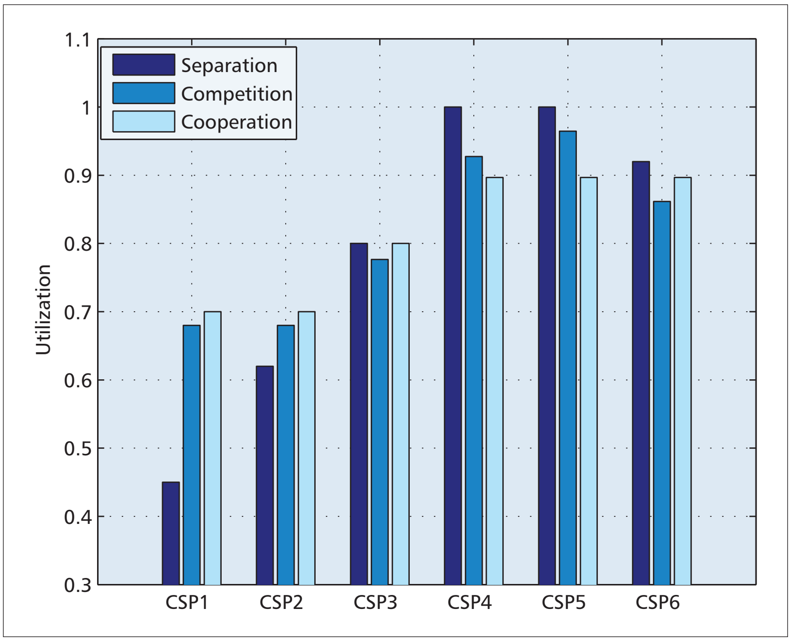


Figure 5. Utility in separation, competition, and cooperation.

On the contrary, in the resource pool, the resource utilization of *CSP*1 and *CSP*2 is improved as a result of leasing resources to other CSPs. The cooperation policy can generate more benefits than the competition. The result of competition shows that CSPs spend more money to rent fewer resources (e.g., *CSP*4 spends 322.00 for 0.0234 in competition and spends 288.29 for 0.0926 in cooperation). Besides, *CSP*1 and *CSP*2 can lease more resources and improve utilization in the cooperation policy. In the cooperation policy, all CSP buyers will buy resources at the same price. The CSPs with higher utilization can buy more resources (e.g., *CSP*4). On the contrary, CSPs with lower utilization can buy less resources (e.g., *CSP*6) or buy nothing (e.g., *CSP*3). As a consequence, the cooperation result is able to balance the resource utilization of all CSPs.

Conclusions

In this article, we propose a new software-defined wireless network framework in cloud computing environments. With the vertical separating framework, geo-distributed cloud resource can be supervised and controlled by the centralized software program. We discuss two resource configuration policies in our proposed cloud-based SDWN environment: the competition policy and the cooperation policy. Results show the cooperation policy can obtain more revenue for cloud service providers than the competition policy. Besides, the cooperation policy is able to achieve two goals: flexible resource management and demand-driven resource distribution. Future research can be conducted in several directions. For example, we can apply the proposed cloud-based SDWN framework in a specific distributed network (e.g., a vehicular network). Advanced mathematical tools can be employed to further improve network performance.

> *Future research can be conducted in several directions. For example, we can apply the proposed cloud-based SDWNs framework in a specific distributed network, e.g. a vehicular network. Advanced mathematical tools can be employed to further improve network performance.*

Acknowledgment

The work is supported in part by programs of NSFC under Grant nos. 61422201, 61370159, and U1201253, U1301255, Guangdong Province Natural Science Foundation under Grant no. S2011030002886, Special-Support Project of Guangdong Province under grant no. 2014TQ01X100, High Education Excellent Young Teacher Program of Guangdong Province under grant no. YQ2013057, and Science and Technology Program of Guangzhou under grant no. 2014J2200097 (Zhujiang New Star Program). This research is partially supported by the projects 240079/F20 funded by the Research Council of Norway, and the European Commission FP7 Project CROWN (grant no. PIRSES-GA-2013-627490.

References

[1] K. Hyojoon and N. Feamster, "Improving Network Management with Software Defined Networking," *IEEE Commun. Mag.*,, vol. 51, Feb. 2013, pp. 114–19.

[2] D. Kreutz *et al.*, "Software-Defined Networking: A Comprehensive Survey," *Proc. IEEE*, vol. 103, pp. 14–76, Jan 2015.

[3] S. Jain *et al.*, "B4: Experience with A Globally-Deployed Software Defined WAN," *ACM SIGCOMM Comp. Commun. Rev.*, vol. 43, 2013, pp. 3–14.

[4] "Vmware, Inc.," NSX Virtualization Platform 2013: https://www.vmware.com/products/nsx/.

[5] G. Aditya *et al.*, "Softran: Software Defined Radio Access Networks," 2013.

[6] G. Ping *et al.*, "A Variable Threshold-Value Authentication Architecture for Wireless Mesh Networks," *J. Internet Tech.*, vol. 15, no. 6, 2014, pp. 929–36.

[7] L. Hongwei *et al.*, "Enabling Efficient Multi-Keyword Ranked Search over Encrypted Mobile Cloud Data through Blind Storage," *IEEE Trans. Parallel and Distrib. Sys.*, vol. 3, no. 1, 2015, pp. 127–38.

[8] D. C. Erdil, "Autonomic Cloud Resource Sharing for Intercloud Federations," *Future Generation Comp. Sys.*, vol. 29, no. 7, 2013, pp. 1700–08.

[9] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Joint Allocation of Computation and Communication Resources in Multiuser Mobile Cloud Computing," *2013 IEEE 14th Wksp. Signal Processing Advances in Wireless Commun.*, June 2013, pp. 26–30.

[10] R. Kaewpuang *et al.*, "A Framework for Cooperative Resource Management in Mobile Cloud Computing," *IEEE JSAC*, vol. 31, Dec. 2013, pp. 2685–2700.

[11] X. Zhihua *et al.*, "A Secure and Dynamic Multi-Keyword Ranked Search Scheme over Encrypted Cloud Data," *IEEE Trans. Parallel and Distrib. Sys.*, Mar. 2015, pp. 1–1.

[12] A. Bo *et al.*, "Automated Negotiation with Decommitment for Dynamic Resource Allocation in Cloud Computing," *Proc. 9th Int'l. Conf. Autonomous Agents and Multiagent Sys.*, 2010, pp. 981–88.

[13] T. Ma *et al.*, "Social Network and Tag Sources Based Augmenting Collaborative Recommender System," *IEICE Trans. Info. and Sys.*, vol. 98, 2015, pp. 902–10.

[14] Z. Fu *et al.*, "Achieving Efficient Cloud Search Services: Multi-Keyword Ranked Search over Encrypted Cloud Data Supporting Parallel Computing," *IEICE Trans. Commun.*, vol. E98-B, no. 1, 2015, pp. 190–200.

Biographies

Jiefei Ding (jiefeiding@gmail.com) received her M.S. degree from Guangdong University of Technology, China. She spent five months studying at Hong Kong University of Science and Technology as a postgraduate visiting internship student in 2015. Her research interests include cloud computing resource management, software-defined networks, and demand response management in smart grid.

Rong Yu [M] (yurong@ieee.org) received his Ph.D. degree from Tsinghua University, China, in 2007. He is a full professor at Guangdong University of Technology (GDUT). His research interests mainly focus on wireless communications and networking, including cognitive radio, wireless sensor networks, and home networking. He is the co-inventor of over 10 patents, and author or co-author of over 70 international journal and conference papers. He is currently serving as the Deputy Secretary General of the Internet of Things (IoT) Industry Alliance, Guangdong, China, and Deputy Head of the IoT Engineering Center, Guangdong, China. He is a member of the Home Networking Standard Committee in China, where he leads the work on three standards.

Yan Zhang [SM] (yanzhang@ieee.org) received a Ph.D. degree from Nanyang Technological University, Singapore. From August 2006, he has been working with Simula Research Laboratory, Norway. He is currently head of the Department of Networks at Simula Research Laboratory, Norway, and an adjunct associate professor in the Department of Informatics, University of Oslo, Norway. He is an Associate Editor, Guest Editor, or on the Editorial Board of a number of well established scientific international journals, such as Wiley's *Wireless Communications and Mobile Computing* and *IEEE Transactions on Industrial Informatics*. He serves in chair positions for a number of conferences. His recent research interests include wireless networks, and reliable and secure cyber-physical systems. He has received 7 Best Paper Awards. He is a Senior Member of IEEE ComSoc and IEEE VT society.

Stein Gjessing [SM] (steing@ifi.uio.no) is an adjunct researcher at Simula Research Laboratory and a professor of computer science in the Department of Informatics, University of Oslo. He received his Ph.D. degree from the University of Oslo in 1985. He acted as head of the Department of Informatics for four years from 1987. From February 1996 to October 2001, he was chairman of the national research program Distributed IT-System founded by the Research Council of Norway. He has participated in three European funded projects: Macrame, Arches, and Ascissa. His current research interests are routing, transport protocols, and wireless networks, including cognitive radio and smart grid applications.

Danny H. K. Tsang [SM] (eetsang@ece.ust.hk) received his Ph.D. degree from the University of Pennsylvania in 1989. He has joined the Department of Electronic & Computer Engineering at the Hong Kong University of Science and Technology since summer 1992 and is now a professor in the department. He was a Guest Editor for the *IEEE Journal of Selected Areas in Communications*' Special Issue on Advances in P2P Streaming Systems, an Associate Editor for the *Journal of Optical Networking* published by the Optical Society of America, and a Guest Editor for the *IEEE Systems Journal*. He currently serves as a Technical Editor for *IEEE Communications Magazine*. He was nominated to become an IEEE Fellow in 2012. His current research interests include cloud computing, cognitive radio networks, coordinated electric vehicle charging, and smart grids.

CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE
WIRELESS TECHNOLOGIES FOR DEVELOPMENT (W4D)

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 societies 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 topics:

- "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 (ICN), 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 (such as 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 (FT) 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 in 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 4,500 words. All articles to be considered for publication must be submitted through the 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

An Intelligent SDN Framework for 5G Heterogeneous Networks

Songlin Sun, Liang Gong, Bo Rong, and Kejie Lu

Songlin Sun is with Beijing University of Posts and Telecommunications.

Liang Gong is with Heirot-Watt University.

Bo Rong is with the Communications Research Center Canada.

Kejie Lu is with Shanghai University of Electric Power and with the University of Puerto Rico at Mayagüez

This work was supported in part by the National Natural Science Foundation of China (Project 61471066), and by the Shanghai Oriental Scholar Program.

ABSTRACT

In fifth-generation (5G) mobile networks, a major challenge is to effectively improve system capacity and meet dynamic service demands. One promising technology to solve this problem is heterogeneous networks (HetNets), which involve a large number of densified low power nodes (LPNs). This article proposes a software defined network (SDN) based intelligent model that can efficiently manage the heterogeneous infrastructure and resources. In particular, we first review the latest SDN standards and discuss the possible extensions. We then discuss the advantages of SDN in meeting the dynamic nature of services and requirements in 5G HetNets. Finally, we develop a variety of schemes to improve traffic control, subscriber management, and resource allocation. Performance analysis shows that our proposed system is reliable, scalable, and implementable.

INTRODUCTION

In the past few years, the fourth-generation (4G) mobile communication system has been rapidly deployed and successfully operated in many countries. To provide enhanced connectivity for more diversified devices with higher mobility, the entire industry is focusing on the fifth-generation (5G) mobile system, and related technologies in 5G have become popular research topics.

A major challenge in the 5G era is to efficiently meet the increasing demand for network capacity while spectrum resource remains scarce. For instance, it is expected that future networks should be capable of handling the complex context of operations characterized by a tenfold increase in traffic [1] with guaranteed quality of service (QoS). In addition, energy efficiency has also been recognized as an urgent issue, since service providers are inceasingly concerned about operating expenses (OPEX) as well as the impact on global climate change and air pollution [2].

To address these challenges, one of the most promising technologies is the heterogeneous network (HetNet) architecture, which consists of a large number of densified low power nodes (LPNs). In HetNets, LPNs can provide high data rates to nearby mobile stations and can improve system capacity with frequency reuse. Moreover, LPNs can transmit signals with low power, leading to a significant reduction in energy consumption.

Despite the promising features of HetNets, with the drastic increase in the ultra-dense deployment of small cells in 5G networks, total energy consumption may easily exceed acceptable levels [3]. Moreover, fluctuating volumes of traffic also bring a considerable increase in energy consumption, which will also cause extra OPEX to service providers. In short, wireless network operators are now facing an unprecedent expansion in the number of users and the size of network.

In recent years various attempts have been made to address the issues of spectrum scarcity and energy consumption. For example, cognitive radio (CR) technology combined with software defined radio (SDR) is an effective solution to the problem of insufficient spectrum resources [5]. Adaptive resource allocation techniques, such as bit and power allocation, can become an efficient method to save energy [6]. Methods of subscriber churn prediction provide a viable method for service providers to balance their service load and provide network evolution insight [7]. Although the aforementioned solutions are feasible, they are mainly targeting a single part of the entire problem.

In this article we consider an intelligent management framework based on the concept of software defined networks (SDN), which features the decoupling of the control plane from the data plane [8]. In essence, SDN recognizes the network as an operating system and abstracts the applications from the hardware [8]. It enables the management-related functions to be implemented in a centralized manner, and thus network intelligence can be realized logically in a centralized SDN controller that maintains the entire system globally. More specifically, an SDN controller is programmed to learn about the physical topology of the network and the status of individual network elements through certain discovery mechanisms or appropriate databases, so as to orchestrate the entire network in a cost-efficient and energy-efficient manner.

In this article we propose the intelligent schemes of HetNet SDN to handle the dynamic environment of 5G. Specifically, our intelligent SDN framework will provide a viable approach to face the existing challenges in a unified manner. The rest of the work is organized as follows. We first review the existing SDN standards and pro-

0163-6804/15/$25.00 © 2015 IEEE

pose possible extensions. We then present our system model and develop a variety of adaptive schemes to make SDN work smartly with environmental change. We finally conclude the article.

SDN Standards and Possible Extensions

Current Standardization Progress

The concept of SDN was first introduced in the 1990s and became popular in the 21st century. The architecture of SDN was formally defined by the open networking forum (ONF) [9]. The ONF is also responsible for the maintenance of OpenFlow standards, technical specifications of the OpenFlow Swich, and the conformance testing of SDN enabled devices. The architecture of SDN consists of three layers: the application layer, the control layer, and the infrastructure layer. The function of the two lower layers are called OpenFlow controller and OpenFlow Switch, corresponding to the control and data planes of traditional IP/MPLS network switches and routers. With the OpenFlow standard, the OpenFlow controller instructs the OpenFlow Switch to define the standard functional messages such as packet-received, send-packet-out, modify-forwarding-table, get-stats, etc.

In addition to the ONF, the Internet Engineering Task Force (IETF), the International Telecommunications Union Telecommunication Standardization Sector (ITU-T), the European Telecommunications Standards Institute (ETSI), and the China Communications Standards Association (CCSA) have also started standardization work on SDN. IETF issued an RFC concerning the requirements and application issues relating to SDN from an operator's perspective [10], while ITU-T has not published a formal recommendation since the project began in 2012. Table 1 summarizes the formally published SDN standards.

Currently, SDN has found its best practices in campus networks and data centers, and has drawn increasing attention recently. SDN is regarded as a promising technology to solve current and emerging problems. However, necessary extensions are still required for future 5G HetNets.

Necessary Standard Extensions for Future 5G Networks

As mentioned previously, future 5G networks are characterized by their heterogeneity. At the same time, more effective technologies need to be employed for spectrum management, traffic control, resource allocation, density management, security, etc., in order to support the interconnection of more diversified user equipment and devices. The exploitation of SDN technology will largely reduce the complexity of 5G networks, lower the cost of network construction, facilitate future network evolution, and empower the network with intellectualization. However, current standardization work has not taken 5G networks into consideration. Thus, we aim to present the necessary extensions to current standards.

Table 2 summarizes and lists the suggested extensions to the standards of SDN. For instance, inter-ISP handoff and control plane security will be the key points for the development of SDN. Since the control plane is decoupled from the data plane, the switching of multi-homing user equipment between HetNets can be easily managed without sudden loss of connectivity or service interruption. The existing standardized solutions include the IEEE 802.11f, 802.11k, 802.11r for the homogeneous handover situations, and the IETF mobile IP (MIP) for handoff between the networks of different service providers. The implementation of SDN is straightforward, as OpenFlow controllers can communicate with one another to update and maintain a roaming IP-address table globally. Also, security is an important issue that must be emphasized in future SDN standards. Specifically, we must pay attention to some key aspects, such as the protection of subscribers' privacy, and robust defenses against cyber and malware attacks in different roaming scenarios.

Standardization organization	Main related standards and activities	Functionality
ONF	Software-Defined Networking: The New Norm for Networks (white paper) Interoperability Event Technical Paper v0.4/v1.0	Definition and interoperability
ITU	ITU-T Resolution 77	Standardization for SDN
IETF	IETF RFC 7149	Perspective
ETSI	NFV ISG	Use cases and framework and requirements
CCSA	TC6 WG1	Application scenarios and framework protocol

Table 1. Current main standards for SDN.

Extension for 4G/5G	Challenge
Mobility management	Seamless handoff between service providers.
Heterogeneous network support	Indiscriminate service regardless of location or type of network access.
Security	SDN control and privacy issues.

Table 2. Standard extensions for SDN.

System Model of Intelligent SDN

This section presents the system model that integrates intelligent SDN into the infrastructure of 5G network.

The 5G network will inherit the system architecture evolution (SAE) of 4G LTE, which is an evolved network framework that consists of a core network and a radio access network, as illustrated in Fig. 1. The core network (CN) of the system, known as the evolved packet core (EPC), is comprised of the serving gateway (S-GW), the

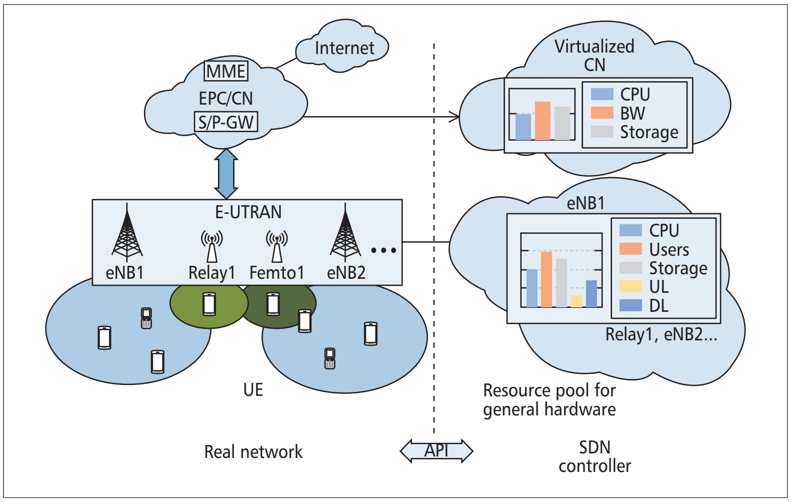


Figure 1. The proposed system model of intelligent SDN.

mobility management entity (MME), and the packet data network (PDN) gateway (P-GW). The radio access network (RAN), known as the evolved-universal terrestrial radio access network (E-UTRAN), is made up of macrocell BSs (also known as evolved NodeBs (eNBs)), pico, and femto (or home) relay nodes. For 5G core networks, SDN allows a high-level abstract, to which a set of underlying network resources are automatically and dynamically mapped. As for E-UTRAN, the eNBs and relay nodes can be implemented in a virtualized manner on general hardware coordinated and managed centrally by SDN controllers. The SDN controller, which is physically deployed on centralized servers, abstracts current resource usage and operates the network elements with intelligent strategy through standard APIs. In that case, different QoS requirements could be satisfied at the same time, and different kinds of service could be fulfilled with the most suitable resources.

Scenarios and Schemes for Network Intelligence

Resource Virtualization Vector

The architecture of intelligent SDN integrated with a 5G network consists of two layers: the SDN controller layer and the underlying physical network layer. The SDN controller instantiates the virtualized functional model CNs and the virtual base stations (VBSs) so that the signaling information can be exchanged locally. In this manner, we can achieve more adaptive and efficient management of the entire network. Specifically, the resources of the network can actually be categorized by computation (CPU utility), storage, bandwidth, etc. There are two vectors of a CN and a VBS representing the resource information respectively for convenience of the processing and modeling.

The CN supports two types of operation: the request for network association from a UE and the network's response. Thus, the resources in the CN mainly include three parameters, including the utility of the CPU, the size of the storage, the and data bandwidth. We can further define $\vec{V}_1 = (c_1, b, s_1)$, where c_1 and s_1 represent the percentage of CPU and random-access memory usage respectively, and b represents the available bandwidth. Similarly, $\vec{V}_2 = (c_2\ , f_{up}, f_{dn}, s_2, n_u, p)$ denotes the vector representing the resource parameters of a virtual VBS with four extra elements, f_{up}, f_{dn}, n_u and p, respectively, describing the uplink and downlink data flows, the number of users attached to the nodes, and the transmitted power.

Application Scenarios

In this section we introduce and discuss specific scenarios that demand dynamic and intelligent resource coordination over the entire network.

Scenario 1–Multimedia Services: Figure 2 shows the resource consumption of three types of multimedia services: live video conferencing, live TV streaming, and OTT with HTTP adaptive streaming (HAS). For video conferencing service, the network should provide a multi-point to multipoint connection among the attendees with low latency and delay variation to ensure a sufficient quality of experience (QoE). It indicates fast data forwarding and stable bandwidth, although data storage and processing may not be required as much. In other words, the demand for uplink and downlink data transmission should be much more urgent than for the CPU and storage. As for live TV streaming, which is a one-point to multi-point service, downlink bandwidth consumption is the dominant issue, while uplink only transmits the signaling data. Similar to video conferencing, live TV streaming only requires limited CPU and storage resources. However, it will become quite different with over-the-top (OTT) services. As the transcoder is involved in generating video slices of dynamic bit rates tailored for various screen sizes, all four aspects of resource requirement should be relatively more intense than the foregoing cases. The scenario of multimedia services presents an explicit example of how service deployment impacts the resource allocation requirement.

Scenario 2–Network Load Variation: Network load usually varies dynamically over time. The variation is usually regular, but sometimes sudden fluctuations may be observed.

Figure 3a illustrates the curves of daily network load fluctuation for business and residential areas. This partially reflects the resource utilization differences in time and locations. For a business area, the network load becomes relatively high in the daytime but low at night, while the situation for residential areas shows the opposite. The explanation is simply that most residents go to work in the daytime and come home in the evening. This scenario indicates the network must be aware of and adaptive to traffic fluctuations in both time and geographic dimensions.

When an emergency or a hotspot event happens, a sudden burst in network load will occur, as shown in Fig. 3b. Thus, the network should be intelligent enough to handle this situation. A viable method is to automatically allocate dedicated resources and to switch to distribute the hotspot contents by eMBMS technology. Thus, we can largely suppress the burst in resource consumption.

Moreover, the network load can vary from area to area. Figure 4a shows the radial traffic demand under random traffic distribution around a certain LPN. With traffic-aware schemes, we can divide the cell into two or more layers and assign each layer different power according to their traffic demands, so the system achieves lower power consumption as more layers become involved in the algorithm. Figure 4b presents the radial traffic demand under exponential traffic distribution, which leads to a similar result as that shown in Fig. 4a. We can conclude that traffic-aware schemes can better fit the network load variation and make full use of resources.

Different Intelligent Schemes

In this section we analyze in detail strategies of intelligent SDN concerning four aspects of 5G networks: traffic prediction, load balancing, user density prediction, and radio resource allocation.

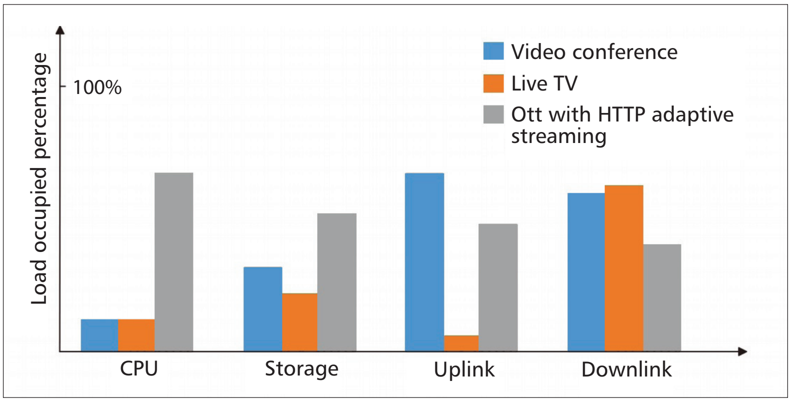


Figure 2. Scenario 1: different multimedia services.

Traffic Behavior Prediction: With the increasing traffic demand of users in the larger and more complex network, network management and monitoring seems to be a vital step of solving the problems. It can be carried out by traffic control, also known as traffic prediction, which predicts network traffic over a certain period of time. Accurate prediction of traffic requires the building of practical network traffic models according to different kinds of networks based on tracking fluctuations of data flows. Different from the conventional network models, the 5G network with intelligent SDN technology shows significant irregularity over time due to many non-linear factors in the network. Therefore, we can no longer describe the system only by combination and linear recursion techniques.

Considering the architecture proposed in this article, many methods [11] can be applied such as the real-time method, time series analysis, machine learning methods, etc. For the real-time method, the basic idea is that the system takes the latest obtained data profile as the one that will hold into the future. Under this circumstance, only in the case of slight variation from time to time can we achieve good performance. For the time series analysis, it takes the advantage of the fact that consecutive measurements can be represented by successive values in the data file. Nevertheless, the model becomes unreliable when trying to forecast further than the observation. Compared with the two approaches above, the methods referring to machine learning seem to be better solutions, and we next discuss two featured schemes: back-propagation (BP) and support vector machine (SVM).

BP is a neural network technology based on a multi-layer hierarchy, in which the upper neurons are full associated with lower neurons [12]. For traffic prediction, a hierarchy of three layers is sufficient, including an input layer, a middle layer, and an output layer. The learning samples are supplied from the input layer to the output layer through the middle layer neurons. The BP algorithm is implemented on the hidden middle layer by identifying the non-linear pattern, through which the traffic characteristics can be characterized and further predicted for the output layer. Applying this model could shorten the time of the training process with better accuracy of prediction.

SVM is in the category of learning algorithms, exploring better prediction techniques even under the traffic condition if intensive variability. The non-linear factor of the network will make most of the models useless, but as a relatively new machine learning technique, SVM can overcome the non-linear issue. The main idea of SVM for traffic prediction is to optimize the support vector that reflects the error of training data. This model applies three parameters, stream speed, volume, and density, to compose the training data and characterize the traffic stream. With training data, we can employ the support vector regression by using a ε-insensitivity loss function with linear kernel to approximately predict the traffic.

Load Balancing: For a cellular network, it is necessary to allocate time and frequency resources efficiently due to insufficient spectrum consumption. Normally, each macro cell is provisioned to possess an equal amount of resources. However, users in motion may spread in a dynamic and non-uniform manner, which often causes competition for resources in user populated cells while leaving a large amount of residual resources wasted in other cells. As a result, a homogeneous quality of service cannot be ensured over the entire network. Therefore, an efficient load balancing algorithm that makes full use of resources globally is vital for a cellular network. For 5G networks with intelligent SDN technology, the load balancing between macrocells and femtocells is a particularly crucial issue to deal with.

Hopefully, the classic single server processor sharing queue [13] can provide a feasible and dynamic solution to intercell load balancing. First, the optimal probabilistic dispatching policy is able to minimize the mean sojourn time; second, the dynamic load balancing polices are considered to acquire and adapt to the time-varying traffic. This two-stage approach determines whether to serve the flow locally in the microcell or to dispatch it to anther macrocell.

User Density Prediction: The density distribution of users in a cellular network provides valuable information that influences location based services and applications. With accurate user density, the network is able to allocate resources more wisely and purposefully. User density can be predicted with power allocation distribution, i.e. a high power sensed in one area usually indicates high user density in that area. Application of the mapping vector between the power in one

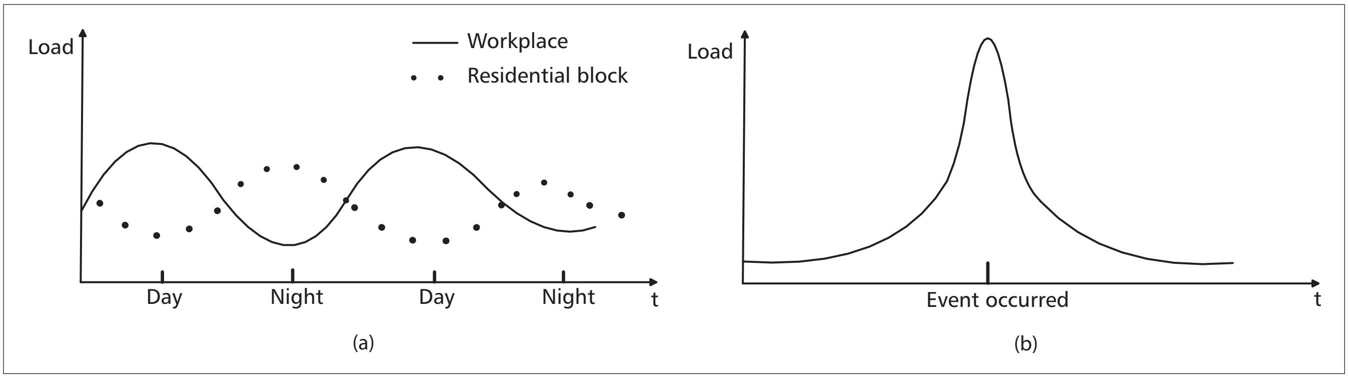


Figure 3. Scenario 2: Network load variation: a) daily network load variation; b) emergent and hotspot events.

area and the user density in that area can realize the density measurement without any other complex algorithm, but cannot ensure accuracy.

In practice, user density prediction usually works together with an enhanced version of the SFR scheme, also called SFFR, which is a traffic demand oriented scheme. This scheme not only pursues the overall throughput but also takes traffic demands in different area in consideration. The basic idea is to perform the power allocation constrained by the traffic demand. According to all the constrained conditions in the network, it can be further modeled as an optimized problem. Thus, we can realize the power allocation scheme with a power upper bound imposed in case of overlarge traffic demand. It is necessary to note that the scheme could lead to a serious interference with the adjacent cells when the distribution of the traffic demand becomes overwhelming.

Radio Resource Allocation: In HetNets, radio resource management (RRM) is a key to jointly managing resources from multiple access networks. The SDN controller works as a global optimizer to jointly manage resources among the networks to maximize network utility, and the UEs with multiple interfaces have the flexibility to select the cell to attach to so as to improve local capacity. Moreover, traffic that is tidal in nature, which means the traffic regularly varies with a large time-scale (i.e. hours), makes RRM and performance analysis more complex and difficult. The following two intelligent schemes are proposed to improve the efficiency of resource allocation.

Genetic Algorithm Based Joint Resource Allocation: The joint resource allocation of bandwidth and power transmission aims at maximizing the overall system capacity of HetNets. In effect, it is a multidimensional nonlinear constrained optimization problem with high complexity. The genetic algorithm (GA) provides an effective means to solve this nonlinear optimization problem. Actually, the GA features significant global search capacity and robustness in problem solving [14]. Aided by the GA, the joint resource allocation problem targeting the maximization of the system capacity can be achieved efficiently and intelligently.

•**Inter-Tier Spectrum Sharing Using Game Theory**: In 5G HetNets, the interference between macro-eNodeBs (MeNBs) and pico-eNodeB (PeNBs) needs to be coordinated to maximize the overall macro-cell throughput. This problem can be solved in the architecture of game theory. In paricular, by applying the Stackelberg game, the MeNB is the leader of the game and PeNBs the followers. The MeNB shall impose a price on the PeNBs for shared frequency bandwidth. The more bandwidth the PeNB demands, the higher expenditure PeNB has to pay. However, the PeNBs could serve some victim macro-users during ABS subframes in their expanded CRE range to reduce the PeNB expenditure. Both MeNBs and PeNBs intend to maximize their utilities. Through the interaction of the two kinds of nodes, frequency resources are allocated dynamically and efficiently with ICI well contained.

Performance Improvements

This subsection addresses the performance improvement achieved from our proposed intelligent SDN in 5G mobile network. These advantages result from four factors: reliability, scalability, complexity, and availability.

For reliability, the basic architecture illustrated in this article is based on models resulting from highly authoritative studies. Moreover, the maturity of SDN standardization also ensures reliability. The proposed architecture is also scalable, as an inherent characteristic of SDN. Specifically, the hardware involves general-purpose devices with functions and features defined by software, facilitating changes in configuration without replacing underlying physical devices.

The algorithms proposed in this article are studied in depth to guarantee both optimality and efficiency. Even for non-linear problems with unpredicted factors, we suggest low complexity solutions such as neural network technology and the GA. For availability, the proposed strategy is in general flexible and intelligent. Several smart schemes are involved to manage different categories of resources to enhance system performance in terms of traffic control, subscriber management, and resource allocation. In addition, the proposed schemes aim at distinct application scenarios.

Conclusions

This article presents an intelligent system that equips SDN with awareness and adaptiveness for dynamic networking in 5G HetNets. In our design, the SDN controller takes the responsibility of managing the infrastructure and optimizing resources. We first identify the key problems in 5G HetNets

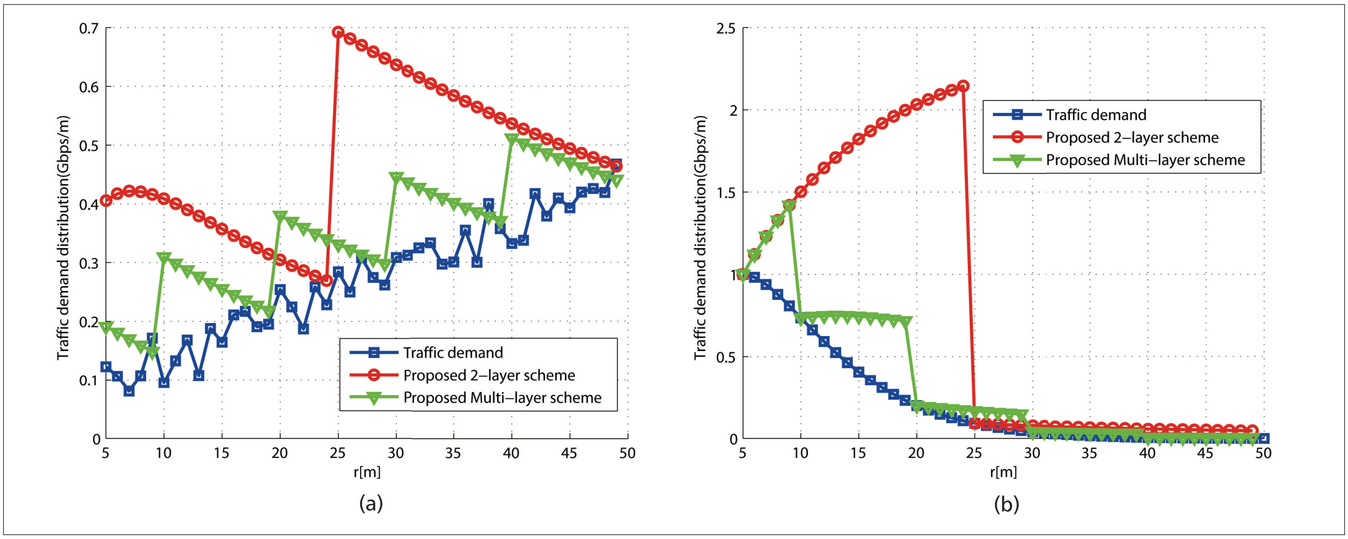


Figure 4. Actual traffic demand and the resulting available traffic from different power allocation schemes: a) random distribution; and b) exponential distribution.

as traffic control, load balancing, density prediction, and resource allocation. We then develop a number of smart schemes to overcome these challenges. Finally, we conduct performance analysis based on practical scenarios and demostrate that our schemes can deal with 5G network dynamics with low complexity and high flexibility.

References

[1] M. Palkovic *et al.*, "Future Software-Defined Radio Platforms and Mapping Flows," *IEEE Signal Proc. Mag.*, vol. 23, no. 4, Mar. 2010, pp. 22–33.

[2] I. Chih-Lin *et al.*, "Toward Green and Soft: A 5G Perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 66–73.

[3] R.L.G. Cavalcante *et al.*, "Toward Energy-Efficient 5G Wireless Communications Technologies: Tools for Decoupling the Scaling of Networks from the Growth of Operating Power," *IEEE Signal Proc. Mag.*, vol. 31, no. 6, Nov. 2014, pp. 24–34.

[4] H. Masutani and Yokosuka, "Requirements and Design of Flexible NFV Network Infrastructure Node Leveraging SDN/OpenFlow," *IEEE Optical Network Design and Modeling*, May 2014, pp. 258–63.

[5] A. Dalvi, P. K. Swamy, and B. B. Meshram, "Cognitive Radio: Emerging Trend of Next Generation Communication System," *Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE)*, 2011.

[6] M. Qian, J. Lu, and L. Wang, "Adaptive OFDM Mixed-service Resource Allocation for Broadband Power Line Communication," *Electricity Distribution (CICED)*, 2012, pp. 1–4.

[7] U. Yabas and H. C. Cankaya, "Churn Prediction in Subscriber Management for Mobile and Wireless Communications Services," *IEEE GLOBECOM Wksps. (GC Wksps.)*, 2013, pp. 991–95.

[8] B.A.A. Nunes *et al.*, "A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 3, 3rd Quarter 2014, pp. 1617–34.

[9] "Open Networking Foundation," Interoperability Event Technical Paper, vol. 4, Feb. 7, 2013.

[10] "Software-Defined Networking: A Perspective from within a Service Provider Environment," Internet Engineering Task Force, Mar. 2014.

[11] X. Yan *et al.*, "Research on Event Prediction in Time-series Data," *IEEE Machine Learning and Cybernetics*, vol. 5, Aug. 2004, pp. 2874–78.

[12] X. Pan, B. Lee, and C. Zhang, "A Comparison of Neural Network Backpropagation Algorithms for Electricity Load Forecasting," *IEEE Intelligent Energy Systems*, Nov. 2013, pp. 22–27.

[13] R. Q. Hu and Y. Qian, "*Heterogeneous Cellular Networks*," John Wiley and Sons, 2013.

[14] L. Tseng and S. Chen, "Two-Phase Genetic Local Search Algorithm for the Multimode Resource-Constrained Project Scheduling Problem," *IEEE Evolutionary Computation*, vol. 13, no. 4, Aug. 2009, pp. 848–57.

Biographies

SongLin Sun (slsun@bupt.edu.cn) received B.S. and M.S. degrees from Shandong University of Technology in 1997 and 2000, respectively, and his Ph.D. degree from Beijing University of Posts and Telecommunications in 2003. He is currently an associate professor at Beijing University of Posts and Telecommunications. His has long been engaged in the research and development and teaching work of wireless communications, multimedia communication, the Internet of Things, and embedded systems.

Liang Gong (lg15@hw.ac.uk) received his B.Sc. degree in communication engineering and the M.Sc. degree in communication and information system from Shandong University in 2003 and 2006, respectively, and his Ph.D. in communication and information system from Shanghai Jiao Tong University in 2011. From 2011 to 2015 he worked as senior engineer with the Academy of Broadcasting Planning (ABP), SAPPRFT, China. Since May 2015 he has been working as a postdoctoral research associate with Heriot-Watt University (UK) on the EPSRC project "TOward Ultimate Convergence of All Networks (TOUCAN)." His research interests include key technologies for 5G communication systems, network resource optimization, and network convergence.

Bo Rong (bo.rong@ieee.org) received his B.S. degree from Shandong University in 1993, his M.S. degree from Beijing University of Aeronautics and Astronautics in 1997, and his Ph.D. degree from Beijing University of Posts and Telecommunications in 2001. He is currently a research scientist with the Communications Research Centre Canada, Ottawa, Ontario. He is also an adjunct professor at Ecole de Technologie Superieure (ETS), Universite du Quebec, Canada. His research interests include modeling, simulation, and performance analysis of next-generation wireless networks.

Kejie Lu (Kejie.lu@upr.edu) received the B.Sc. and M.Sc. degrees in telecommunications engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 1994 and 1997, respectively. He received the Ph.D. degree in electrical engineering from the University of Texas at Dallas in 2003. In July 2005 he joined the Department of Electrical and Computer Engineering, University of Puerto Rico at Mayagüez, where he is currently an associate professor. Since January 2014 he has been an Oriental Scholar with the School of Computer Engineering, Shanghai University of Electric Power, Shanghai, China. His research interests include architecture and protocol design for computer and communication networks, performance analysis, network security, and wireless communications.

GREEN COMMUNICATIONS AND COMPUTING NETWORKS

Jinsong Wu

John Thompson

Honggang Zhang

Daniel C. Kilper

The concept of green information and communications technologies (ICT) is relevant to both environmental sustainability and ICT. Green ICT is an interdisciplinary field relevant to a number of areas and topics, such as information systems, computer science and technologies, communications and networking, power and energy systems, electronics, environmental and civil engineering, industrial engineering and project management, social sciences, and so on. The green ICT could be considered a coined-term from two highly overlapped terms, green communications [1] and green computing [2]. Basically, two mainstream ideas may be addressed in green ICT, i.e., greening ICT and ICT for green objectives. Even if the field of green ICT has been receiving more and more attentions in research communities as well as industrial, governmental, and international organizations [1–5], more and more technical investigations and applications are still expected due to the relevant long term concerns and challenges. In early 2015, the IEEE Technical Committee on Green Communications and Computing (TCGCC), IEEE Research Project on Vision for Green ICT Standardizations, and IEEE Technical SubCommittee on Big Data (TSCBD) jointly initialized the efforts of the first IEEE International Workshop on Green Standardization and Industry Issues for ICT and Relevant Technologies (GSICT), which will be held in conjunction with the 2015 IEEE Global Communications Conference (Globecom 2015), and the motivation of the efforts is to promote more relevant activities, especially standardization, towards green objectives in industry. Surely, the IEEE Series on Green Communications and Computing Networks can expect more submissions on the relevant industry issues and standardization activities in the future.

The fourth, November 2015, issue of IEEE Series on Green Communications and Computing Networks includes 6 articles addressing different topics relevant to green ICT.

The article "Green Energy Optimization in Energy Harvesting Wireless Sensor Networks," written .by Jianchao Zheng, Yueming Cai, *et al.*, investigated green energy optimization in wireless sensor network with energy harvesting capability, taking into account the dynamic and unknown environments on the target distribution and energy arrival.

The article "On Balancing Energy Efficiency for Network Operators and Mobile Users in Dynamic Planning," written by Muhammad Ismail, Mohamed Kashef, *et al.*, proposed an energy-aware dynamic planning scheme by taking both downlink and uplink energy consumptions into account, and elaborated the design issues for mobile terminal association and base station operation along with their trade-offs.

The article "Post-Peak ICT: Graceful Degradation for Communication Networks in an Energy Constrained Future," written by Sofie Lambert, Margot Deruyck, *et al.*, explored the interdependency of temporally varying power supply when renewable and intermittent energies provide major energy shares in the future ICT networks, and presented an introduction and discussion on the relevant problem for different ICT network sections and a case study for a wireless access network.

The article "A Survey on Green Scheduling Schemes for Homogeneous and Heterogeneous Cellular Networks," written by Ting Yang, Fabien Heliot, and Chuan Foh, provided a recent survey on green scheduling schemes for heterogeneous cellular networks and a performance analysis on some of the existing schemes.

The article "Assessing the Energy Consumption of Mobile Applications," written by Chien Chan, Wenwen Li, *et al.*, presented the energy consumption evaluation of smartphone applications in a real LTE (Long Term Evolution) network, provided a methodology for analyzing and processing large volumes of data from wireless networks, and studied the energy trade-offs between data and signaling traffic energy consumption, highlighting the effects of signaling on the network overall.

The article "Towards Green Data Center as an Interruptible Load for Grid Stabilization in Singapore," written by Wenfeng Xia, Yonggang Wen, *et al.*, addressed an interruptible load of data center caused from renewable energy and distributed generation, and proposed real-time power analytics framework embedded software as sensors (ESaS), which monitor resource usage, in order to stabilize power grids.

ACKNOWLEDGMENTS

We would like to acknowledge the great support from Osman S. Gebizlioglu, the current Editor-in-Chief of *IEEE Communications Magazine*, and S. Charis Scoggins, the Managing Editor of IEEE Communications Society Magazines, Jennifer Porcello, Production Specialist, and Joseph Milizzo, Assistant Publisher, and the other IEEE Communications Society publication staff. We also highlight the great support of this Green Series from the members of the IEEE Technical Committee on Green Communications and Computing (TCGCC) of the IEEE Communications Society.

REFERENCES

[1] J. Wu, S. Rangan, and H. Zhang, *Green Communications:Theoretical Fundamentals, Algorithms, and Applications*, CRC Press, September 2012.

[2] S. Murugesan and G. R. Gangadharan, *Harnessing Green IT: Principles and Practices,* Wiley, October 2012.

[3] R. Brown, C. Webber, and J. Koomey, "Status and Future Directions of the Energy Star Program," *Elsevier Energy*, vol. 27, no. 5, May 2002.

[4] "Recommendations for Measuring and Reporting Overall Data Center Efficiency Version 2 – Measuring PUE for Data Centers," the collective report of 7 × 24 Exchange, ASHRAE, The Green Grid, Silicon Valley Leadership Group, United States Department of Energy Save Energy Now Program, United States Environmental Protection Agency ENERGY STAR Program, United States Green Building Council, and Uptime Institute, May 2011.

[5] D. Kilper, "Greentouch Consortium: Building the Roadmap," *GreenTouch Consortium*, 2012.

BIOGRAPHIES

JINSONG WU [SM] (wujs@ieee.org) is associate professor in Department of Electrical Engineering, Universidad de Chile, Santiago, Chile. He is the founder and founding Chair of the IEEE Technical Committee on Green Communications and Computing (TCGCC). He is an Editor of IEEE Journal on Selected Areas in Communications (JSAC) Series on Green Communications and Networking. He was the leading Editor and co-author of the comprehensive book Green Communications: Theoretical Fundamentals, Algorithms, and Applications (CRC Press, 2012).

JOHN THOMPSON [SM] (john.thompson@ed.ac.uk) currently holds a personal chair in Signal Processing and Communications at the School of Engineering in the University of Edinburgh, UK. He was deputy academic coordinator for the recent Mobile Virtual Centre of Excellence Green Radio project and now leads the UK SERAN project which studies spectrum issues for 5G wireless. He also currently leads the European Marie Curie Training Network ADVANTAGE which trains 13 PhD students in the area of Smart Grid Technology. He is also a distinguished lecturer on green topics for ComSoc in 2014-2015.

HONGGANG ZHANG [SM] (honggangzhang@zju.edu.cn) is a Full Professor with the Zhejiang University, China. He was the International Chair Professor of Excellence for Université Européenne de Bretagne (UEB) and Supélec, France (2012-2014). He served as the Chair of the Technical Committee on Cognitive Networks (TCCN) in ComSoc during 2011-2012. He was the Lead Guest Editor of IEEE Communications Magazine Feature Topic Issues on "Green Communications". He has served as the General Co-Chair of 2010 IEEE GreenCom and the Co-Chair of IEEE Online GreenComm 2015. He is the book co-editor/co-author of "Green Communications: Theoretical Fundamentals, Algorithms and Applications" (CRC Press).

DANIEL C. KILPER [SM] (dkilper@optics.arizona.edu) is with University of Arizona. He served as the founding Technical Committee Chair of GreenTouch Consortium, and was the Bell Labs Liaison Executive for the Center for Energy Efficient Telecommunications at the University of Melbourne, Australia. While at Bell Labs, he received the President's Gold Medal Award in 2004 and was a member of the President's Advisory Council on Research. He has served as the General Co-Chair of IEEE Online Green Communications Conference 2014 and 2015.

Green Energy Optimization in Energy Harvesting Wireless Sensor Networks

Jianchao Zheng, Yueming Cai, Xuemin (Sherman) Shen, Zhongming Zheng, and Weiwei Yang

Jianchao Zheng, Yueming Cai, and Weiwei Yang are with PLA University of Science and Technology.

Xuemin (Sherman) Shen and Zhongming Zheng are with the University of Waterloo.

ABSTRACT

This article studies the sensor activation control for the optimization of green energy utilization in an EH-WSN, where both energy generation and target distribution exhibit temporal and spatial diversities. Decentralized operation is considered for the green energy optimization in the EH-WSN. The optimization is achieved in two dimensions: dynamic (activation) mode adaptation in the temporal dimension and energy balancing in the spatial dimension. Due to the interactions among autonomous distributed sensors, game theory is applied to the local information based decentralized optimization for the spatial energy balancing problem. In addition, reinforcement learning techniques are proposed to address the temporal mode adaptation in the dynamic and unknown environment. Simulation results are provided to demonstrate the effectiveness of the proposed approaches.

INTRODUCTION

Wireless sensor networks (WSNs) have profound significance toward environmental surveillance and monitoring by spreading throughout factories, forests, oceans, battlefields, and so on [1–3]. However, the limited network lifetime constrained by the battery capacity is a major deployment barrier for traditional WSNs. Recently, energy harvesting (EH) has emerged as a promising technology to extend the lifetime of communication networks by continuously harvesting green energy from environmental sources, such as the sun, wind, and vibrations [4, 5].

Due to the uncertain and dynamically changing environmental conditions, the intermittent and random nature is the most typical characteristics of the EH process. Thus, efficient energy management becomes critical to ensure continuous and reliable network operation [6, 7]. Most existing works assume that either the transmitter has non-causal information on the exact data/energy arrival instants and amounts, or the transmitter knows the statistics of underlying EH and data arrival processes [6]. Nonetheless, in many practical scenarios, the characteristics of EH and data arrival processes may change over time. Moreover, it may not be possible to have reliable statistical information about these processes before deploying the nodes. Hence, non-causal information about the data/energy arrival instants and amounts may be infeasible, so offline optimization frameworks may not be satisfactory in most practical scenarios. Besides, existing research on EH mainly focuses on a point-to-point communication system [6], while the network of multiple EH nodes is more challenging to study [8].

This article studies a general EH-WSN, where multiple energy harvesting sensors (EHSs) are deployed to monitor a target area, as depicted in Fig. 1. Energy harvested by sensors at different locations at different times are usually varying, which reflects both temporal and spatial diversities. Besides, due to the moving characteristics and random (usually not uniform) scattering of targets (e.g., in the battlefield environment), the target distribution also exhibits both temporal and spatial diversities. We focus on the dynamic and online sensor activation (activate/sleep) scheduling for green energy management.

According to characteristics of energy generation and target distribution, green energy optimization in the EH-WSN is a challenging problem that involves optimization in two dimensions: dynamic (activation) mode adaptation in the temporal dimension and energy balancing in the spatial dimension. Specifically, dynamic mode adaptation aims to optimize the green energy usage in multiple time slots to adapt to temporal dynamics of green energy generation and target mobility, while spatial energy balancing maximizes the utilization of green energy in each time slot by balancing the energy consumption among sensors to adapt to the spatial diversity of the energy generation and target distribution.

In general, the complexity of centralized schemes to achieve the optimal energy utilization increases significantly with the number of nodes in the network [8]. Moreover, the optimal solutions depend heavily on the available knowledge of the EH profiles across different sensors, which is difficult to obtain or even unattainable. Therefore, decentralized opti-

0163-6804/15/$25.00 © 2015 IEEE

mization based only on local information has drawn more attention [2, 3]. However, the existing works either consider a restrictive energy generation/event occurrence model, or simplify the study of the interactions among distributed sensors [9, 10].

Due to the complex interactions among individual sensors, we adopt game theory [11–13] to investigate the local-information-based decentralized green energy optimization in the spatial dimension. By carefully designing the utility function for each sensor, the network nodes can be made to exhibit desired behaviors while individual nodes simply perform local tasks. Moreover, to address the dynamic mode adaptation problem in temporal dimension, reinforcement learning techniques are proposed. After multiple iterative learning, self-regulating sensors can adapt their behaviors to the dynamic and unknown environment, and thus obtain a satisfactory solution that maximizes the green energy utilization.

The rest of this article is organized as follows. In the next section, we present an overview of the studied EH-WSN. Then we discuss key issues and technical challenges of green energy optimization. Following that, reinforcement learning techniques are incorporated into game theory to deal with the complex optimization for green energy utilization. Research directions are then discussed, followed by the conclusion.

Overview of the EH-WSN

In the EH-WSN, multiple EHSs are deployed for target monitoring, as shown in Fig. 1. Generally, both the energy generation and target distribution exhibits temporal and spatial diversities. The temporal diversity of target distribution indicates that the target distribution in the network varies in different time slots due to the moving characteristics of targets and probably the joining of new targets. Moreover, targets are randomly distributed in the area; thus, sensors at different locations may experience different target distribution intensities, which reflects the spatial diversity. Besides, green energy generation also possesses both temporal and spatial diversities. For example, solar energy generation depends on many factors such as temperature, sunlight intensity, the geographical location of the solar panel, and so on [7]. Therefore, energy generation by sensors at different locations is different. Moreover, the daily solar energy generation in a given area exhibits temporal dynamics that peak around noon and bottom out during the night.

In order to keep continuous and sustaining target monitoring, green energy utilization should be optimized by coping with the temporal and spatial diversities of green energy generation and target distribution. The characteristics of the energy arrival and target distribution in the current time slots as well as in future time slots need to be considered. The key problems for green energy optimization in the EH-WSN include medium access control (MAC) [1], power control [2], topology control, activation scheduling [3], and so on. We focus on the activation scheduling problem in this article.

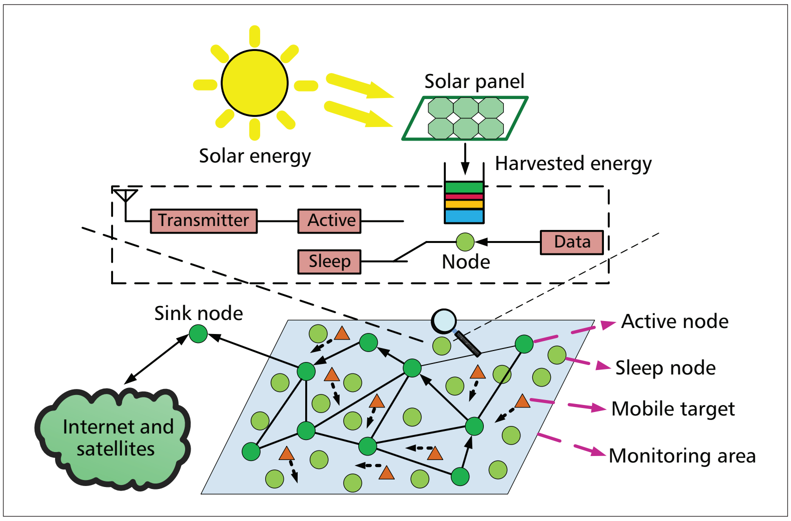


Figure 1. Energy harvesting wireless sensor network.

Green Energy Optimization in the EH-WSN

In this section, we discuss the motivation for activation scheduling based green energy optimization, and then introduce its key issues and technical challenges in the EH-WSN.

Motivation of Green Energy Optimization

Due to the seamless deployment of sensors, a target is often covered by multiple sensors. In the traditional battery-operated WSN, optimally turning some of these sensors to sleep mode will prolong the lifetime of the network while maintaining complete target coverage. In essence, the energy efficiency is improved only by optimizing the energy consumption. However, to improve the energy efficiency in the EH-WSN, not only energy consumption but also green EH should be taken into consideration. In other words, we should minimize the energy consumption and maximize the green energy collection at the same time.

Without loss of generality, we take an example with three sensors and three targets, as shown in Fig. 2a. The node sensing area is the disk centered at the sensor, with the radius equal to the sensing range. A sensor covers a target if the Euclidean distance between the sensor and the target is smaller than or equal to a predefined sensing range [14].

Assume each sensor has two units of energy in storage, and one unit can supply each sensor to be active for one time slot. Thus, if all sensors are active continuously, the network lifetime is two time slots. To prolong the network lifetime, we can permit each sensor to sleep alternately to save energy while ensuring all targets are monitored continuously by at least one sensor. In order to cover all the targets, at least two sensors need to keep active at any time. Therefore, we can divide the sensors into three cover sets: $C_1 = \{s_1, s_2\}$, $C_2 = \{s_2, s_3\}$, $C_3 = \{s_1, s_3\}$, and let each cover set be active for one time slot. This scheme will achieve a longer lifetime (i.e., $1 \times 3 = 3$ time slots) irrespective of the activation order of the cover sets.

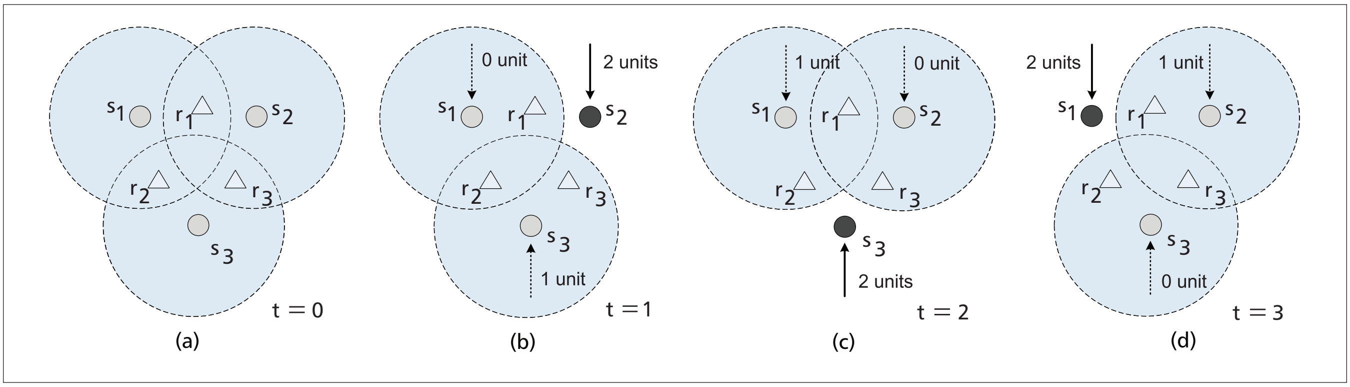


Figure 2. An example with three sensors $C = \{s_1, s_2, s_3\}$ and three targets $R = \{r_1, r_2, r_3\}$.

However, if sensors are equipped with the EH ability, they can get the recharging opportunity for energy supplementation.[1] Take 3 time slots for study, and assume the energy arrival processes for sensor 1, sensor 2, and sensor 3 are (0, 1, 2),[2] (2, 0, 1) and (1, 2, 0), respectively. By exhaustive searching, we can derive the optimal activation scheduling scheme as (C_3, C_1, C_2), as depicted in Figs. 2b–d. That is, C_3, C_1, and C_2 are active in the first, second, and third time slots, respectively. We can see that each sensor enters sleep state to collect energy when its energy arrival reaches the maximum. Although each sensor consumes two units of energy to activate for two time slots, it also harvests two units of energy from the environment. Therefore, through green energy optimization, not only is the energy consumption minimized, but the green energy collection is also maximized. Each sensor obtains sustaining supplementation for its energy consumption, which further enhances the energy efficiency and prolongs the lifetime of the network.

Key Issues and Technical Challenges

According to the above example, the optimal usage of green energy depends on characteristics of green energy generation and target distribution, both of which exhibit temporal and spatial diversity.[3] Therefore, the green energy optimization is a challenging problem that involves optimization in two dimensions: the temporal dimension and the spatial dimension.

Dynamic Mode Adaptation: Since mobile targets show temporal dynamics, sensors' energy demands change over time. Moreover, green energy supplements vary along the time horizon. Thus, in order to optimize their performance, sensors should adjust their activation modes adaptively, that is, determine when to activate and when to sleep. If a sensor stays active for more time at the current stage, it can provide better coverage, but more energy is utilized, and it may suffer from tracking discontinuity due to energy shortages in future stages. To solve the temporal mode adaptation problem, parameters such as the current energy arrival and consumption, and estimations of future energy arrival and consumption should be considered. However, characteristics of energy generation and target distribution change over time, and it is usually not possible to have statistical information before deploying the nodes. Thus, offline optimization frameworks cannot apply to EH-WSNs.

Spatial Energy Balancing: Due to the seamless deployment of sensors, a sensor can enter sleep mode by offloading its covering target to neighboring sensors. In this way, sensors' power consumption is adapted while ensuring complete target coverage. In order to maximize network sustainability, green energy utilization should be optimized by balancing the power consumption among sensors according to the availability of green energy. The power consumption of sensors is balanced by properly deciding the sleep time of each sensor. Sensors that have more harvested energy can keep working longer while letting sensors that are energy deficient enter sleep mode for energy conservation and new energy collection.

In practice, it is difficult to collect global information for centralized operation due to the time-varying characteristics of energy generation and target distribution. Alternatively, decentralized schemes for spatial energy balancing can be considered, which can achieve robust, scalable, and energy-efficient operation.

In summary, the main challenges of green energy optimization in the EH-WSN are listed below:

- Due to the complex coupling of the optimization in temporal and spatial dimensions, it is challenging to achieve optimal green energy utilization.
- The existing static and offline optimization schemes cannot adapt to the dynamics of energy generation and target distribution, which need dynamic and online optimization techniques.
- Energy arrival and target moving are stochastic, which means the information about future energy arrival and target distribution is nondeterministic and unknown.
- Self-organizing sensors perform decentralized information processing for proper operation only based on the local observed information.

[1] It is assumed that sensors cannot harvest green energy when they are active for target monitoring, since monitoring targets and collecting green energy simultaneously complicates the hardware design of the sensors.

[2] The ith element denotes the amount of energy arrival in the ith time slot.

[3] For simplicity, the temporal and spatial diversity of the target distribution are not studied in the above example.

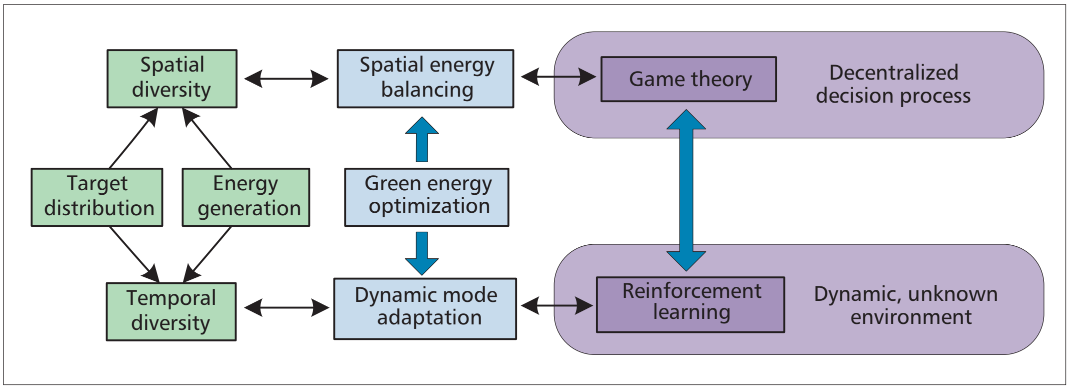


Figure 3. Schematic of green energy optimization in the EH-WSN.

Recently, there has been a great deal of interest in using game theory for analyzing communication networks. It is driven by the need for developing autonomous and flexible network structures as well as designing low-complexity distributed algorithms.

Spatio-Temporal Optimization in a Decentralized, Dynamic, and Unknown Environment

In this section, we incorporate reinforcement learning techniques into game theory to deal with the complex coupling of optimization in the temporal and spatial dimensions. Specifically, game theory is adopted to investigate local information based decentralized optimization for the spatial energy balancing problem, while reinforcement learning techniques are employed to address the temporal mode adaptation problem in a dynamic and unknown environment, as shown in Fig. 3.

Decentralized Optimization Using Game Theory

Game theory is a mathematical tool applied to model and analyze interactive decision making processes [11–13]. Recently, there has been a great deal of interest in using game theory for analyzing communication networks. It is driven by the need to develop autonomous and flexible network structures as well as design low-complexity distributed algorithms. Generally, a game model consists of three components: a set of players, a set of available actions for each player, and a set of utility functions mapping the action profiles into real numbers. By using game theory, the interactions among multiple interdependent decision makers can be well modeled and analyzed, and the outcome of complex interactions is predictable, and thus can be improved by properly designing the utility function and action update rule of each player.

In WSNs, the channel quality, required packet transmission energy, and acquired data value of each sensor all depend on the activity of other sensors. Due to the complex interactions among individual sensors, game theory becomes an attractive tool to investigate decentralized green energy optimization in the spatial dimension. By formulating a game, each sensor acts as an autonomous game player that observes and reacts to other sensors' behavior in an optimal fashion. This sets up a dynamic system wherein each sensor and its environment (i.e., other sensors) continuously self-adjust to adapt to each other, instead of treating other sensors as static entities [12]. This reactive behavior is also the reason game-theoretic optimization works better than a deterministic scheme for greedy optimization. In the literature, a lot of works use game theory to perform distributed optimization for battery-powered WSNs, but only a few study EH-WSNs.

Michelusi and Zorzi in [1] consider a multi-access game to design an optimal MAC protocol for maximizing the network utility in the EH-WSN, while [2] designs a power control game to maximize sensors' throughput. As for activation scheduling, Niyato *et al.* combine queuing theory and bargaining game to formulate a model for solar-powered WSNs[3], but the interactions among sensors are not analyzed.

For the game design, there are both similarities and differences between the battery-powered WSN and the EH-WSN. In both networks, each sensor's utility depends on its acquired data value and the associated energy cost, which are both affected by the activities of its neighboring sensors. Taking sensor *i*, for instance: if too many of sensor *i*'s neighbors activate simultaneously, excessive energy is consumed due to the spatio-temporal correlation of sensors' measurements, and the value of data collected by sensor *i* decreases. Moreover, the probability of successful transmission drops due to channel congestion, which means more energy for packet transmission is required to keep success rate fixed. Therefore, sensors are motivated to activate when the majority of neighbors are in sleep mode and/or its measurement is far from the local aggregated parameter [13]. However, the activation strategy of each sensor in the EH-WSN also depends on its time-varying energy state. If the residual energy is sufficient, it is natural for the sensor to take a positive activation strategy for monitoring targets, which takes the burden off the shoulders of other sensors that are short of energy. On the other hand, when the sensor has less energy in storage, it should be able to enter sleep mode for energy saving as well as new energy collection.

Besides, due to the temporal diversities of the green energy generation and target distribution in the EH-WSN, sensors' energy states, energy consumption, and acquired data values all vary dynamically. Thus, unlike the battery-powered WSN, mainly adopting static and deterministic game models, the dynamic and Bayesian games

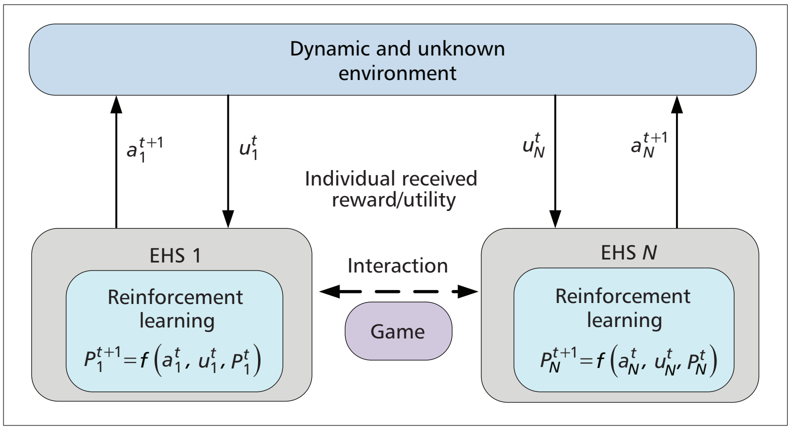

Dynamic and unknown environment
a_1^{t+1}
u_1^t
u_N^t
a_N^{t+1}
Individual received reward/utility
EHS 1
Interaction
EHS N
Reinforcement learning
$P_1^{t+1}=f\left(a_1^t, u_1^t, P_1^t\right)$
Game
Reinforcement learning
$P_N^{t+1}=f\left(a_N^t, u_N^t, P_N^t\right)$

Figure 4. Diagram of the reinforcement learning in the dynamic and unknown EH-WSN. $P_i^t = (P_i^t(0), P_i^t(1))$ is the player i's probability vector for activation strategy at time t, and $P_i^{t+1} = f(a_i^t, u_i^t, P_i^t)$ represents the probability updating rule that depends on the current strategy a_i^t and the received utility u_i^t.

are more suitable to the EH-WSN, which can be formally given by $\mathcal{G} = [\mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \mathbf{X}, \{\bar{u}_i\}_{i\in\mathcal{N}}]$. Here, $\mathcal{N} = \{1,2, \ldots, N\}$ is the set of players[4] (i.e., sensors), $\mathcal{A}_i = \{0, 1\}$ is the set of activation strategies (0 denotes sleep and 1 represents activate) for each player, $\mathbf{X}$ is a random variable characterizing the dynamic and unknown environment, and $\bar{u}_i = \mathbb{E}_{\mathbf{X}}[u_i(\mathbf{X})]$ is the mathematical expectation of the state-based utility function u_i, which is designed to trade off the energy cost of acquiring data against its value, based on the current energy state, that is,

$$u_i = \begin{cases} D_i^t - \gamma \dfrac{E_i^t}{\phi_i^t}, & \text{Activate,} \\ 0, & \text{Sleep,} \end{cases} \qquad (1)$$

where D_i^t denotes the value of data collected by sensor i at time t, E_i^t is the amount of energy consumption for activation, ϕ_i^t represents the current energy state of sensor i, and γ is a parameter that weighs the energy cost against its performance. Each player repeatedly plays a game where the actions are whether to activate or sleep and accordingly receives a reward/utility u_i. No pre-computed strategy is given, and players learn their activation strategies through repeated play, continuously adapting their strategies to maximize the expected utility $\bar{u}_i$.

Reinforcement Learning Techniques

Adapting to various unknown and time-varying characteristics in EH systems is a challenging research topic. The results available so far are few and limited. In [6, 10], the authors introduce the Markov decision problem (MDP) to address the dynamic process of data and energy arrival. However, the MDP requires the data/energy state transition to follow the Markov model and the state transition probability to be known. In this subsection, we propose two reinforcement learning techniques for a completely unknown dynamic environment. Reinforcement learning techniques are artificial intelligence tools that provide a system with the necessary information in order to plan its actions to maximize the reward it receives from the environment [15]. By adapting to the dynamic and stochastic characteristics of the wireless environment, reinforcement learning can result in a significant improvement in network performance.

No-Regret Reinforcement Learning: The no-regret procedure [11] is a regret-based reinforcement learning approach for optimization in a dynamic and unknown environment. In each time period, a player may either continue playing the same activation strategy as in the previous period or switch to another strategy, with probabilities that are proportional to the difference in the accumulated payoff caused by the strategy change, which we call the regret value. Taking time period t, for instance, each player first calculates the utility of the current strategy $a_i \in \mathcal{A}_i$ and the utility of choosing the other strategy $a_i' \in \mathcal{A}_i$, and then updates its regret value R_i^t by

$$R_i^t(a_i, a_i') = \frac{1}{t} \sum_{\tau \leqslant t:\, a_i^{(\tau)} = a_i} \left[u_i^\tau\left(a_i', a_{-i}^{(\tau)}\right) - u_i^\tau\left(a_i^{(\tau)}, a_{-i}^{(\tau)}\right) \right], \qquad (2)$$

where $a_{-i}^{(\tau)}$ is the joint strategy of all the players excluding i at time τ. By regret, the player compares its average utility to that of the other activation strategy, and then makes an intelligent decision on the optimal strategy for the next period according to the probability $P_i^{t+1}(a_i') = 1/\mu\,[R_i^t(a_i, a_i')]^+$, where $[R_i^t(a_i, a_i')]^+ = \max\{R_i^t(a_i, a_i'), 0\}$, and μ is a application-dependent normalization coefficient ensuring the probability in interval [0, 1]. It was proven in [11] that the average regret vanishes at the rate of $O(T^{-1/2})$, where T is the number of time periods. Having no regret means that no other strategies would significantly improve the player's utility.

However, existing no regret procedures mainly focus on the *full information model*, in which the utility of every action is observed at each time period, and all the history information needs to be exchanged among neighboring players by necessary communication [12]. This would create heavy signaling overhead for energy-hungry sensor networks. Therefore, we are more concerned with the *partial information model*, where at each time period only the utility/reward of the selected action is observed, as shown in Fig. 4. Each player may not even know the number of players participating in the game, let alone the actions the other players choose. Learning automata (LA) [15] are powerful learning tools that can be used toward this goal.

Reward-Based Learning Automata: At each period, a LA player that resides in a certain state chooses one of the available actions, performs it, and receives a new state from the environment as well as the environmental response. By repeating the above procedure, the LA player continuously interacts with the random operating environment in order to find the optimal strategy among the available actions. Specifically, the probability vector for activation strategy is updated according to the following rules:

[4] We will use sensor and game player interchangeably in this article.

$$\begin{cases} P_i^{t+1}(j) = P_i^t(j) + b\tilde{u}_i^t(1-P_i^t(j)), & \text{if } j = a_i^t, \\ P_i^{t+1}(j) = P_i^t(j) - b\tilde{u}_i^t P_i^t(j), & \text{otherwise}, \end{cases} \qquad (3)$$

where $a_i^t \in \mathcal{A}_i$ denotes the activation strategy played by player i at time t, $\tilde{u}_i^t$ is the normalized utility value for choosing a_i^t, $P_i^t(j)$ is the probability of choosing strategy $j \in \mathcal{A}_i$ for player i at time t, and b is the step size controlling the learning rate. The player operates entirely on the basis of their own strategies and the corresponding response from the environment, without any knowledge of other players in the network, and without prior knowledge of state transition probabilities or rewards. Therefore, it is particularly attractive to apply LA to address the dynamic and unknown characteristics in the EH-WSN.

Besides, although game theory copes with the spatial optimization while reinforcement learning deals with the temporal adaptation, the two techniques are executed at the same time to address the complex coupling of the optimization in the temporal and spatial domains, as shown in Fig. 4. The game is played repeatedly in the time horizon, and the player's reward in the dynamic learning process is essentially the game-theoretic utility. In this way, the proposed game-theoretic learning approach effectively solves the two-dimensional optimization of green energy utilization in the EH-WSN.

Performance Evaluation

Unlike the traditional battery-operated WSN, wherein both the target monitoring performance and network lifetime should be considered, the performance evaluation for an EH-WSN is mainly from the perspective of target monitoring, since the capability of harvesting green energy promises a potentially infinite lifetime [8]. In simulations, we consider an EH-WSN where multiple EHSs and 10 targets are randomly scattered in a 100 m^2 area. The system operates in a time-slotted fashion, where slot t is the time interval $[t\Delta_{TS}, t\Delta_{TS} + \Delta_{TS})$, $t \in \mathbb{Z}^+$, and $\Delta_{TS} = 10$ ms is the time slot duration. The target's location evolves according to a slow Markov process. Specifically, in every slot, each of the 10 targets randomly jumps to a new location with probability $\rho = 0.01$. Besides, we denote the energy harvested in slot t by B_i^t, which is modeled as a Bernoulli random process taking values in {0, 1}, and the probability of harvesting one energy quantum $\Pr(B_i^t = 1)$ varies in {0.1, 0.5}. In addition, the normalization parameter of the no regret procedure is set to be $\mu = 10$, and the step size for the reward-based LA is set to be $b = 0.1$. Similar to [1, 12], the network utility quantifying the performance of target monitoring is defined as the aggregate value of data collected by all sensors, that is, $U = \Sigma_i D_i$, where D_i denotes the value/importance of data reported by sensor i.

Figure 5 plots the convergence behavior of the two reinforcement learning algorithms when 300 EHSs are deployed. It can be seen that the no regret learning procedure converges faster to a better equilibrium, while the reward-based LA converges with more fluctuations to an inferior solution. However, the no regret learning procedure requires more information exchange for strategy updating to achieve the *no regret* play, while the LA does not need inter-node communication for information exchange. Besides, according to [11, 15], the computational complexity for both algorithms is $O(|\mathcal{A}_i|)$, where $|\mathcal{A}_i|$ is the cardinality of $\mathcal{A}_i$.

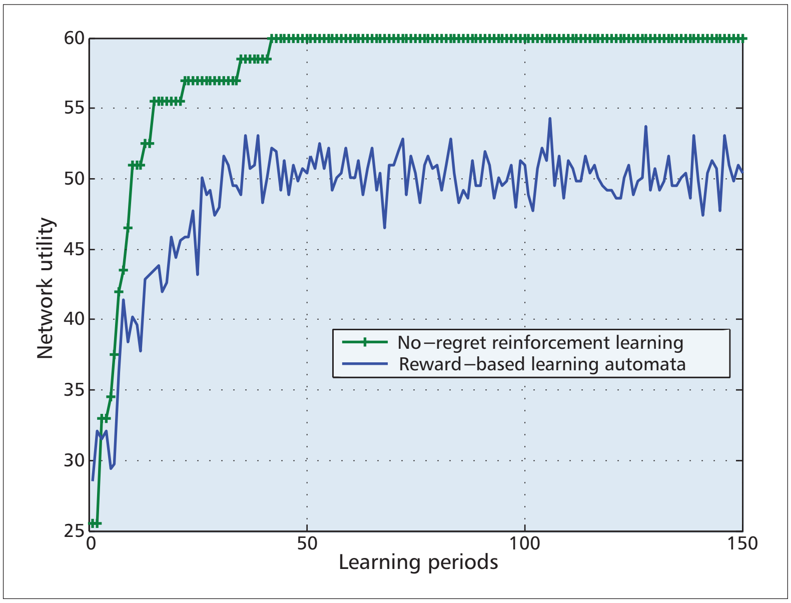


Figure 5. Convergence behavior of the reinforcement learning algorithms.

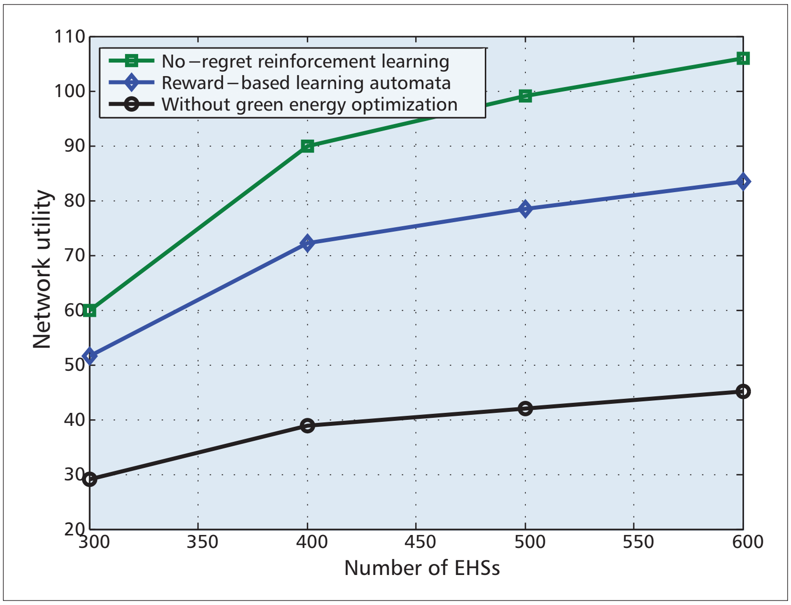


Figure 6. Utility performance comparison.

Figure 6 presents a performance comparison for different solutions in terms of the obtained network utility when the number of EHSs varies from 300 to 600. As the number of EHSs increases, the target monitoring performance obtained by all the solutions gets improved, but the proposed reinforcement learning algorithms are much better than the scheme in [12] without green energy optimization. The reinforcement learning algorithms achieve significant perfor-

Simulation results have demonstrated the effectiveness of the proposed approaches. In addition, several research directions have been identified and discussed, aiming to provide insights and guidelines for the researchers in this field.

mance improvement due to their advantages of dealing with the dynamic and stochastic EH environment. Moreover, since the regret/reward value for the algorithm implementation is based on the game-theoretic utility, the green energy utilization is also optimized/balanced among sensors in the spatial dimension.

Research Directions

To better utilize the green energy in EH-WSNs and further improve the network performance, the following potential research topics can be studied.

Designing More Efficient Game Models: The efficiency of the game-theoretic solution depends largely on the design of utility function for each player. This article only considers a simple and intuitive utility function to study the key problem. Further research can improve the game efficiency by carefully designing each player's utility function. Furthermore, different from the non-cooperative game models adopted in this article, cooperative game models can be applied to decrease competition among interactive players and improve energy cooperation among individual players.

Investigating the Trade-off between the Performance and the Cost of the Learning Techniques: The no regret algorithm achieves better performance at the expense of heavier information exchange overhead than the LA algorithm. As can be expected, the green energy optimization performance can be improved at the cost of convergence speed, computational complexity, signaling overhead, and so on. Therefore, it is important to investigate the trade-off between the performance and the cost as well as design proper algorithms based on specific applications.

Studying Energy Cooperation among Different Energy Sources: This article only studies solar-powered systems. However, due to the intermittent and random nature of the EH process, it is better to incorporated multiple energy sources into the network to increase the energy robustness. For example, at night, solar energy may not be available, but there can be wind for energy generation. Therefore, how to properly deploy sensors with different energy supplies and promote energy cooperation among them is a critical and interesting topic.

Studying Multihop Routing, Medium Access Control, Transmission Power Control, and Topology Control in the EH-WSN: Due to the dynamic EH process, these typical problems that are inherent in the traditional battery-operated WSN become quite different for the EH-WSN. However, existing works have not addressed these problems well; hence, they need further investigation.

Conclusion

In this article, we have investigated the green energy optimization in an EH-WSN, which involves two subproblems: the dynamic mode adaptation problem in the temporal dimension and the energy balancing problem in the spatial dimension. We have proposed game-theoretic methods with reinforcement learning techniques to deal with the complex coupling of the two-dimensional optimization in the EH-WSN. Simulation results have demonstrated the effectiveness of the proposed approaches. In addition, several research directions have been identified and discussed, aiming to provide insights and guidelines for researchers in this field.

Acknowledgment

This research work is supported by the Project of Natural Science Foundations of China under Grant No. 61301163, No. 61301162 and the Jiangsu Provincial Nature Science Foundation of China under Grant No. BK20130067.

References

[1] N. Michelusi and M. Zorzi, "Optimal Random Multiaccess in Energy Harvesting Wireless Sensor Networks," *Proc. IEEE ICC*, 2013.

[2] F. Tsuo *et al.*, "Energy-Aware Transmission Control for Wireless Sensor Networks Powered by Ambient Energy Harvesting: A Game-Theoretic Approach," *Proc. IEEE ICC*, 2011.

[3] D. Niyato et al., "Sleep and Wakeup Strategies in Solar-Powered Wireless Sensor/Mesh Networks: Performance Analysis and Optimization," *IEEE Trans. Mobile Computing*, vol. 6, no. 2, Feb. 2007, pp. 221–36.

[4] Z. Zheng et al., "Sustainable Communication and Networking in Two-tier Green Cellular Networks," *IEEE Wireless Commun*., Aug. 2014, pp. 47–53.

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

[6] P. Blasco *et al.*, "A learning Theoretic Approach to Energy Harvesting Communication System Optimization," *IEEE Trans. Wireless Commun*., vol. 12, no. 4, Apr. 2013, pp. 1872–82.

[7] T. Han and N. Ansari, "On Optimizing Green Energy Utilization for Cellular Networks with Hybrid Energy Supplies," *IEEE Trans. Wireless Commun*., vol. 12, no. 8, Aug. 2013, pp. 3872–82.

[8] D. Gündüz *et al.*, "Designing Intelligent Energy Harvesting Communications Systems," *IEEE Commun. Mag*., Jan. 2014, pp. 210–16.

[9] K. Kar *et al.*, "Dynamic Node Activation in Networks of Rechargeable Sensors," *IEEE/ACM Trans. Networking*, vol. 14, no. 1, Feb. 2006, pp. 15–26.

[10] Z. Ren *et al.*, "Dynamic Activation Policies for Event Capture in Rechargeable Sensor Network," *IEEE Trans. Parallel and Distrib. Sys.*, vol. 25, no. 12, Dec. 2014, pp. 3124–34.

[11] J. Zheng *et al.*, "Distributed Channel Selection for Interference Mitigation in Dynamic Environment: A Game-Theoretic Stochastic Learning Solution," *IEEE Trans. Vehic. Tech*., vol. 63, no. 9, Nov. 2014, pp. 4757–62.

[12] V. Krishnamurthy *et al.*, "Decentralized Adaptive Filtering Algorithms for Sensor Activation in an Unattended Ground Sensor Network," *IEEE Trans. Signal Process*., vol. 56, no. 12, Dec. 2008, pp. 6086–6101.

[13] O. Gharehshiran *et al.*, "Distributed Energy-Aware Diffusion Least Mean Squares: Game-Theoretic Learning," *IEEE J. Sel.Topics Signal Processing*, vol. 7, no. 5, Oct. 2013, pp. 821–36.

[14] M. Cardei *et al.*, "Energy-Efficient Target Coverage in Wireless Sensor Networks," *Proc. IEEE INFOCOM*, 2005.

[15] P. Nicopolitidis *et al.*, "Adaptive Wireless Networks Using Learning Automata," *IEEE Wireless Commun*., Apr. 2011, pp. 75–81.

Biography

Jianchao Zheng [S'12] (longxingren.zjc.s@163.com) received a B.S. degree in electronic engineering from the College of Communications Engineering, PLA University of Science and Technology, Nanjing, China, in 2010. He is currently pursuing a Ph.D. degree in communications and information systems in the College of Communications Engineering, PLA University of Science and Technology. His research interests focus on interference mitigation techniques, learning theory, game theory, and optimization techniques.

YUEMING CAI [M'05,SM'12] (caiym@vip.sina.com) received a B.S. degree in physics from Xiamen University, China, in 1982, and an M.S. degree in micro-electronics engineering and a Ph.D. degree in communications and information systems, both from Southeast University, Nanjing, China, in 1988 and 1996, respectively. His current research interests include cooperative communications, signal processing in communications, wireless sensor networks, and physical layer security.

XUEMIN (SHERMAN) SHEN [M'97, SM'02, F'09] (sshen@uwaterloo.ca) received a B.Sc.(1982) degree from Dalian Maritime University, China, and M.Sc. (1987) and Ph.D. (1990) degrees from Rutgers University, New Jersey, all in electrical engineering. He is a professor and University Research Chair, Department of Electrical and Computer Engineering, University of Waterloo, 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 in wireless communications and networks, control, and filtering. He is an elected member of the IEEE ComSoc Board of Governors and Chair of the Distinguished Lecturers Selection Committee. He served as TPC Chair/Co-Chair for IEEE INFOCOM '14, IEEE VTC-Fall '10, and IEEE GLOBECOM '07; Symposia Chair for IEEE ICC '10; Tutorial Chair for IEEE VTC-Spring '11 and IEEE ICC '08, General Co-Chair for Chinacom '07 and QShine '06; and as Chair of the IEEE Communications Society Technical Committees on Wireless Communications, and P2P Communications and Networking. He also serves/has served as Editor-in-Chief of *IEEE Network*, *Peer-to-Peer Networking and Application*, and *IET Communications*; a Founding Area Editor for *IEEE Transactions on Wireless Communications*; an Associate Editor for *IEEE Transactions on Vehicular Technology*, *Computer Networks*, *ACM/Wireless Networks*, and others; and Guest Editor for *IEEE JSAC*, *IEEE Wireless Communications*, *IEEE Communications Magazine*, *ACM Mobile Networks and Applications*, and so on. He received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004, 2007, and 2010 from the University of Waterloo; the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada; and the Distinguished Performance Award in 2002 and 2007 from the Faculty of Engineering, University of Waterloo. He is a registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, and a Distinguished Lecturer of the IEEE Vehicular Technology and Communications Societies.

ZHONGMING ZHENG (forezero@gmail.com) received B.Eng. (2007) and M.Sc. (2010) degrees from the City University of Hong Kong. Currently, he is pursuing his Ph.D degree in electrical and computer engineering at the University of Waterloo in the Broadband Communication Research Group. His research focuses on green wireless communication, smart grid, and wireless sensor networks.

WEIWEI YANG (wwyang1981@163.com) received B.S., M.S. and Ph.D. degrees from the Institute of Communications Engineering, PLA University of Science and Technology, in 2003, 2006, and 2011, respectively. He is assistant professor with the College of Communications Engineering, PLA University of Science and Technology. His research interests are mainly in OFDM systems, signal processing in communications, cooperative communications, and physical layer security.

On Balancing Energy Efficiency for Network Operators and Mobile Users in Dynamic Planning

Muhammad Ismail, Mohamed Kashef, Erchin Serpedin, and Khalid Qaraqe

ABSTRACT

The high energy consumption of network operators and mobile users has raised environmental, financial, and quality-of-experience concerns. These concerns have renewed the research efforts in developing green communication strategies for energy efficient wireless network operation. Network operators employ dynamic planning to save energy at low call traffic load by switching off some of their base stations (BSs), and mobile users are served by the remaining active BSs. The existing research investigates energy efficiency of dynamic planning approaches only from the network operator perspective. Dynamic planning, if not carefully designed, can lead to higher energy consumption for the mobile users in the uplink, which in turn degrades the uplink service quality due to mobile terminals' battery depletion. In this article we propose a dynamic planning framework with balanced energy efficiency that accounts for the energy consumption of the mobile users in the uplink as well as that of the network operators in the downlink. We discuss the associated challenges and implementation issues. A dynamic planning approach based on a multi-time scale decision process is proposed to achieve the balanced energy efficiency framework. Numerical results demonstrate the improved energy efficiency performance for the uplink mobile users as compared with the traditional dynamic planning approach.

INTRODUCTION

The great advancement in wireless communication services has resulted in high energy consumption for network operators and mobile users. Overall, there exist three million base stations (BSs) and approximately three billion mobile terminals (MTs) worldwide that consume 4.5 GW and 0.2 GW to 0.4 GW of power, respectively. The high energy consumption of network operators and mobile users has resulted in environmental, financial, and quality-of-experience (QoE) concerns.

From an environmental point of view, the telecommunication industry is responsible for about 2 percent of the total CO_2 emissions worldwide, and the percentage is expected to double by 2020 [1]. From a financial standpoint, it has been estimated that the energy bills of service providers cost about 18 to 32 percent of their operating expenditures [2]. From a QoE consideration, it has been estimated that almost 60 percent of mobile users suffer from a limited battery capacity, a problem that is further complicated by the slow advance in battery technology. The aforementioned concerns have motivated research in *green* communications for energy efficient wireless network operation.

In general, the green solutions that can be deployed by network operators can be classified based on the call traffic load condition. At a low call traffic load, network operators switch off some of their BSs to save energy and MTs are served by the remaining active BSs, which is referred to as *dynamic planning* [1]. In this context, one limitation associated with the existing research is that it investigates energy efficiency only from the network operator perspective. Dynamic planning approaches, if not carefully designed, can lead to higher energy consumption for the MTs in the uplink due to larger transmission distances. In such a case, dynamic planning would only shift the energy consumption burden from the BSs to the MTs, which results in battery drain for MTs at a faster rate. Consequently, this will degrade the service quality perceived by the mobile users because of, for example, lower throughput, higher call dropping rate, and so on. Hence, the future design of dynamic planning should capture and balance the trade-off in energy efficiency among network operators and mobile users.

In this article we shed the light on such a trade-off to motivate research in this direction. We first review the fundamentals of the traditional dynamic planning approaches. Then we present a dynamic planning framework with balanced energy efficiency and discuss the associated challenging design and implementation issues. Finally, we propose a dynamic planning approach

Muhammad Ismail, Mohamed Kashef, and Khalid Qaraqe are with Texas A&M University at Qatar.

Erchin Serpedin is with Texas A&M University, College Station.

0163-6804/15/$25.00 © 2015 IEEE

with balanced energy efficiency based on a multi-time scale decision process, and we present numerical results to illustrate the performance of the proposed approach.

Dynamic Planning in Green Networks

In this section we discuss the motivation for adopting dynamic planning for energy efficient network operation along with design fundamentals.

Motivations for Dynamic Planning

In cellular network planning, the cell size and capacity are traditionally designed to satisfy the peak call traffic load. However, it has been shown that the call traffic load exhibits temporal and spatial fluctuations [2]. Consequently, networks are over-provisioned at a low call traffic load, which leads to wasted energy. Thus, BS on-off switching is proposed to adapt network energy consumption according to the call traffic load condition, an approach that is referred to as dynamic planning. Hence, at a low call traffic load, lightly loaded BSs are switched off and MTs are connected to the remaining active BSs.

Two fundamental issues must be tackled while designing a dynamic planning mechanism, namely MT association and BS operation, as discussed in the following.

MT Association

This issue deals with associating each MT to a given BS for service. In the following, we discuss the design objectives of the MT association mechanism in the context of dynamic planning. Then we review the MT association mechanism decision criteria.

Design Objectives: In literature, the MT association serves two objectives for dynamic planning deployment. The first objective is related to concentrating the MTs' traffic in a few BSs to switch-off the remaining BSs. The second objective aims to balance the trade-off between network energy efficiency and flow level performance of the MTs (e.g. data rate, time delay, and so on.) [3]. The rationale behind such an objective is not to jeopardize the target quality-of-service (QoS) of MTs while reducing the network energy consumption.

Decision Criteria: The simplest decision criterion for the MT association problem is MT-BS distance based, where MTs with downlink traffic are associated with the nearest BSs [4]. As a result, low transmission power is consumed due to the short transmission distances. The network impact is another decision criterion introduced in [5] where MTs' traffic is concentrated in the BSs that lead to lower inter-cell interference. Furthermore, coverage hole avoidance is a decision criterion that aims to concentrate the call traffic load in a subset of BSs that can provide acceptable network coverage [6].

BS Operation

The BS operation specifies which BSs are switched off, when to wake up a given BS, and how to implement the switching decisions. The associated design issues are discussed in the following.

Prediction of Future Traffic Demands: The BS mode (on or off) lasts for a long duration (i.e. in hours) to avoid frequent BS on-off switching. Hence, the BS on-off switching decision should consider not only the current traffic load (through MT association) but also future demands so as to guarantee an acceptable QoS [1]. As a result, extra resources should be reserved at the active BSs to satisfy future traffic demands [7]. Information regarding future demands can be inferred from historical call traffic load pattern [1] or via prediction using an online stochastic game [8].

BS Wake-Up Design: When the call traffic load served by the active BSs increases beyond their capacity limitation, some of the switched off BSs have to be turned on. Hence, in dynamic planning, it is necessary to specify the wake-up instants for the inactive BSs. For instance, in [9] N-based and V-based wake-up schemes are proposed. Specifically, in the N-based scheme an inactive BS wakes up only when N MTs request service. On the other hand, in the V-based scheme, an inactive BS wakes up after a vacation time V. Similarly, different wake-up schemes are proposed for femto-cell BSs, which can be BS, MT, or network controlled [10].

Switching off Mode Entrance and Exit: An MT may not be able to complete its handoff procedure to an active BS and suffers from call dropping if its associated BS is switched off too quickly. This is due to the low received signal from the neighboring BSs and the limited signaling channel capacity. In [11] BS wilting is proposed for a smooth switch-off mode entrance, where the BS transmit power is progressively halved until the BS is turned off. Similarly, MTs in service can suffer from strong interference if a BS is switched on too quickly. A BS blossoming process is proposed in [11] for smooth switch-off mode exit, where the BS transmit power is progressively doubled until the BS is turned on.

An MT may not be able to complete its handoff procedure to an active BS and suffers from call dropping if its associated BS is switched off too fast. This is due to the low received signal from the neighboring BSs and the limited signaling channel capacity.

Dynamic Planning with Balanced Energy Efficiency

The main goal of dynamic planning is to reduce BS energy consumption while ensuring an acceptable downlink service quality for mobile users. For instance, in [3] the objective is to balance BS energy consumption performance with the MTs' downlink flow-level performance. However, no attention is paid to the relation between the mobile users' perceived service quality and their incurred uplink energy consumption. For instance, if an MT is consuming high energy in the uplink, it is expected to drain its battery at a fast rate, eventually leading to call dropping. Hence, dynamic planning approaches with balanced energy efficiency among network operators and mobile users should be investigated. In the following, the motivation for dynamic planning with balanced energy efficiency is presented. Then challenging design and implementation issues are discussed.

Motivation

When the main scope of dynamic planning is to enhance the energy efficiency of network operators, which is the case for the existing research, the BSs' switch-off decisions can result in energy

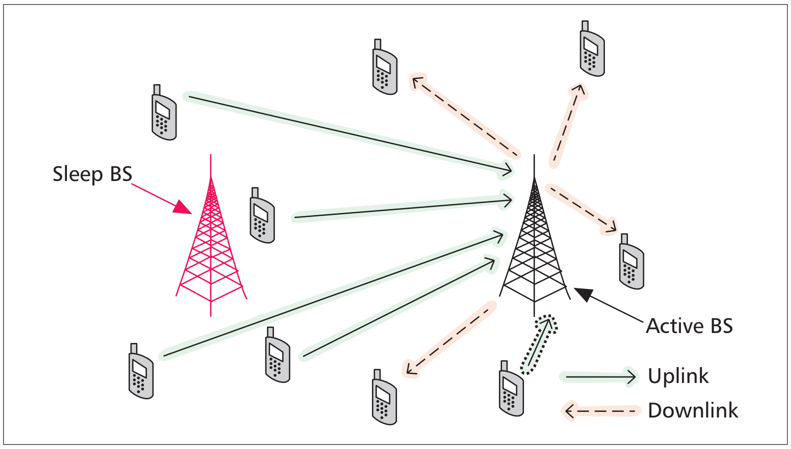


Figure 1. Dynamic planning with unbalanced energy saving. MTs with uplink traffic are associated with faraway BSs.

inefficient user association from the mobile users' standpoint. As shown in Fig. 1, accounting only for the downlink performance, MTs with uplink traffic can be associated with a faraway BS, due to a switched off nearby BS. Because of the long transmission distance, high energy consumption of MTs in the uplink is expected, which leads to energy depletion for MTs at a faster rate. Although energy consumption for MTs is not that much compared with BS energy consumption, a fast rate battery depletion for MTs still results in a high rate of dropped services in the uplink, which jeopardizes the mobile users' perceived service quality. Hence, the dynamic planning approach should be designed to capture and balance the trade-off in the resulting energy efficiency for network operators and mobile users.

Challenging Issues

In this subsection we discuss the challenging design and implementation issues toward developing a dynamic planning approach with balanced energy efficiency performance between network operators and mobile users.

The Coupling Between MT Association and BS Operation: One challenge with dynamic planning is that the switching decisions of BSs are coupled with MT associations. Specifically, when a BS is switched off, the MTs associated with it need to perform a handover process to another BS. Similarly, when a BS is turned-on, the nearby MTs can perform a handover process to this BS. Also, newly incoming MTs are associated with a subset of active BSs to obtain service. However, the BS operation (i.e. on-off switching) does not occur at the same rate as MT association. Hence, dynamic planning is a two time scale problem. At a high level, the BS operation occurs at a slow rate (with a scale of hours) that depends on the call traffic load density. At a low level, the MT association takes place at a faster rate (with a scale of minutes) based on user arrivals and departures. When only downlink traffic is considered, as in the existing research, the decisions at both levels are determined based only on BS energy consumption. With the coexistence of uplink and downlink traffic, the decisions at both levels are determined based on the expected energy consumption at the BSs and the MTs.

New Switch-off and Wake-up Decision Criteria: When uplink traffic is considered, the BS switch-off decision criteria should be revised. Specifically, the switch-off decision criteria should capture the impact of MTs' battery drain on uplink service degradation, for example, lower throughput, higher latency, higher call dropping rate, and so on. Hence, the uplink service degradation is due to two factors, namely unavailability of radio resources at the BSs (due to BS switch-off) and MTs' battery drain (due to communicating with faraway BSs). The BS switch-off decision metric should balance BS energy consumption with uplink service quality due to MTs' battery drain. Similarly, the existing mechanisms employ the call traffic load increase as a wake-up decision criterion for a switched off BS [5]. However, in the presence of uplink traffic, the wake-up decision criteria should include, besides the call traffic load measure, a measure of MTs' service degradation due to battery drain. As a result, if the MTs' service quality is degraded due to battery drain, a nearby inactive BS should be turned on to avoid MTs' battery depletion and hence dropping of uplink calls.

A Two Time Scale Dynamic Planning Approach with Balanced Energy Efficiency

In this section we propose a two time scale dynamic planning approach with balanced energy efficiency.

System Model

For illustration purposes, we consider a geographical region that is covered by two BSs from different networks, as shown in Fig. 2. Let s denote the BS index, with $s \in \{1, 2\}$. The BSs operate in separate frequency bands, hence no interference is considered. Interference management schemes (e.g. frequency reuse [12]) are employed for interference mitigation among BSs within each network. The distance between the two BSs is given by D and each BS has a height of H_s, as shown in Fig. 2. The network operators cooperate with each other for energy saving by alternately switching on and off their BSs according to the call traffic load condition [1]. The active BSs carry the call traffic load in the geographical region. Each BS can control its coverage area through antenna tilting [7]. For simplicity, two tilting angles are considered per BS, which corresponds to two cell coverage areas for each BS, namely A_{s,k_s}, where $k_s = 1$ and $k_s = 2$ denote the cell coverage area corresponding to the first and second tilting angles, respectively, as shown in Fig. 2.

In this article both downlink and uplink video traffic loads are considered. The uplink traffic includes mobile users who capture videos on their MTs and transmit them for online posting, while the downlink traffic includes mobile users performing video streaming. The user arrival rates, λ_{UL} and λ_{DL} for the uplink and downlink traffic loads, respectively, vary through the day [1]. The user service time has an average duration of μ_{UL} and μ_{DL} for the uplink and downlink traffic loads, respectively. The minimum required data rate is R_{UL} and R_{DL} for uplink and downlink users, respectively. A frequency division duplex (FDD) technique is considered at the BSs with available bandwidth of B_s^{UL} and B_s^{DL} at BS s for uplink and downlink traffic, respectively. Hence, BS s can support in total N_s and M_s users simultaneously on the uplink and downlink, respectively.

The channel average power gain, G_{UL} and G_{DL} in the uplink and downlink, respectively, is determined based on the path loss. To determine the average path loss, for BS s with coverage area $A_{s,1}$, the average transmission distance is approximated[1] by H_s, while for coverage area $A_{s,2}$, the transmission distance is approximated by

$$\sqrt{H_s^2 + D^2}\,.$$

The power consumption of a given BS s in the downlink is determined according to its mode of operation, i.e. active or inactive. For an inactive BS, the power consumption is $P_{\mathrm{L},s}$, which is much smaller than the active power consumption. For an active BS, the power consumption has two parts. The first is a fixed power component $P_{\mathrm{F},s}$, which captures the power consumed by the power supply, cooling, backhaul, and other circuits. The second component is proportional to the call traffic load, and is given by the multiplication of Δ_s (a slope of the load-dependent power consumption of BS s) by the BS transmission power. The BS average transmission power consumption, $P_{s,\mathrm{tx}}$, is a function of the average channel power gain (and hence the tilting angle index k_s) and the number of mobile users associated with BS s in the downlink $m_s (\le M_s)$. Specifically, to support m_s MTs by BS s each with minimum required data rate of R_{DL}, the BS downlink transmission capacity should at least be $m_s R_{\mathrm{DL}}$. Using the Shannon formula, the minimum average transmission power $P_{s,\mathrm{tx}}(k_s, m_s)$ can be computed. The total average power consumption of an BS s is denoted by $P_{s,\mathrm{DL}}(k_s, m_s)$, which has a maximum of $P_{s,\mathrm{mx}}$. The power consumption of a given MT in the uplink has two components. The first is a circuit power consumption part, while the second is the average transmission power component. For $n_s (\le N_s)$ MTs supported by BS s each with minimum required data rate of R_{UL}, the BS uplink transmission capacity should be $n_s R_{\mathrm{UL}}$. Using the Shannon formula, the minimum average transmission power consumption per MT associated with BS s, $P_{\mathrm{UL,tx}}(k_s, n_s)$, can be computed. The total average power consumption of an MT supported by BS s is denoted by $P_{\mathrm{UL},s}(k_s, n_s)$, which has a maximum of $P_{\mathrm{UL,mx}}$.

The number of MTs in the geographical region is denoted by $m = \Sigma_s m_s$ and $n = \Sigma_s n_s$ in the downlink and uplink, respectively. The spatial distributions of MTs in the geographical region follow probability mass functions (PMFs) of $\rho_{\mathrm{DL}}(A_{s,1})$ and $\rho_{\mathrm{UL}}(A_{s,1})$ for MTs with downlink and uplink traffic, respectively. Table 1 summarizes important mathematical symbols.

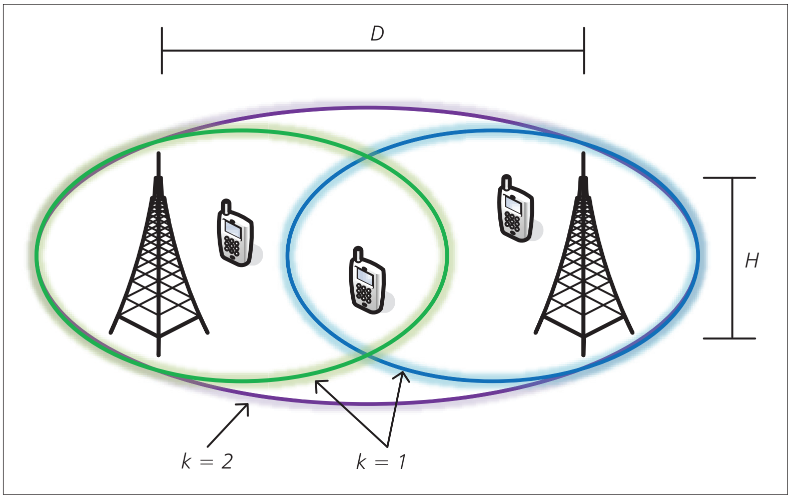


Figure 2. An example of dynamic planning cluster consisting of two BSs. For simplicity, two tilting angles are assumed per BS, leading to two coverage areas per BS.

The Two Time Scale Decision Problem Formulation

Time is divided into two scales, namely slow and fast scales, as shown in Fig. 3. At the slow scale, time is partitioned into a set of periods, $\mathcal{T} = \{1, 2, \dots, T\}$, with fixed duration τ, that covers the 24 hours of the day. The slow time scale captures the variation in the call traffic density. During one period $t \in \{\mathcal{T}\}$, the uplink and downlink arrival rates $\lambda_{\mathrm{UL}}(t)$ and $\lambda_{\mathrm{DL}}(t)$ are fixed, and vary from one period to another. At the fast scale, time is partitioned into a set of periods, $\mathcal{I} = \{1, 2, \dots, I\}$, of equal duration ω, $I = \lceil \tau/\omega \rceil$. The fast time scale captures the MT arrivals and departures. Hence, during one period $i \in \mathcal{I}$, the number of mobile users in the uplink and downlink in the geographical region, $n(i)$ and $m(i)$, are fixed, and may vary from one period to another.

The decision problem model has five elements: decision epochs, states, actions, transition probabilities, and cost. These are discussed for the two time scales in the following.

Slow Time Scale: At the beginning of every period t, the network operators make a decision regarding their BS operation mode and the tilting angle. The system state $\Upsilon(t) = (\gamma_{\mathrm{UL}}(t), \gamma_{\mathrm{DL}}(t))$ is given by the call traffic load density in the uplink and downlink, $\gamma_{\mathrm{UL}}(t) = \lambda_{\mathrm{UL}}(t)/\mu_{\mathrm{UL}}$ and $\gamma_{\mathrm{DL}}(t) = \lambda_{\mathrm{DL}}(t)/\mu_{\mathrm{DL}}$. The set of system states are pre-known from historical load patterns [1]. Given the uplink and downlink call traffic load densities, the actions specify the BS operation mode and tilting for the current period, that is, $W(t) = (k_1(t), k_2(t))$, where an inactive BS has $k_s(t) = 0$, otherwise $k_s(t) \in \{1, 2\}$. Given the system transmission capacity and users' minimum required data rate, the chosen action $W(t)$ should result in an acceptable service quality within t in the uplink and downlink, for example, $W(t)$ satisfies η_{UL} and η_{DL} target upper bounds on call blocking probabilities in the uplink and downlink, respectively. To provide radio coverage guarantees in the geographical region, both BSs are not allowed to be switched off simultaneously. Furthermore, if one BS is switched off, the

[1] More accurate expressions can be obtained using the distance distribution between BSs and mobile users [13].

other BS must provide radio coverage for the whole geographical region. For the system state $\Upsilon(t)$, the next state transition probability is deterministic and derived from the historical load patterns.

Fast Time Scale: At the beginning of every period i, the network operators decide the BSs' transmission power in the downlink and control the MTs' transmission power in the uplink. The system state $X(i, t)$ gives the number of uplink and downlink MTs in the geographical region, that is, $X(i, t) = (n(i, t), m(i, t))$. For a discrete-time decision problem, the fast time scale evolves as a discrete queuing system. Specifically, within one period i, the arrivals of MTs in the uplink and downlink are described by Bernoulli processes with probability $\lambda_{\mathrm{UL}}(t)\omega/\tau$ in the uplink and $\lambda_{\mathrm{DL}}(t)\omega/\tau$ in the downlink. The service processes follow geometric distributions with parameter $\mu_{\mathrm{UL}}\omega/\tau$ in the uplink and $\mu_{\mathrm{DL}}\omega/\tau$ in the downlink. Hence, the number of MTs being served in uplink and downlink are described by *Geo*/*Geo*/*N*/*N* and *Geo*/*Geo*/*M*/*M* queues, respectively [14]. The steady state probabilities of having $n(i, t)$ and $m(i, t)$ MTs in the uplink and downlink and the system transition probabilities can be found from the analysis of the discrete queues in [14]. The fast scale actions are to set the BSs' and MTs' transmission powers, that is, $J(i, t) = (P_{\mathrm{DL},s}(k_s(t), m_s(i, t)), P_{\mathrm{UL},s}(k_s(t), n_s(i, t)))$ for all s. Given the actions taken at period i, the BSs' expected energy consumption, $E_{\mathrm{DL}}(m(i, t))$, and the total expected energy consumption for the MTs with uplink traffic, $E_{\mathrm{UL}}(n(i, t))$, can be obtained. In order to account for the MT energy efficiency in the dynamic planning problem, the fast time scale cost function is modeled as a weighted function of the total downlink and uplink energy consumption, with a weighting factor β, that is,

$$C_{\mathrm{f}}(X(i, t), J(i, t)) = E_{\mathrm{DL}}(m(i, t)) \\ + \beta E_{\mathrm{UL}}(n(i, t)). \qquad (1)$$

The weighting factor β captures the impact of the MT energy consumption on the uplink service quality degradation (in terms of call dropping, throughput, etc.). Large values of β imply high impact of MT uplink energy consumption on uplink service quality degradation.

Optimal Solution: Let π_t denote the policy of the fast time scale decision problem in time slot t, that is, π_t is the set of actions taken for all $i \in \mathcal{I}$ at a given t, with a policy space of Π_t. In time slot t with system state $\Upsilon(t)$ and action $W(t)$, let $V_{\pi_t}(X_0(t))$ denote the total value function given initial system state $X_0(t)$ for the fast time scale decision problem. Hence, $V_{\pi_t}(X_0(t))$ is given by averaging $C_f(X(i, t), J(i, t))$ given some action $W(t)$ over system states and time. For the slow time scale decision problem, the immediate cost in time slot t, $C_s(\Upsilon(t), W(t), \pi_t)$, is determined by finding the expectation of $V_{\pi_t}(X_0(t))$ over the initial state $X_0(t)$. The slow time scale policy is denoted by $\pi = \{W(1), \ldots, W(T)\}$, with a policy space of Π. The dynamic planning approach with balanced energy efficiency follows the policies π and π_t for all $t \in \mathcal{T}$ that minimize the total uplink and downlink expected energy cost, that is,

$$\min_{\pi \in \Pi} \min_{\pi_1, \pi_2, \ldots \pi_t} \mathbb{E} \sum_{t=1}^{T} C_s\big(\Upsilon(t), W(t), \pi_t\big) \qquad (2)$$

where $\mathbb{E}$ denotes the expectation, which is taken over the states $\Upsilon(t)$. In order to solve Eq. 2, we first find the optimal fast time scale policy π_t that minimizes the expected total energy consumption $C_f(X(i, t), J(i, t))$ given the slow time scale action $W(t)$. Then we find the optimal action $W(t)$ that minimizes the expected total energy consumption $C_s(\Upsilon(t), W(t), \pi_t)$.

Numerical Results and Discussions

In this section we evaluate the performance of the proposed dynamic planning approach with balanced energy efficiency, by solving Eq. 2, as compared with a traditional dynamic planning approach that does not account for the energy consumption of the mobile users (i.e., with $\beta = 0$), which resembles existing mechanisms, for example, [1] and [7]. The system model is given in

Symbol	Definition
$B_s^{\mathrm{UL/DL}}$	Uplink/downlink available bandwidth at BS s
$E_{\mathrm{UL}}(n(i, t))$	Uplink total energy consumption at $n(i, t)$
$E_{\mathrm{DL}}(m(i, t))$	Downlink total energy consumption at $m(i, t)$
$J(i, t)$	Fast time scale action
k_s	Tilting angle index for BS s
m_s	Number of mobile users associated with BS s in the downlink
n_s	Number of mobile users associated with BS s in the uplink
$P_{s,\mathrm{DL}}(k_s, m_s)$	Total average power consumption of a BS s
$P_{\mathrm{UL},s}(k_s, n_s)$	Total average power consumption of an MT supported by BS s
$R_{\mathrm{UL/DL}}$	Uplink/downlink minimum required data rate
s	BS index
$W(t)$	Slow time scale action
$X(i, t)$	Fast time scale system state
β	Weighting factor
$\lambda_{\mathrm{UL/DL}}$	Uplink/downlink user arrival rate
$\mu_{\mathrm{UL/DL}}$	Uplink/downlink service time average duration
$\rho_{\mathrm{UL/DL}}(A_{s,1})$	PMF of spatial user distributions for MTs with uplink/downlik traffic
$\Upsilon(t)$	Slow time scale system state

Table 1. Summary of important symbols.

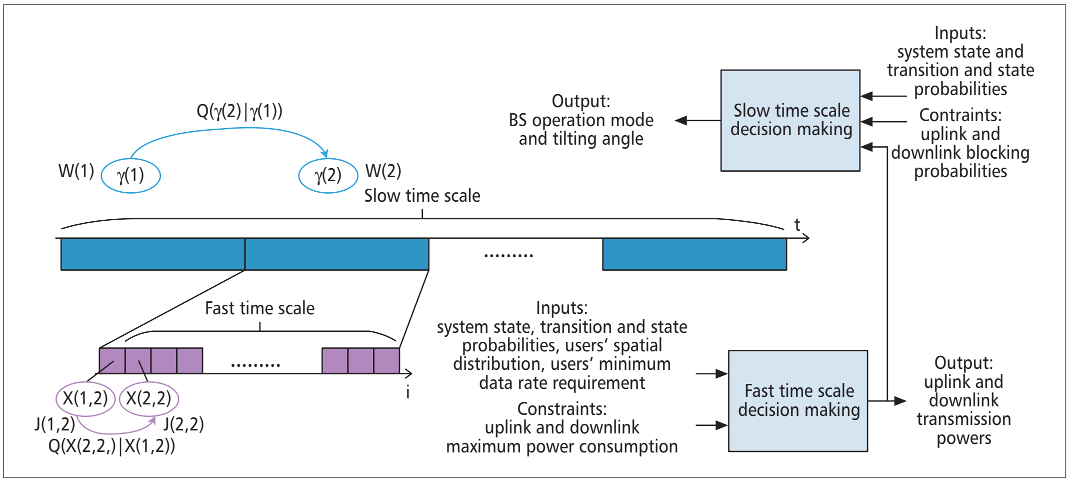


Figure 3. An illustration of the fast and slow time scales under consideration, the system states, actions, transition probabilities, and the decision making process.

Even if the unbalanced approach follows a random or round-robin BS switching off policy to decide which BS should be switched off, the unbalanced approach still will lead to higher expected uplink energy consumption compared with the balanced approach.

Fig. 2. The two BSs are identical and the system parameters are given by $H_s = 100$ meter, $D = 150$ meter, $\eta_{\mathrm{UL}} = \eta_{\mathrm{DL}} = 0.01$, $M_s = N_s = 7$, $R_{\mathrm{UL}} = R_{\mathrm{DL}} = 5$ Mb/s, $B_s^{\mathrm{UL}} = B_s^{\mathrm{DL}} = 5$ MHz, $P_{\mathrm{F},s} = 390$ watts, $\Delta_s = 4.7$, and $P_{\mathrm{L},s} = 75$ watts, $\tau = 1$ hour, $\omega = 5$ minutes, and $\beta = 50$. The fast time scale arrival rate in the downlink is 0.5, and the average service duration is 0.2 for both the uplink and downlink.

Figures 4a and 4b show the expected downlink energy consumption for both balanced and unbalanced dynamic planning. The unbalanced dynamic planning energy consumption performance does not vary with the weighting factor β since it does not account for the MTs' incurred energy consumption. It is affected only by the arrival rate. At low arrival rates [0.3, 0.6], only one BS is kept active to serve the MTs, while at higher arrival rates greater than 0.6, both BSs are switched on to satisfy the target service quality in terms of minimum required data rates (and hence upper bound on call blocking probabilities). On the other hand, the balanced dynamic planning energy consumption performance is affected by both β and the arrival rate. For low arrival rates and low β, a single BS is kept active to serve the MTs. As the arrival rate increases, a second BS is switched on to satisfy the users' target service quality (in terms of minimum required data rate and call blocking probabilities). In addition, large β values force the second BS activation to avoid uplink service degradation (e.g. higher call dropping rate, lower throughput, and so on) due to MTs' battery depletion. At low arrival rate values, the second BS activation is dominated by large β values, since the uplink service degradation due to MTs' battery depletion is more pronounced than users' call blocking due to limited radio resources, and the opposite is true at high arrival rate values. In Figures 4c and 4d, when a single BS is switched on at arrival rates [0.3, 0.6] and β = 5550 to 1400, respectively, the balanced approach decides which BS should be kept active based on the spatial distribution of the uplink users. Hence, for the balanced approach the second BS is kept active while the first BS is switched off. However, for the unbalanced approach, the expected energy consumption of the uplink users is not accounted for and hence the first BS is kept active while the second BS is switched off. Even if the unbalanced approach follows a random or round-robin BS switching off policy to decide which BS should be switched off, the unbalanced approach still will lead to higher expected uplink energy consumption compared with the balanced approach.

Figure 5 shows the expected uplink energy consumption versus the spatial distribution of uplink users near the proximity of the first BS. The arrival rate for the uplink users is fixed at 0.4. Due to the low arrival rate, only a single BS is kept active (Fig. 4a). As shown in Fig. 5, with more uplink users concentrated around the second BS ($\rho_{\mathrm{UL}}(A_{1,1}) \in [0.1, 0.5]$), the balanced dynamic planning approach keeps the second BS active and switches off the first BS, resulting in low expected energy consumption for the uplink users, unlike the unbalanced approach, which keeps the first BS active and switches off the second BS. As uplink users become more concentrated around the first BS ($\rho_{\mathrm{UL}}(A_{1,1}) > 0.5$), the balanced approach switches off the second BS and keeps the first BS active to keep the expected energy consumption of the uplink users as low as possible.

Conclusions and Future Research

In this article we have proposed a dynamic planning approach with balanced energy efficiency between network operators and mobile users. The proposed approach decouples the MT association and BS operation phases based on a two time scale decision problem. It introduces a new BS switch off metric based on the uplink mobile users' spatial distribution. Also, it accounts for the mobile users' uplink energy consumption in BS wake up, which avoids service quality degradation in the uplink. Our future research will focus on modeling the impact of uplink energy consumption on service quality degradation (e.g. in terms of higher call dropping, lower throughput, and so on) to provide a better representation of the weighting factor β. In addition,

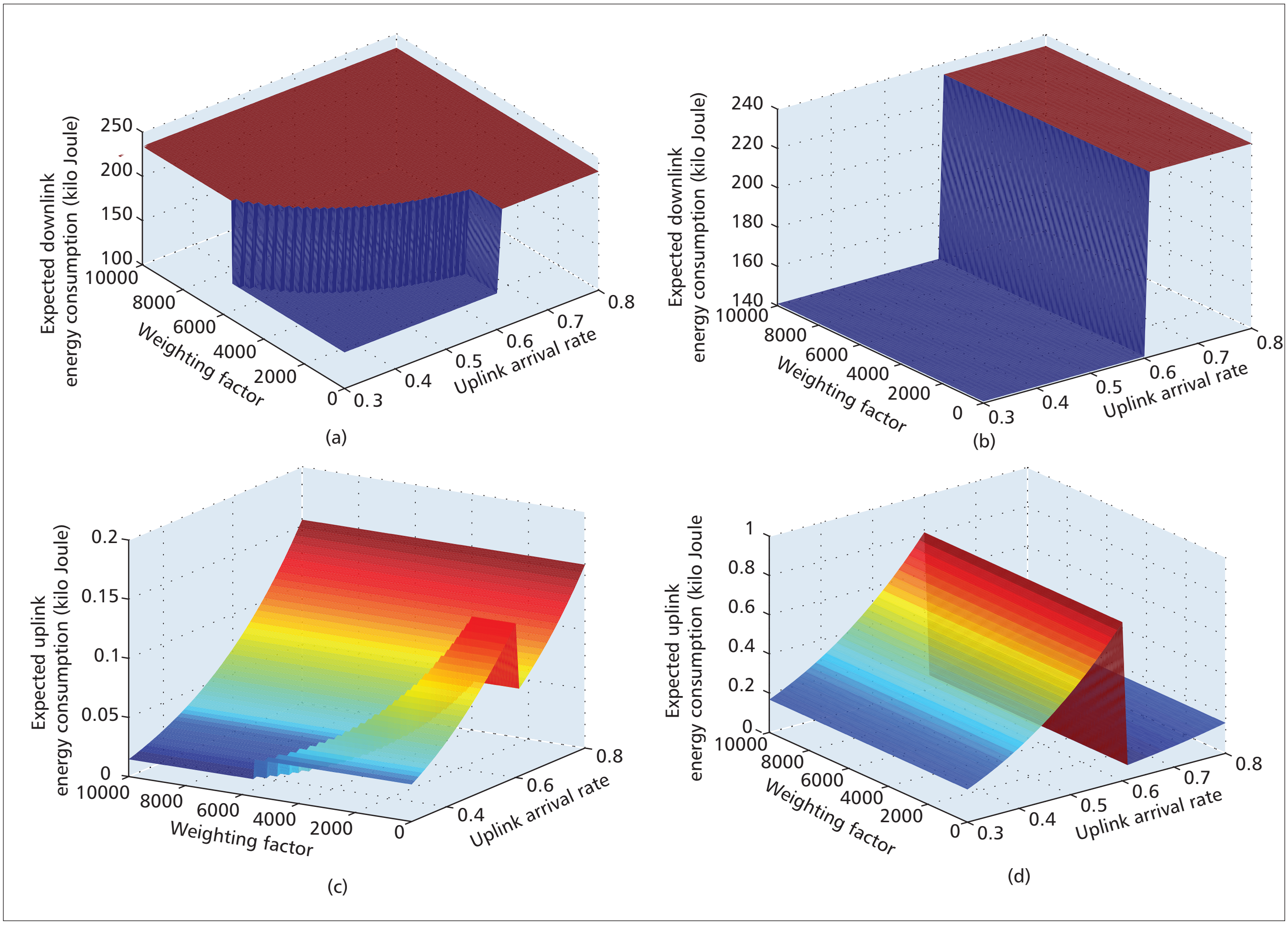


Figure 4. The expected energy consumption versus the arrival rate of uplink users and the weighting factor β: a) downlink expected energy for balanced approach; b) downlink expected energy for unbalanced approach; c) uplink expected energy for balanced approach; d) uplink expected energy for unbalanced approach. The spatial distributions are $\rho_{UL}(A_{2,1}) = 0.7$ and $\rho_{DL}(A1,1) = 0.8$ for uplink and downlink users, respectively.

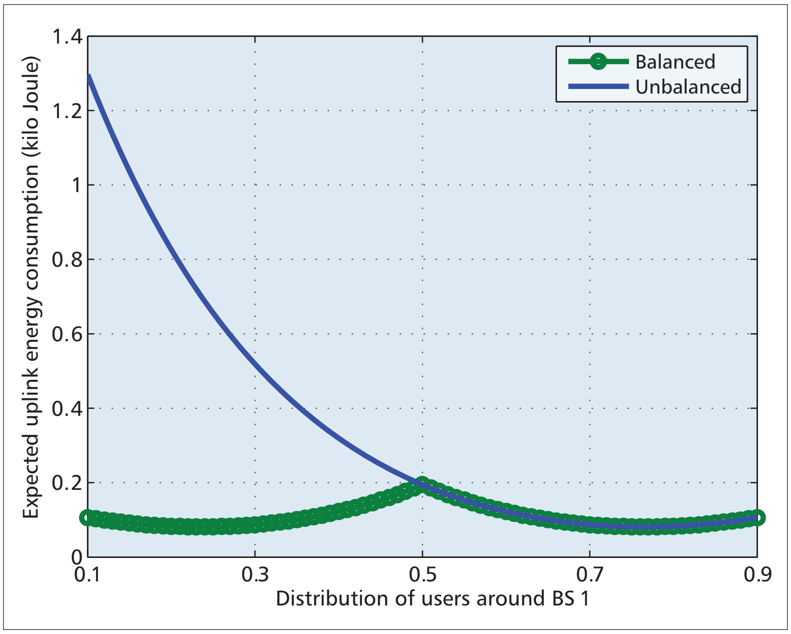


Figure 5. The expected energy consumption of uplink users versus the spatial distribution of the uplink users near the proximity of the first BS. The uplink users' arrival rate is 0.4.

cooperation incentives will be investigated to motivate different networks' cooperation in dynamic planning. Finally, future work will aim to balance backhaul links for improved energy efficiency of cellular networks [15].

Acknowledgment

This publication was made possible by NPRP grant # NPRP 6-415-3-111 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] M. Ismail and W. Zhuang, "Network Cooperation for Energy Saving in Green Radio Communications," *IEEE Wireless Commun.*, vol. 18, no. 5, Oct. 2011, pp. 76–81.

[2] L. Budzisz *et al.*, "Dynamic Resource Provisioning for Energy Efficiency in Wireless Access Networks: A Survey and an Outlook," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 4, 2014, pp. 2259–85.

[3] K. Son *et al.*, "Base Station Operation and User Association Mechanisms for Energy-Delay Trade-Offs in Green Cellular Networks," *IEEE JSAC*, vol. 29, no. 8, Sep. 2011, pp. 1525–36.

[4] T. Han and N. Ansari, "On Greening Cellular Networks via Multicell Cooperation," *IEEE Wireless Commun.*, vol. 20, no. 1, Feb. 2013, pp. 82–89.

[5] E. Oh, K. Son, and B. Krishnamachari, "Dynamic Base Station Switching-On/Off Strategies for Green Cellular Networks," *IEEE Trans. Wireless Commun*., vol. 12, no. 5, May 2013, pp. 2126–36.

[6] C. Y. Chang *et al*., "On Optimal Cell Activation for Coverage Preservation in Green Cellular Networks," *IEEE Trans. Mobile Computing*, to appear.

[7] Z. Niu *et al*., "Cell Zooming for Cost-Efficient Green Cellular Networks," *IEEE Commun. Mag*., vol. 48, no. 11, Nov. 2010, pp. 74–79.

[8] N. Saxena, B. J. R. Sahu, and Y.S. Han, "Traffic-Aware Energy Optimization in Green LTE Cellular Systems," *IEEE Commun. Lett*., vol. 18, no. 1, Jan. 2014, pp. 38–41.

[9] J. Wu, S. Zhou, and Z. Niu, "Traffic-Aware Base Station Sleeping Control and Power Matching for Energy-Delay Trade-Offs in Green Cellular Networks," *IEEE Trans. Wireless Commun*., vol. 12, no. 8, Aug. 2013, pp. 4196–4209.

[10] S. Navaratnarajah, A. Saeed, M. Dianati, and M. A. Imran, "Energy Efficiency in Heterogeneous Wireless Access Networks," *IEEE Wireless Commun*., vol. 20, no. 5, Oct. 2013, pp. 37–43.

[11] A. Conte *et al*., "Cell Wilting and Blossoming for Energy Efficiency," *IEEE Wireless Commun*., vol. 18, no. 5, Oct. 2011, pp. 50–57.

[12] D. L. Perez *et al*., "OFDMA Femtocells: A Roadmap on Interference Avoidance," *IEEE Commun. Mag.*, vol. 47, no. 9, Sept. 2009, pp. 41–48.

[13] Y. Zhuang *et al*., "A Geometric Probability Model for Capacity Analysis and Interference Estimation in Wireless Mobile Cellular Systems," *IEEE Globecom*, Dec. 2011, pp. 1–6.

[14] V. Goswami and U. C. Gupta, "Analyzing the Discrete-Time Multiserver Queue Geom/Geom/M Queue with Late and Early Arrivals," *Information and Management Sciences*, vol. 9, no. 2, June 1998, pp. 55–66.

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

Biographies

Muhammad Ismail [S'10, M'13] is an assistant research scientist in the Electrical and Computer Engineering Department at Texas A&M University at Qatar. He received his Ph.D. degree in electrical and computer engineering from the University of Waterloo, Canada in 2013, and the M.Sc. and B.Sc. degrees with honors in electrical engineering (electronics and communications) from Ain Shams University, Cairo, Egypt in 2007 and 2009, respectively. Dr. Ismail's research interests include distributed resource allocation, green wireless networks, cooperative networking, smart grid, and biomedical signal processing. He is a co-recipient of the best paper awards at IEEE ICC 2014, IEEE Globecom 2014, and SGRE 2015. He is a co-author of two research monographs by Springer and Wiley/IEEE Press. He served as a TPC member for the ICWMC in 2010–2014 and IEEE ICC 2014 and 2015. He served on the IEEE INFOCOM 2014 organizing committee. He has been an associate editor with *IET Communications* since 2014. He was an editorial assistant for *IEEE Transactions on Vehicular Technology* from January 2011 to July 2013.

Mohamed Kashef [M'14] is a postdoctoral research associate with Texas A&M University at Qatar. He obtained his Ph.D. in electrical engineering from the University of Maryland at College Park in 2013. He received the B.Sc. and the M.S. degrees with honors in electronics and electrical communications engineering from Cairo University, Cairo, Egypt, in June 2006 and August 2009, respectively. His research interests generally include wireless communication systems and networks, visible light communication networks, and optimization of stochastic systems.

Erchin Serpedin [F'13] is a professor in the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. He received the specialization degree in signal processing and transmission of information from Ecole Superieure D'Electricite (SUPELEC), Paris, France, in 1992, the M.Sc. degree from the Georgia Institute of Technology, Atlanta, USA, in 1992, and the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, USA, in January 1999. Dr. Serpedin is the author of two research monographs, one textbook, nine book chapters, 110 journal papers, and 180 conference papers. His research interests include signal processing, biomedical engineering, bioinformatics, and machine learning. Dr. Serpedin is currently serving as an associate editor of *IEEE Signal Processing Magazine* and as the Editor-in-Chief of *EURASIP Journal on Bioinformatics and Systems Biology*, an online journal edited by Springer. He has served as an associate editor of dozens of journals, such as *IEEE Transactions on Information Theory*, *IEEE Transactions on Signal Processing*, *IEEE Transactions on Communications*, *IEEE Signal Processing Letters*, *IEEE Communications Letters*, *IEEE Transactions on Wireless Communications*, *Signal Processing* (Elsevier), *Physical Communications* (Elsevier), *EURASIP Journal on Advances in Signal Processing*, and as a Technical Chair for five major conferences. He received numerous awards and research grants.

Khalid Qaraqe [M'97, SM'00] received the B.S. degree in EE from the University of Technology, Iraq, in 1986, with honors. He received the M.S. degree in EE from the University of Jordan, Jordan, in 1989, and he earned his Ph.D. degree in EE from Texas A&M University, College Station, TX, in 1997. From 1989 to 2004 Dr. Qaraqe held a variety of positions in many companies, and he has more than 12 years of experience in the telecommunication industry. Dr. Qaraqe has worked for Qualcomm, Enad Design Systems, Cadence Design Systems/Tality Corporation, STC, SBC, and Ericsson. He has worked on numerous GSM, CDMA, and WCDMA projects, and has experience in product development, design, deployments, testing, and integration. In July 2004 Dr. Qaraqe joined the Department of Electrical and Computer Engineering at Texas A&M University at Qatar, where he is now a professor. Dr. Qaraqe's research interests include communication theory and its application to design and performance, analysis of cellular systems and indoor communication systems. Particular interests are in mobile networks, broadband wireless access, cooperative networks, cognitive radio, diversity techniques, and beyond 4G systems.

The proposed approach decouples the MT association and BS operation phases based on a two time scale decision problem. It introduces a new BS switch off metric based on the uplink mobile users' spatial distribution. Also, it accounts for the mobile users' uplink energy consumption in BS wake up, which avoids service quality degradation in the uplink.

GREEN COMMUNICATIONS AND COMPUTING NETWORKS

Post-Peak ICT: Graceful Degradation for Communication Networks in an Energy Constrained Future

Sofie Lambert, Margot Deruyck, Ward Van Heddeghem, Bart Lannoo, Wout Joseph, Didier Colle, Mario Pickavet, and Piet Demeester

The authors are with Ghent University-iMinds.

ABSTRACT

In recent years, rising energy prices and increasing environmental concerns have boosted research in the so called *green ICT* and *green networking* research tracks, aimed at improving the energy efficiency of communications while still offering maximal functionality. In this article we explore a future scenario in which low power networking is no longer optional, but instead becomes a necessity due to fluctuating energy availability. The contribution of this work is twofold. First, we argue why a so called post-peak future scenario, in which we can no longer rely on fossil fuels as our main resource for electricity production, is not unlikely, and what it might entail. Second, we explore the consequences of such a scenario for ICT: How well can current and future infrastructures cope with temporary energy limitations? As an illustration, we present a case study showing the impact of reduced energy availability on a wireless access network.

INTRODUCTION

In this article we consider a post-peak future scenario in which fossil fuels are no longer the main energy source for electricity production (fossil fuels are "past their peak"), but instead are increasingly replaced by alternative energy sources. We start our description by discussing why a post-peak scenario may be imminent, and how this could result in temporary energy restrictions for information and communications technology (ICT) networks. Next, we outline a framework to evaluate the post-peak potential of technologies and propose new avenues of research to prepare ICT infrastructures for a post-peak situation. In a specific case study for a wireless access network, we modified a wireless planning tool to make optimal use of energy in a post-peak situation.

WHY SHOULD WE ANTICIPATE A POST-PEAK FUTURE?

Present-day societies and economies run mostly on fossil fuels. Oil (petroleum), coal, and natural gas made up 82 percent of the world's primary energy supply in 2012, and 68 percent of electricity was generated by burning fossil fuels, as shown in Fig. 1a.

It is likely that in the coming years societies will be forced to move from fossil fuels to alternative sources for their energy supply. We see three reasons why this may happen in the near future, in order of increasing importance: fossil fuel depletion, security of energy supply, and the impact of fossil fuels on climate change. Some scientists argue that current trends in fossil fuel consumption will result in a peak of conventional oil production before 2030 (referred to as *peak oil*), leading to a global fuel shortage and steep increase in oil prices [1]. Whether this peak oil will indeed occur in the near future is a contested point in scientific literature, but even adversaries of the peak oil theory agree that political instability in oil-producing countries can result in severe oil price shocks, with devastating effects for fossil fuel-based economies [2]. This brings us to the second argument: security of the energy supply. Countries that switch to renewable energy sources reduce their vulnerability to oil price shocks [3]. Finally, and most importantly, if policy makers take the climate change challenge seriously, renewables offer one of the few virtually carbon-neutral alternatives to fossil fuels (Fig. 1b).

Nuclear energy could make an increasing contribution to low-carbon energy supply, but a variety of barriers and risks exist. It is costly compared to alternatives, particularly when the high risks involved are factored in. In addition, a nuclear power renaissance would also increase the risk of nuclear terrorism and make efforts to control the spread of nuclear weapons much more difficult.

0163-6804/15/$25.00 © 2015 IEEE

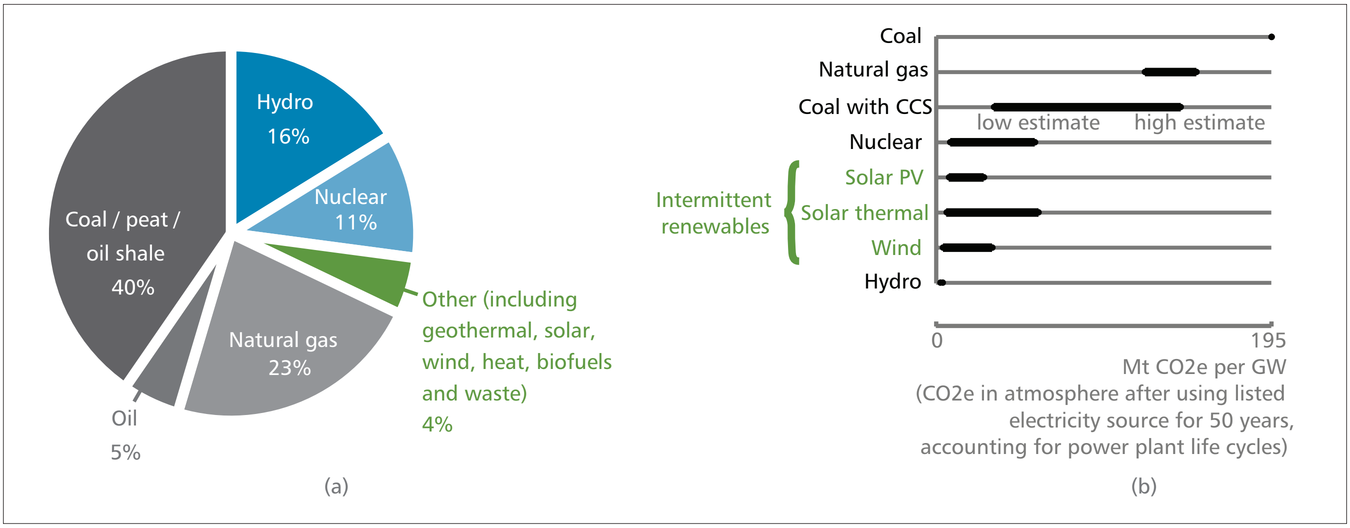


Figure 1. In order to reduce the carbon footprint of electricity generation, the share of fossil fuels must be reduced, and intermittent renewables (solar photovoltaic, solar thermal and wind energy) will need to contribute a bigger share in the energy mix. a) Worldwide fuel shares of electricity generation in 2012 (source: Key world energy statistics by the International Energy Agency, published in 2014); b) Carbon footprint of various electricity sources (high and low estimates taken from [4]) CCS = carbon capture and storage.

Due to their intermittent nature, renewables replacing fossil fuels will not be able to offer the same supply continuity as present-day energy provisioning systems. Some of the most mature renewable energy technologies that can be deployed on a large scale and at a relatively low cost, such as wind and solar energy, have varying outputs that depend on fluctuating weather conditions.

Since operation of the electricity grid requires energy production and energy consumption to be in balance at all times, low production periods must either be matched with low consumption, or the production deficit must be compensated with previously stored energy and/or energy imports. In an ideal smart-grid scenario, temporary low production can be matched on the demand side by postponing non-urgent energy consumption by homes and industries. The intermittency of renewable sources could also be (partly) compensated by constructing large energy storage facilities, e.g. using hydro pumped storage or batteries, or by interconnecting production capacities over a large geographical area, taking advantage of weather locality to reduce the chance of an overall low production.

However, bringing these upgrades to the power provisioning system in place will take time, as they require a widespread introduction of smart meters and controllable devices, and the construction of large energy storage facilities and long-distance high-capacity power lines. Meanwhile, the consequences of the transition away from carbon-heavy and nuclear sources may already manifest themselves in the very near future. Looming electricity shortages have been reported in news outlets of several developed nations in the past year (examples include Belgium, Germany, the United Kingdom, and Japan). Moreover, even after future utility networks are properly dimensioned, there may still be temporary power shortages in rare periods of extreme weather conditions.

To conclude this discussion, we remark that even though we cannot be certain that we are headed for an energy intermittent future, the outlined trends indicate that the possibility is real, and we can only deal with the consequences properly if we are prepared. Even today, this knowledge could already be useful in other energy-constrained situations, such as disaster recovery or off-grid installations in developing countries, to make optimal use of the energy available in emergency generators and back-up batteries.

Consequences for ICT

When there is a drop in energy production (lasting from a couple of hours to several days), and the energy gap cannot be filled by burning abundant fossil fuels, governments may impose temporary restrictions on power consumption to avoid a collapse of the grid. A number of measures can be taken: cutting off (geographical) clusters of consumers,[1] applying dynamic pricing schemes that reflect energy availability,[2] or forcing large energy consumers to curb their consumption.

If governments decide to impose energy restrictions on large consumers, ICT service providers and telecom operators will also be targeted, as their extensive infrastructures use considerable amounts of power. For example, Telecom Italia is Italy's second biggest electricity consumer, and British Telecom consumed 0.76 percent of the U.K.'s national electricity consumption in 2011/12, making it the largest single electricity consumer in the country. Globally, communication networks account for approximately 1.7 percent of total electricity consumption; data centers consume another 1.4 percent [5].

Evidently, pulling the plug on ICT infrastructure altogether is not a desirable option as it supports several critical applications. But some of the services on offer may be more dispensable

[1] The Belgian government installed a "disconnection plan" (Dutch: afschakelplan) after the unforeseen shutdown of two nuclear reactors in 2014. The plan allows the electricity provider to temporarily cut off clusters of consumers when demands exceed supply, for example during peak hours (17h–20h) on a cold winter day.

[2] Throughout this article, when we talk about energy restrictions or shortages, this does not necessarily mean there is no energy available at all, but rather, that scarcity can make it prohibitively expensive.

Reducing the maximum power consumption and improving the power proportionality (making power consumption scale with load, for example, through load-adaptive network operation) have been important tracks in Green ICT research, and will become even more important when targeting graceful degradation for post-peak ICT.

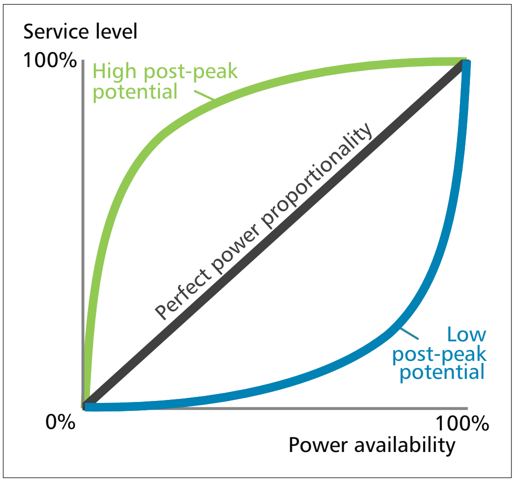


Figure 2. Depending on how their service level scales with power availability, technologies may have *high post-peak potential*, offering relatively good service for a fraction of their normal power, or *low post-peak potential*, losing service level rapidly even for a small power reduction. This relationship is influenced by the choice of service level metric, which can be based on criteria such as speed, reliability, user coverage.

than others (e.g. omnipresent broadband Internet access vs. lifeline communications). Therefore, we introduce the concept of *graceful degradation* under energy constraints. If the available energy for ICT is drastically reduced, falling back to 50 percent, 25 percent, or a mere 10 percent of regular power levels, can we still offer minimal functionality and connectivity? In order to answer this question, we need to determine what part of the functionality is truly indispensable, and how much power is needed to keep it running. This will be the focus of our evaluation of ICT technologies in the post-peak context.

Related Work

This is a relatively new research domain as we are only just starting to see signs of an impending post-peak future. The most notable existing work in the field is by B. Raghavan. His 2011 publication [6] was the first to consider the implications of a permanent energy crisis for ICT, listing a number of post-peak design principles and research questions. The main difference with our work is that Raghavan considered a scenario where energy demands would exceed supply overall and the current Internet architecture could no longer be used, whereas we focus on dealing with short term energy limitations, assuming the current network architecture is still in place and functioning normally most of the time.

The Concept of Graceful Degradation

In a post-peak context, we want to know how the delivered service level of a network or device scales with the available power. Three generic power profiles are shown in Fig. 2. Without graceful degradation, the delivered service level will quickly drop to (almost) zero when the available power decreases, as indicated by the lower line in Fig. 2. These infrastructures have *low post-peak potential*, as they are unable to function under energy constraints. The upper two lines in Fig. 2 represent devices or networks that do allow graceful degradation. In an infrastructure with perfect power proportionality, the service level decreases at the same rate as the available power; infrastructures with *high post-peak potential* are able to offer a high service level even if the available power is decreased significantly.

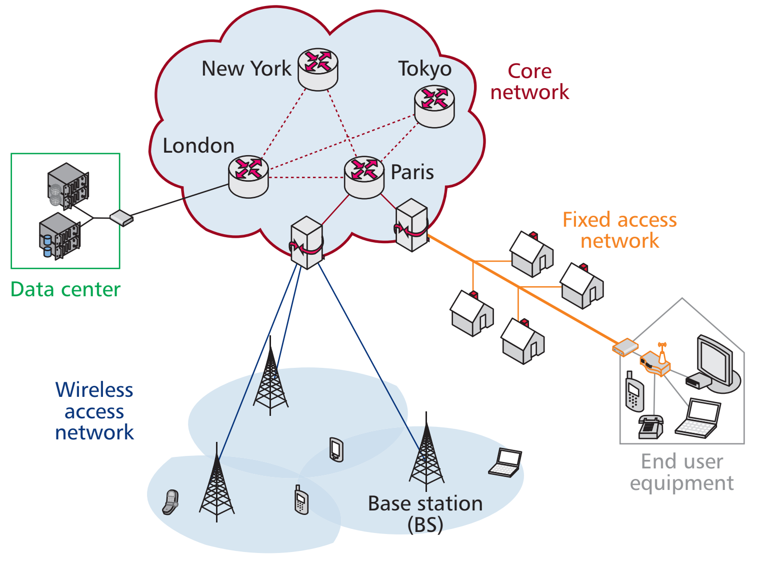
As we will discuss in more detail later, a low post-peak potential is typically associated with dedicated devices, whereas a high post-peak potential is typically available when multiple resources can be flexibly shared. The shape of the power profile of the device or network under study depends on the power proportionality that it exhibits, as well as on how its service level metrics are defined, as explained below.

Power Proportionality

In Fig. 2, 100 percent power availability corresponds to what the device or network consumes in a normal energy situation to offer maximal service. This maximum power consumption will depend on the specific equipment under study, and whether it uses current (energy-hungry) or future (greener) technologies. The profile is also influenced by the impact on power consumption of a service reduction. Currently deployed networks typically exhibit a power consumption that is relatively constant despite strong diurnal variations in the traffic load or service, corresponding to the lower curve of Fig. 2. Reducing the maximum power consumption and improving the power proportionality (making power consumption scale with load, for example, through load-adaptive network operation) have been important tracks in Green ICT research [7], and will become even more important when targeting graceful degradation for post-peak ICT.

Service Level Metrics

For a given device or network, the power profile will also depend strongly on the metrics used to define the service level. This can be a quantitative assessment, based on a weighted function of data throughput, bandwidth, uptime, error rate, user coverage, computation capacity, and so on; or it can be more qualitative, when certain applications or services are considered critical and thus part of the *minimal service*, while others are dispensable. For an existing example of graceful degradation, consider a smartphone with a battery that is running low. Certain applications, such as the camera functionality, may be temporarily disabled to ensure that the minimal service, texting and phone calls, is guaranteed for as long as possible. If we define the service level of the smartphone as the number of calculations per second, the service under energy constraints may be quite poor; but if we define the service level as the ability to communicate, it is still relatively high.



Examples of post-peak power reduction measures

Fixed access network

- Improve customer premises equipment (CPE) power proportionality
- Centralized control of CPE power states

Wireless access network

- Selectively switch off some BSs, based on different strategies:
  - Lowered capacity or reduced covered area
  - Support fewer wireless technologies

Core network

- Reduce level of resilience or overprovisioning
- Change of virtual topology
- Block specific traffic

Data center

- Reschedule user requests:
  - To a later time
  - To another location

Figure 3. Schematic of the network showing the areas under study (fixed and wireless access networks, core networks and data centers), along with examples of post-peak solutions for these areas.

Graceful Degradation in ICT Infrastructures

In this section we assess the potential for graceful degradation in various ICT infrastructures. We start by introducing some terminology concerning the areas under study, then we look into specific solutions for each of these areas.

Overview of ICT Infrastructures

Figure 3 gives an overview of a typical network infrastructure. The access network provides a physical connection to the end users through which they can connect to the Internet. In *fixed access networks*, a physical wire runs to the end user's premises, where the signal is decoded by a modem and distributed further in the local home or enterprise network. In *wireless access networks*, subscribers use standardized radio signals to connect to the nearest base station (BS), from where the signal is forwarded through a dedicated backhaul link to an aggregation point. Traffic from the access networks is aggregated and transmitted further through the backbone or *core network*, which provides high-capacity, low-latency connections across large geographical distances.

The focus of this article is on devices that can be controlled centrally by the Internet service provider (ISP). This means that, despite their significant contribution to ICT power consumption, end devices controlled by the consumer such as TVs, personal computers, or mobile phones are outside the scope of this work. Modems, routers, and wireless access points (AP), installed at the customer premises, may be considered part of the fixed access network, depending on whether they can be controlled remotely by the ISP. *Data centers* can also be controlled centrally, and form the last important contributors to network power consumption. These facilities house large numbers of computer and storage systems to host a wide range of applications, from websites over search engines to cloud computing.

Field-Specific Solutions and Limitations

In the following, we explore the opportunities and limitations for graceful degradation in the four fields introduced above. The list of proposed measures is by no means exhaustive, but rather intended as a starting point for further research. While we discuss the solutions for these fields separately, post-peak strategy design should also keep a holistic overview of the interactions between different fields, and see whether some services can be substituted for others. For example, guaranteeing both wired and wireless connectivity may not be feasible under certain energy constraints, but as long as one of the communication channels remains available, this may be sufficient as a minimal service.

Fixed Access Networks: Since most of the power in fixed access networks is consumed by the customer premises equipment (CPE) [8], this is where we direct our first efforts to try to optimize the use of limited power resources. Currently deployed modems, wired and wireless routers typically show bad power proportionality with respect to data rates, corresponding to the low post-peak potential profile in Fig. 2.

However, if we define the service level as the number of customers being served, and consider the power consumption of the access network as a whole instead of on a single device level, switching off a selection of CPEs will scale down power consumption as the service level decreases, corresponding to a more power-proportional profile (the straight line in Fig. 2).

Depending on how the minimal service is defined, several strategies can be used to power off a well-chosen selection of access equipment. A *time division based strategy* could divide the access network into regions, and assign time slots for each region when connectivity would be

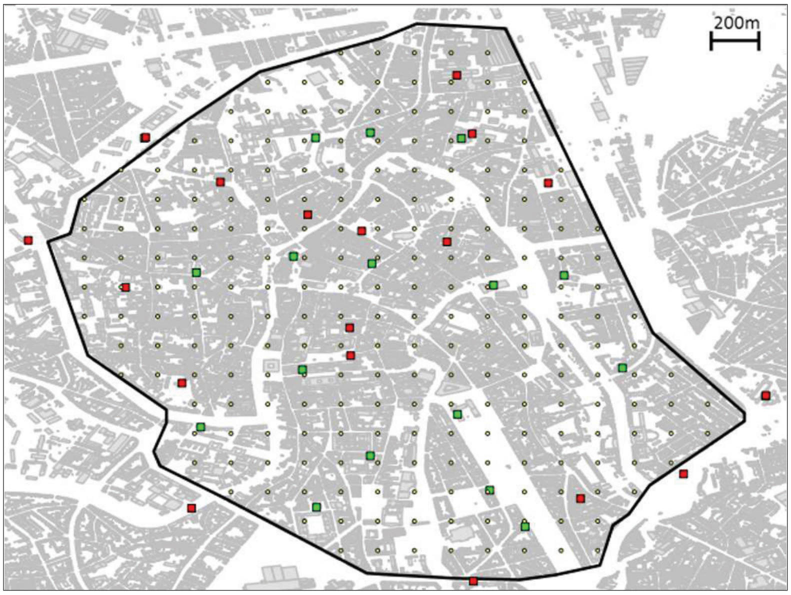


Figure 4. The selected suburban area of 6.85 km^2 in Ghent, Belgium. Users are located on the yellow dots; squares indicate base station (BS) sites. As an illustration of the sleep mode principle, the active and inactive BSs for a simulation with 70 percent available power and random user distribution are colored green and red, respectively.

available, powering off all access equipment in that region for the remaining time. This would give users periodic, deterministic access to the network. Alternatively, in dense urban networks a *wireless ad hoc strategy* could be applied, taking advantage of the relatively dense deployment of wireless AP. Here, a large fraction of the modems and AP could be powered down, while almost complete wireless coverage of the area (at reduced data rates) could be maintained by connecting users to neighboring hot spots. Note that for these strategies to work, the network provider should be able to power down modems and routers remotely.

In larger office environments, local area networks (LANs) can be structured so that less critical parts can be selectively powered off while critical connections are maintained.

Wireless Access Networks: In wireless access networks, BSs are the main power consumers. The power consumption of a single BS does not scale well with the traffic load, corresponding to low post-peak potential in Fig. 2. As a consequence, just as in fixed access networks, the only way to reduce power consumption in current networks is by switching off some of the equipment and looking at the service level on the network scale instead of the device scale. We propose two strategies for BS switch-off below, designed to provide the maximum possible service out of the limited available energy during a post-peak shortage.

The first is to *cut capacity by reducing network density*. There is a trade-off between the capacity and the range of a single BS. For a given input power, decreasing the capacity (bit rate) results in longer ranges and thus a larger area covered by the BS. In a post-peak situation, the capacity per user can be limited so that the required network capacity would be reduced and, since the range of the BSs in a low-capacity network is bigger, the number of active BSs could be decreased. The achievable savings are calculated for a realistic case below. Depending on what fraction of the normal operating power is available, this strategy may result in reduced coverage. A wireless ad-hoc functionality for emergency communications, where information can be relayed through other end users' devices to reach the access network, may be worth investigating.

The second strategy is to *fall back to single-standard support*. In current mobile networks, overlapping coverage is offered for different standards. The most widespread technology is the second generation (2G) standard Global System for Mobile Communications (GSM). Third-generation (3G) Universal Mobile Telecommunications System (UMTS) and fourth-generation (4G) Long-Term Evolution (LTE) sites are often built at pre-existing 2G sites to increase peak data rates and the maximum number of user connections. During a post-peak temporary energy shortage, power consumption might, for example, be decreased by disabling the 3G and 4G BSs, falling back to the GSM network (GSM is currently still the most widely supported technology on handsets). This could result in a drastic reduction in power consumption, since BS consumption is similar across technologies (in the order of 1–2 kW/BS) [9]. For subscribers, this measure would result in a noticeable reduction in quality of experience, as network capacity and peak data rates would be reduced, but coverage would still be guaranteed.

Core Networks: During a post-peak shortage, the traffic load in the core network will decrease significantly as a consequence of traffic reductions realized in access networks and data centers. Unfortunately, because of the current weak power proportionality in response to changes in the traffic load, power consumption in the core network will remain almost constant, once again corresponding to a low post-peak potential profile in Fig. 2. In future deployments we might see new equipment and techniques that improve power proportionality, such as bit rate variable transponders and power aware routing schemes that increase the potential for sleeping interfaces [10]. In anticipation of these developments, we can think of four other approaches that could keep core networks running under post-peak power reductions.

The first is through *reduced resilience*. Typically, protection mechanisms are in place to almost instantly switch traffic over a secondary path in case of a link or node failure. Turning off backup equipment could reduce power consumption by a factor of two. As a result, recovery would no longer be instant (within 50 ms), but would take a couple of seconds when a recovery path is discovered and set-up automatically, or several minutes or hours when links need to be brought back online manually (in case of insufficient capacity). A second, similar approach is to *temporarily reduce the overcapacity* that is installed to handle peak-to-mean traffic variations and unexpected traffic spikes [11].

This could decrease power consumption in the core by another factor of two (or more). However, unexpected traffic spikes would no longer be handled flawlessly and traffic bottlenecks would occur in anticipation of extra capacity being (manually) brought online. Third, we could *change the virtual network topology*. Several works have indicated that with traffic demands currently being higher than the equipment line rates, fully-meshed virtual (IP) topologies are more energy efficient than ring topologies [12]. Further research is needed to investigate which topology (mesh, ring, etc.) would be most efficient in a post-peak scenario with reduced traffic demands, and how it would affect latency. The fourth and last approach we propose is *blocking service-specific traffic*, where non-critical services are temporarily blocked in the core, insofar as this is not already done at the network edge. This technique could also be used to reduce traffic demand to a level where a topology reconfiguration (introduced above) can bring additional savings.

Data Centers: In [13] the authors discuss how data center management should be revised to maximize the use of off-grid renewables. These techniques can be adapted for use in a post-peak scenario, taking into account the more restrictive energy limitations. Below we list a number of ways to temporarily reduce the workload in data centers. Note that these measures can only reduce power consumption significantly if the data center manager is able to plan capacity by turning off selected groups of machines when the load is reduced, thus achieving high post-peak potential (Fig. 2).

The first logical step is to delay system maintenance tasks such as system updates and backups. Incoming user requests can also be *rescheduled* to a later time or *migrated*, if possible, to another location where energy is more abundant. If workloads cannot be migrated but must be reduced, a *priority label* could be given to the most critical data and services, which would then be placed on machines that are kept on at all times to guarantee the minimal service. This priority label could be assigned manually based on service level agreements (SLAs) with customers, or automatically to the most frequently accessed content. Alternatively, all user requests could be handled with the same priority, but with less resources, keeping all services available, but at the cost of longer response times, including that of critical services.

Case Study: Wireless Access Network with Reduced Power

Realistic Scenario for a Suburban Area

As a concrete case study, we evaluate the post-peak potential of a wireless access network by simulating an LTE (4G) network in a suburban area of Ghent, Belgium. Figure 4 shows the selected area together with the locations of the 35 simulated BS sites (based on real BS positions). Table 1 lists the LTE characteristics.

We assume 224 candidate users who want to be simultaneously active, each requesting 64 kb/s, which is more than enough to make a phone call. The location of the users within the selected area is known (a realistic mix of indoor and outdoor users is assumed), but the order in which they connect to the network varies randomly across simulation runs. A heuristic deployment algorithm, based on the one described in [14], chooses which BSs are active and which are in sleep state. The objective of the algorithm is to maximize the number of active users, that is, users connected to an active BS, by switching BSs on and off in response to the simulated user locations and load, and setting the power level for each BS based on its optimal reach. The original algorithm in [14] was designed to achieve a predefined QoS and coverage with minimal power consumption. In the post-peak version, we introduce an additional constraint: the overall power consumption (summed over all BSs) cannot exceed a pre-defined power budget. This may result in some loss of QoS and coverage.

Property	Value
Carrier frequency	2600 MHz
Channel bandwidth	5 MHz
Minimum power/active BS (antenna input power 1 dBm)	1204 W
Maximum power/active BS (antenna input power 43 dBm)	1672 W
BS power in sleep state – scenario zero sleep power – scenario 45 percent sleep power	 0 W 752 W
(Coding rates) modulation schemes	(1/3, 1/2, 2/3) QPSK (1/2, 2/3, 4/5) 16-QAM (2/3) 64-QAM
Bit rate/BS [1/3 QPSK – 2/3 64-QAM]	2.8 Mb/s – 16.9 Mb/s
Range/BS [1/3 QPSK – 2/3 64-QAM] (NLOS (non-line-of-sight), @43 dBm)	1090 m – 194 m

Table 1. Properties of the considered LTE (4G) radio technology.

We consider two scenarios for the power consumption of BSs in sleep mode: perfect sleep, where a sleeping BS consumes no power; and a more realistic sleep power, which would typically be approximately 45 percent of the maximum power [15] (see Table 1 for values).

Simulation Results

The heuristic deployment algorithm is applied to the area under study with varying constraints on overall power consumption. This results in the power profiles shown in Fig. 5, where each simulated point corresponds to the median of 40 simulation runs. The horizontal axis in Fig. 5 shows the network power consumption as a percentage of the maximum power consumption (100 percent = power for a network optimized for maximum coverage). The vertical axis shows the service level, which can be defined using two different metrics. The user coverage indicates the percentage of users that can get the minimum bit rate they are requesting from the active BSs (in this case, 64 kb/s). The geometrical coverage

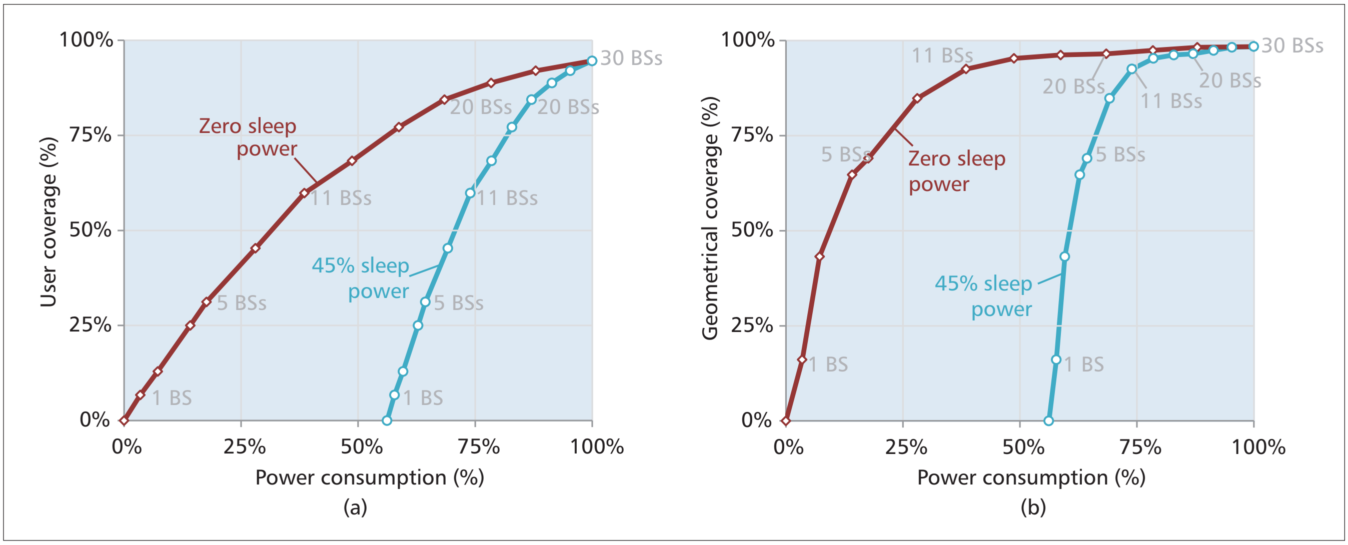


Figure 5. Simulation results of the wireless case study for LTE (4G) technology. When BSs consume negligible power in the inactive state (zero sleep power, dark red lines), both user and geometrical coverage have high post-peak potential. When BSs require a significant fraction of their maximum power in sleep state (45 percent sleep power, light blue lines), most of the post-peak potential is lost.

indicates the percentage of the (outdoor) area that can be reached by signals from the active BSs at the lowest coding rate and modulation scheme (1/3 QPSK), offering a bit rate of 2.8 Mb/s per BS (Table 1). User coverage is a more strict metric, as a geographical location may be within range of a BS while a user in that location may not be covered as such, due to an uncovered indoor position or capacity limitations of the BS.

When BSs consume negligible power in the inactive state (zero sleep power, dark red lines in Fig. 5), the post-peak potential for user coverage is limited; when 25 percent of the power is available, user coverage for LTE is only approximately 30 to 40 percent. The power profile for geometrical coverage looks more promising, since approximately 75 percent coverage can be guaranteed with only 25 percent of the power. This is because after most of the area is covered, the energy cost of adding geometrical coverage increases as filling a small coverage gap will still require a complete additional BS to be switched on, and the minimum power per BS is 1204 W (Table 1).

The geometrical coverage could be considered the fraction of users that can be served when only outdoor coverage is guaranteed and extremely low data rates are allowed, offering basic communications such as text messaging. Depending on how the minimal service is defined — should it include voice communications and possibly even low-data rate exchange; what about indoor coverage? — one could opt for higher geometrical coverage (avoiding large outdoor coverage gaps) or better user coverage in densely populated areas (leaving large coverage gaps in remote areas, but keeping indoor coverage and higher rates available in urban areas).

When BSs require a significant fraction of their active power in sleep state (45 percent sleep power, light blue lines in Fig. 5), most of the post-peak potential is lost, as 56 percent of the maximum power is still needed even when all BSs are in sleep mode (= minimum network power consumption) and thus no service at all is offered. Note that theoretically only 45 percent of the power would be needed if the network were perfectly dimensioned and all BSs were switched on at maximum load. However, activating 30 out of 35 BSs was sufficient to reach the maximum attainable coverage for maximum load; hence, the 100 percent power consumption already corresponds to the consumption of 30 active and five sleeping BSs. In any case, the post-peak potential of a deployment with relatively high sleep power is clearly insufficient, as no service at all can be offered when energy availability is temporarily below 50 percent.

Conclusion

Conclusions of This Work

We studied the effect of post-peak energy shortages on various ICT infrastructures: fixed and wireless access networks, core networks, and data centers. We assessed whether these infrastructures could still offer a basic functionality when the available power was significantly lowered compared to normal operating conditions.

We distinguished between low and high post-peak potential infrastructures and devices (Fig. 2). Based on the ICT fields we studied, we conclude that the post-peak potential on a single-device level is typically low, as most devices are unable to offer half of the normal service (defined as the number of calculations, data rate, etc.) when the available power is halved. Therefore, the post-peak potential of dedicated devices to which users are statically assigned is very limited. On the other hand, in networked environments, when users share resources and flexible resource assignment is possible, the post-peak potential on a wider scale can be much higher. This is, for example, the case in large-scale data centers or wireless access networks, where some of the capacity can be temporarily disabled while still offering a minimal service to all users.

We also considered the concrete case of a suburban wireless access network under varying energy constraints. Simulation results showed that the post-peak potential can only be realized if low power modes are truly low-power and inactive BSs consume negligible power. This is in contrast with present-day sleep modes for BSs, which typically consume approximately 45 percent of the active power.

Moving Forward

This is a relatively new research domain, and although we touched upon a number of strategies already, handling a post-peak situation will no doubt require further research into a broad range of applications and domains. When investigating post-peak features, there are a number of pitfalls that need to be kept in mind.

A first issue is that indirect effects of a post-peak situation on user behavior are hard to predict. For example, offering a slower service could lead to a reduced load if users give up non-urgent activities in response to the reduced quality of experience, but it could also increase the load if users' connection times are increased due to the slower service. Indirect effects on user behavior could also be used to enhance post-peak savings, for example by suppressing TV signals for the indirect effect of having less television sets switched on. Interdisciplinary research is needed to predict these kinds of effects.

A second important issue is that systems will become more vulnerable at the time of energy shortages. While solutions should be chosen carefully to minimize vulnerability, some loss of reliability during a post-peak energy shortage will be inevitable. We should keep in mind that a minimal service level is still preferable to no service at all.

Last but not least, ICT is not only an important power consumer, it can also play a key role in controlling the power consumption of other loads. In a post-peak future, a smart grid that gathers information about suppliers and consumers and helps control them, will be an important instrument to balance electricity supplies and demands. This should be studied in a dedicated research track, considering incentives that shape consumer behavior and opportunities for automated power control. Post-peak measures for ICT specifically will need to guarantee the exchange of smart grid control signals. Further research is needed to assess how smart grid control can be set up independently from existing communications infrastructures, or how it can be guaranteed as part of the minimal service.

Acknowledgments

The first author of this work is funded by the Agency for Innovation by Science and Technology in Flanders (IWT).

References

[1] S. Sorrell *et al*., "Global Oil Depletion: A Review of the Evidence," *Energy Policy*, vol. 38, no. 9, Sep. 2010, pp. 5290–95.

[2] M. Radetzki, "Peak Oil and Other Threatening Peaks — Chimeras without Substance," *Energy Policy*, vol. 38, no. 11, Nov. 2010, pp. 6566–69.

[3] C. Lutz, U. Lehr, and K. S. Wiebe, "Economic Effects of Peak Oil," *Energy Policy*, vol. 48, Sep. 2012, pp. 829–34.

[4] N. P. Myhrvold and K. Caldeira, "Greenhouse Gases, Climate Change and the Transition from Coal to Low-Carbon Electricity," *Environmental Research Lett.*, vol. 7, no. 1, Mar. 2012, p. 014019.

[5] W. Van Heddeghem *et al*., "Trends in Worldwide ICT Electricity Consumption from 2007 to 2012," *Computer Commun.*, vol. 50, Sep. 2014, pp. 64–76.

[6] B. Raghavan and J. Ma, "Networking in the Long Emergency," *Proc. ACM SIGCOMM Workshop on Green Networking*, 2011.

[7] C. Lange *et al*., "Energy Consumption of Telecommunication Networks and Related Improvement Options," *IEEE J. Sel. Topics Quantum Electron.*, vol. 17, no. 2, Mar. 2011, pp. 285–95.

[8] J. Baliga *et al*., "Energy Consumption in Wired and Wireless Access Networks," *IEEE Commun. Mag.*, vol. 49, no. 6, Jun. 2011, pp. 70–77.

[9] M. Deruyck *et al*., "Modelling and Optimization of Power Consumption in Wireless Access Networks," *Computer Commun.*, vol. 34, no. 17, Nov. 2011, pp. 2036–46.

[10] F. Idzikowski *et al*., "TREND in Energy-Aware Adaptive Routing Solutions," *IEEE Commun. Mag.*, vol. 51, no. 11, Nov. 2013, pp. 94–104.

[11] D. Kilper *et al*., "Energy Challenges in Current and Future Optical Transmission Networks," *Proc. IEEE*, vol. 100, no. 5, 2012, pp. 1168–87.

[12] W. Van Heddeghem *et al*., "A Power Consumption Sensitivity Analysis of Circuit-Switched Versus Packet-Switched Backbone Networks," *Computer Networks*, vol. 78, no. 0, 2015, pp. 42–56.

[13] C. Stewart and K. Shen, "Some Joules are More Precious than Others: Managing Renewable Energy in the Datacenter," *Proc. Workshop on Power Aware Computing and Systems*, 2009.

[14] M. Deruyck *et al*., "Reducing the Power Consumption in LTE-Advanced Wireless Access Networks by a Capacity Based Deployment Tool," *Radio Science*, vol. 49, no. 9, 2014, pp. 777–87.

[15] M. Gonzalez *et al*., "Concepts for Energy Efficient LTE Transceiver Systems in Macro Base Stations," *Future Network Mobile Summit* (FutureNetw), 2011, Jun. 2011, pp. 1–8.

Biographies

Sofie Lambert received her MSc degree in photonics in 2011 from Ghent University (Belgium), where she is currently pursuing a Ph.D. in Green ICT in the Department of Information Technology (INTEC). Her research interests include the worldwide electricity consumption in ICT infrastructure, the energy efficiency of various future optical and wireless access network architectures, and energy saving strategies such as sleep modes.

Margot Deruyck received the M.Sc. degree in computer science engineering and the Ph.D. degree from Ghent University, Ghent, Belgium, in 2009 and 2015, respectively. Her scientific work is focused on green wireless access networks with minimal power consumption and minimal exposure for human beings. This work led to the Ph.D. degree. Since January 2015 she has been a postdoctoral researcher at the iMinds/Ugent-INTEC (Ghent University-Department of Information Technology) and is continuing her work in green wireless access networks.

Ward Van Heddeghem received his master degree in applied engineering electromechanics from Hogeschool Gent, Belgium (1999), his M.Sc. degree in computer science engineering from Vrije Universiteit Brussel, Belgium (2009), and a Ph.D. degree in computer science engineering from Ghent University, Belgium (2014). He is the author of more than 20 internationally published papers. His research interests are in the field of the environmental impact of ICT and energy-efficient network architectures.

Bart Lannoo received an M.Sc. degree in electro-technical engineering and a Ph.D. degree from Ghent University (Belgium) in 2002 and 2008, respectively. His main research interests are in the field of fixed and wireless access networks, focusing on MAC protocols, Green ICT, and techno-economics. He is currently a postdoctoral researcher at Ghent University/iMinds and coordinates Green ICT research. He has been involved in various national and European research projects.

In a post-peak future, a smart grid that gathers information about suppliers and consumers and helps control them, will be an important instrument to balance electricity supplies and demands. This should be studied in a dedicated research track, considering incentives that shape consumer behavior and opportunities for automated power control.

WOUT JOSEPH (M'05, SM'12) obtained the Ph.D. degree in 2005. From 2007 to 2013 he was a postdoctoral fellow of FWO-V. Since October 2009 he has been a professor in the domain of experimental characterization of wireless communication systems. His professional interests are electromagnetic field exposure assessment, propagation and wireless performance analysis for wireless communication systems, antennas, calibration, and green networks.

DIDIER COLLE is a full professor at Ghent University. He received a Ph.D. degree in 2002 and a M.Sc. degree in electrotechnical engineering in 1997 from the same university. He is group leader in the iMinds Internet Technologies Department. He is co-responsible for the research cluster on network modelling, design and evaluation (NetMoDeL). This research cluster deals with fixed Internet architectures and optical networks, Green-ICT, design of network algorithms, and techno-economic studies.

MARIO PICKAVET is a full professor at Ghent University, where he is teaching courses on discrete mathematics and network modeling. He is co-leading the research cluster on network modeling, design and evaluation (NetMoDeL). His main research interests are Internet architectures and optical networks, Green ICT, and design of network algorithms. He is involved in several European and national research projects. He has published approximately 300 international publications in journals and in conference proceedings.

PIET DEMEESTER is a professor on the faculty of engineering at Ghent University. He is head of the research group "Internet Based Communication Networks and Services" (IBCN, Ghent University), and he leads the Internet Technologies Department of the strategic research centre iMinds. He is co-author of more than 1000 publications in international journals or conference proceedings. He is a Fellow of the IEEE.

A Survey of Green Scheduling Schemes for Homogeneous and Heterogeneous Cellular Networks

Ting Yang, Fabien Héliot, and Chuan Heng Foh

The authors are with the University of Surrey.

ABSTRACT

Energy efficiency is becoming an important feature for designing the next generation of communication networks, as are the multiplication of access points and the reduction of their coverage area. In this article we survey the latest development in energy-efficient scheduling, a.k.a. green scheduling, for both classic and heterogeneous cellular networks. We first introduce the main system model and framework that are considered in most of the existing green scheduling works. We then describe the main contributions on green scheduling as well as summarize their key findings. For instance, green scheduling schemes have demonstrated that they can significantly reduce transmit power and improve the energy efficiency of cellular systems. We also provide a performance analysis of some of the existing schemes in order to highlight some of the challenges that need to be addressed to make green scheduling more effective in heterogeneous networks. Indeed, the coordination between tiers and the rate fairness between the users of different tiers are important issues that have not yet been addressed. In addition, most existing designs exhibit a computational complexity that is too high for being deployed in a real system.

INTRODUCTION

The explosive growth of smart portable devices in recent years has triggered a shift from desktop to mobile Internet access. It is envisioned that the next generation of mobile networks will support higher area capacity and will improve coverage. One effective solution to do so is to increase cell density for higher spatial reuse. This solution will inevitably introduce additional power consumption, leading to a larger amount of greenhouse gas emission. With the information and communication technology (ICT) industry already contributing 2 percent of worldwide greenhouse emissions [1], seeking solutions to achieve energy efficiency (EE) in mobile networks has become a key consideration especially for the design of future mobile networks.

In a mobile network, base stations (BSs) account for as much as 80 percent of total power consumption [2]. Typically, macrocell BSs use high transmit power to maintain their large cell size (500 m inter site distance (ISD) for urban macrocell, 1732 m ISD for suburban macrocell in a Long Term Evolution (LTE) network [3]). While large cells suit low user density usage and help to reduce capital cost, they consume a large amount of energy to operate. Reducing cell size can help lower energy consumption, but this would require more cells to cover the same area, which, in turn, will increase capital cost and energy usage. The concept of heterogeneous networks (HetNets) is proposed to offer a flexible solution. In HetNets, small cells are deployed within an existing macrocell and operate either concurrently with the macrocell by sharing the same radio resource, or orthogonally by using a different frequency band. Deploying small cells within areas of high traffic load can help reduce the traffic load of the macrocell. Intelligent network deployment strategies, where a high density deployment of low-power small BSs is utilized, are believed to decrease power consumption compared with a low density deployment of high-power macro BSs, the idea being that a BS closer to mobile users requires less transmit power due to advantageous path loss conditions [4].

In a recent survey conducted by Feng *et al.* [4], it has been shown that EE transmissions offering lower power consumption are effective for saving a significant amount of energy in a mobile network. Defining EE to be the ratio of total rate to the total energy consumption of the network, the main objective of EE design is to maximize the value of EE measured in *bit/Joule*. The value of EE represents the amount of information bits that can be transmitted per every Joule of energy consumed by the network.

User scheduling and resource allocation have been found effective in the past for improving spectrum efficiency (SE) or peak rate performance of communication systems [2]. Given the

0163-6804/15/$25.00 © 2015 IEEE

The research on green scheduling generally exploits channel condition information to reduce transmit power while maintaining performance. Different techniques have been developed for different constraints and scenarios in the literature. We classify the existing research efforts by different constraints and then scenarios.

increasing relevance of EE for future communication networks, EE-based scheduling and resource allocation have recently attracted substantial research interest [2, 4, 5]. EE-based scheduling, which is also known as green scheduling (GS), aims at scheduling user transmission while reducing overall power consumption of the networks and guaranteeing an acceptable transmission rate. Traditionally, to achieve higher achievable rate, mobile networks operate at their maximum allowable power in order to achieve higher signal to interference and noise ratio (SINR). However, high transmit power does not necessarily lead to high network throughput; indeed, in a traditional multicell cellular system, higher transmit power will result in higher co-channel interference that can, counterproductively, reduce SINR. This highlights the importance of properly adjusting power according to the actual user requirements and environments, as recently demonstrated in [2, 4], in order to maintain a high transmission rate with lower transmit power. Whereas in a two-tier HetNet setup, given the different EE characteristics of small cells and macrocells, the importance of fine-tuning the power according to user requirements/environment will be even more critical for controlling both inter-tier and intra-tier interference. The HetNet setup provides additional flexibility in green scheduling to achieve further energy saving in transmissions, for example by performing joint scheduling between the two tiers to coordinate their interference and, in turn, improve EE. Some early investigations have already shown encouraging results of green scheduling in HetNets [6, 7].

However, we believe that the full potential of green scheduling to further improve EE in future communication networks, that is, dense small cell networks, has yet to be unlocked. Thus, we provide here a detailed survey of green scheduling schemes for mobile networks and emerging HetNets by identifying and classifying the current trends in green scheduling and summarizing the key results of existing works. Based on the latter, we first identify the future challenges for green scheduling, and then propose our thoughts and ideas for tackling them. The rest of the article is organized as follows. In the following section we present the system model and framework that are common to existing green scheduling research. Then we describe the various existing approaches in green scheduling. The key findings of these approaches are summarized and analyzed, and then both single-tier and HetNet scenarios are simulated as an illustration of green scheduling schemes' performance. Challenges and future directions in green scheduling design are then discussed, followed by important concluding remarks.

System Model

Green scheduling aims to achieve EE in addition to SE. In green scheduling, EE is often measured by the ratio of transmission rate to corresponding used power. Studies of green scheduling mostly consider a downlink orthogonal frequency division multiplexing access (OFDMA) network with a single BS or M coordinated BSs. Each BS accesses the shared spectrum of N subcarriers to serve K users within its radio range. Both BS and user equipment (UE) are equipped with a single antenna. Availability of perfect channel state information (CSI) is often assumed in the process of green scheduling.

Energy efficiency of the system, EE, defined as the ratio of total transmission rate to the total consumed power, can be expressed as [4]

$$\mathrm{EE} \triangleq \frac{R_{total}}{P_{total}} = \frac{\sum_{m=1}^{M} \sum_{k=1}^{K} R_{K(m)}}{\sum_{m=1}^{M} \sum_{n=1}^{N} \Delta_m p_m^n + P_{fix}}, \qquad (1)$$

where $R_{k(m)}$ represents the sum rate of user k served by BS m. Green scheduling aims to maximize the number of transmitted bits with every joule of energy consumption. Power consumption consists of two parts, the transmit power p_m^n and the operating power P_{fix}, and Δ_m accounts for the radio frequency (RF) dependent slope of BS m. In Eq. 1 the quantity P_{fix} captures the operating power consumption including circuit power, cooling system, power for backhaul communications, and others. The above formula is also valid for the small cell scenario with proper adjustments [4].

The objective of green scheduling is to find a particular user scheduling and power allocation such that EE is maximized, given some constraints on power, rate, quality of service (QoS), and so on. It can be remarked that in most existing works on green scheduling for multicell systems [2, 5–10], the CSI is assumed to be available at all BSs. This implicitly corresponds to a coordinated scenario, where BSs have enough backhaul capability to exchange this CSI. As for single-cell systems, CSI is only available within each individual cell for its scheduling decision-making [11, 12].

Green Scheduling Schemes

The research on green scheduling generally exploits channel condition information to reduce transmit power while maintaining performance. Different techniques have been developed for different constraints and scenarios in the literature. We classify the existing research efforts by different constraints and then scenarios, as depicted in Fig. 1. The first track of research works generally imposes a constraint on transmit power when designing techniques for green scheduling [5, 7]. To capture the minimum transmission rate requirement in the system, research works in the second track jointly consider power and rate constraints [2, 6, 8, 11, 12]. Apart from power and rate, the third track of research efforts focuses on fairness-aware green scheduling which aims to provide balanced rates among users [10, 13].

While the majority of the existing works consider the traditional macrocell network as their scenario, some recent works have started to focus on proposing green scheduling schemes for the emerging HetNet scenario. In the case of multicell macro networks, BSs can perform coordinated green scheduling, that is, BSs can coordinate their transmission to improve their EE. The scenario of HetNet is more complicated due to the involvement of two tiers. Ideally, interference among intra cells and between inter cells should be jointly considered when designing a green scheduling scheme, which can be interpreted as full coordination among two tiers. However, due to modeling complexity, many existing works in HetNet are only half-coordinated, that is, developing green scheduling schemes for the small cells only while taking into account interference from the macro tier in their system model. For each track of research and scenario, we present several important works in the following.

Power Constrained Green Scheduling

For cellular networks sharing the same frequency, maximum transmit power for each BS is set to controll interferenc to other cells. This maximum transmit power is often derived during the planning phase with the objective of providing necessary coverage without excessive interference to other neighboring cells. Traditionally, maximum transmit power is divided evenly on each subcarrier in OFDMA systems. However, when channel condition is known, the corresponding transmit power can be calculated during the process of EE optimization, such that maximum transmit power may in effect not necessarily be used. The sum of scheduled power on each subcarrier must not exceed the maximum transmit power for each BS. Given power constraint and CSI knowledge, power constrained green scheduling assigns resources over time to maximize the EE of the system.

With the constraint on maximum transmit power, this track of research focuses on cooperative scheduling of users among neighboring BSs such that downlink transmit power can be reduced without significantly sacrificing data transmission rate. In [5], Venturino *et al.* consider both maximum overall BS transmission power as well as maximum transmit power per subcarrier. In their work, the power on each subcarrier should not exceed the overall BS transmission power divided by the number of subcarriers. Given the power constraint on each subcarrier, the corresponding EE on individual subcarriers is summed up to interchangeably represent the EE of the entire system. In other words, EE given by Eq. 1 is replaced by the following Sum-EE expression:

$$\text{Sum-EE} \triangleq \sum_m \sum_n \frac{R^n_{k(m)}}{\frac{\Delta_m p^n_m}{N} + \frac{P_{fix}}{N}}. \qquad (2)$$

The Sum-EE expression, which adds up individual EE on subcarrier n of BS m, is an approximation of the EE in Eq. 1, and hence it does not always maximize EE. However, Sum-EE simplifies the introduction of weight on each subcarrier, which permits the study of differentiated services [5].

In a HetNet scenario, Zhang *et al*. consider maximizing EE in densely deployed femtocells with maximum transmit power per femtocell BS [7]. In their setup, macro and femto cells operate on different spectrum, and hence there is no interference between the two tiers. The research focuses on maximizing the EE of the femto tier. The technique used in the work is based on game theory. A distributed algorithm is developed to achieve the EE objective. In [9], Xiao *et al.* optimize the EE for HetNets by using Lagrangian dual decomposition, serial carrier, and power allocation. EE optimization is straightforwardly obtained from SE optimization with power constraints because fixed transmit/consumed power is considered in this work. The heterogeneity of BSs is used to further improve the EE of the network.

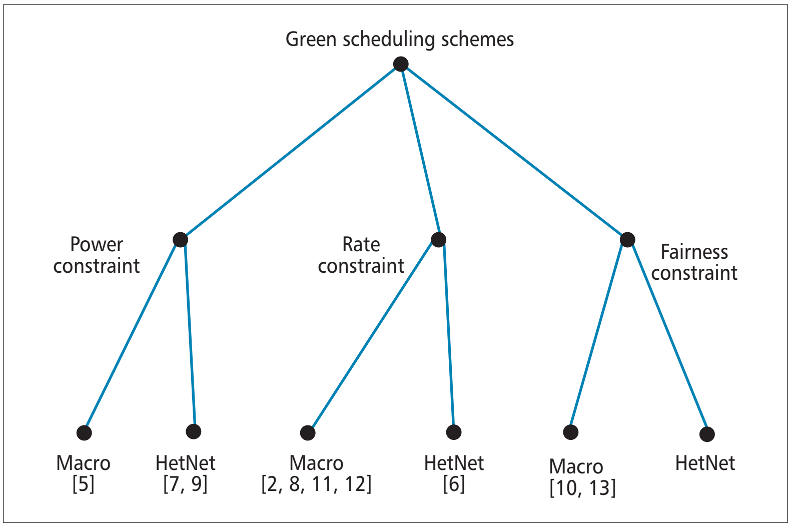


Figure 1. Existing green scheduling schemes.

Rate Constrained Green Scheduling

Many power constrained EE optimization solutions favor users with good channel quality. Their solutions often allocate fewer resources to those users with poor channel quality than to others, which affects the QoS that the system attempts to deliver. A trade-off between EE and SE exists [4]. In the literature, transmission rate is often the metric used to measure QoS. To deliver a minimum rate in the system, that is, to guarantee QoS, rate constraint is introduced to the EE optimization process. Given a scheduling period, the overall transmission of each BS must achieve a certain sum-rate and/or each user must be scheduled with a minimum transmission rate.

In the single-cell scenario, Xiao *et al*. [11] focus on providing an optimal scheduling solution to maximize EE while satisfying the rate requirements requested by the users. The problem is first transformed from a fractional form into an equivalent subtractive form, then Lagrangian duality is used. An approximation of the problem is then used to obtain the optimal subcarrier allocation for the relaxed problem. In order to meet the rate constraint of all the users, the power is next allocated by using a water-filling algorithm according to the obtained subcarrier allocation. For the same scenario, Zheng *et al.* [12] study EE when maintaining a minimum sum-rate for the BS as well as minimum individual user rates. This green scheduling scheme employs a water-filling method to find the best power allocation for each user. They propose a suboptimal algorithm for subcarrier allocation, which guarantees that each user is first allocated one subcarrier. Then, it assigns the rest of the

Many power constraint EE optimization solutions favor users with good channel quality. Their solutions often allocate less resources to those users with poor channel quality than others which affects the QoS that the system attempts to deliver. A trade-off between EE and SE exists.

Scheme	Property	Description
GS-Co	EE-based coordinated	Algorithm 5 of [5]. It has been specifically designed for the classic cellular layout; its generalized formulation makes it readily usable without modifications for the two-tier scenario.
GS-NC	EE-based non-coordinated	The coordinated green scheduler in [2], which has been specifically designed for the classic cellular layout, is utilized in a two-tier HetNet scenario but without cross tier coordination; in other words, each tier is coordinated independently, without being aware of cross-tier interference.
SE-Co	SE-based coordinated	Algorithm 3 of [14], which is a coordinated scheduler that can be readily used for both macro and small cells.
SE-Or	SE-based non-coordinated	Each BS has an equal number of dedicated subcarriers that are orthogonal to each other.

Table 1. Green scheduling and SE-based scheduling schemes.

subcarriers such that power consumption is minimized. Their simulation results show that the algorithm can achieve a better balance between EE and SE with reduced complexity, compared with Xiong's algorithm in [13], whereas in the multicell scenario, Heliot *et al.*'s scheme in [2] schedules a group of users that have similar CSI characteristics instead of individual users to achieve a low-complexity coordinated scheduling design. By comparing the EE performance of the non-coordinated and coordinated multicell approaches, they find that coordination helps improve the EE of cellular systems.

Limited backhaul capacity has a fundamental impact on scheduling schemes, especially for small cells. In [8] Derrick *et al.* factor in the limitation of backhaul capacity in each macrocell BS and constrain the maximum overall sum-rate that each BS can transmit for a scheduling period. In their work, the optimization problem is transformed into the same form as in [11] for deriving an efficient iterative green scheduling algorithm for the multicell scenario.

In the HetNet scenario, interference caused by another tier can influence the design of green scheduling. In [6] Jiang *et al.* propose a green scheduling scheme to maximize EE and satisfy the minimum rate requested by each user of the small cell in a single-cell HetNet, when taking into account the cross-tier interference from the macro-tier. The interference from the macro-tier may vary according to different CSI. In the meantime, an interference threshold is set for the small-cell tier in order to guarantee the QoS of the macrocell users. The optimal solution to the power and resource allocation is obtained by first transforming the EE problem into a subtractive form as in [8, 11], and then utilizing the dual decomposition method.

Fairness Constrained Green Scheduling

Introducing rate constraint in the EE optimization process has provided a basic means to deliver a certain service to BS or users. Since the minimum rate achievable by the system is not yet known during the EE optimization process, it is difficult to set a practical rate constraint for the optimization. Rate fairness that manages relative rates among users with predefined weights offers a practical setting in the EE optimization process.

In [13] Xiong *et al.* formulate EE optimization with predetermined weights to allow differentiated rate allocation among users. Their results show that with an appropriate setting of weights for users, fairness can be achieved, especially for users with low channel-gain-to-noise ratio (CNR). However, this work did not explicitly provide a method to appropriately set the weights that can achieve fairness. In order to support fairness, Ren *et al.* [10] include proportional rate constraint in the EE optimization framework, and design a low complexity algorithm for solving it. The results show that system performances are better when users with higher CNR are scheduled more often, which in turn reflects that the proportional rate constraints influence the fairness between users. Both [10] and [13] demonstrate that without fairness consideration, the system would favor high CNR users.

Performance of Green Scheduling Schemes

In the following we first summarize the key findings of our survey before providing results demonstrating the potential of full coordination for green scheduling.

Key Findings in Existing Green Scheduling Schemes

Green scheduling schemes have been shown to reduce transmit power (e.g. by 90 percent in [2]) and improve the EE of the system (e.g. by 94.2 percent in [9]) compared with traditional scheduling [2, 9–13]. Optimal solutions to EE optimization have been developed for a single-cell scenario. However, these solutions are computationally complex [10, 12, 13]. Additional research efforts have been made to develop low complexity sub-optimal green scheduling with some successes [10–13]. However, it is found that achieving EE targets introduces unbalanced fairness between different groups of users [11, 12]. In particular, users with poor channel condition transmit with unfairly low rates. Additional research efforts have been made to develop solutions for EE with QoS consideration.

For multicell networks where multiple cells share the same frequency band, it has been found that coordination among neighboring BSs can further improve EE performance [2]. However, due to inter-cell interference, finding optimal solutions requires computational complexity algorithms [5]. Suboptimal solutions can be obtained by using various approaches, such as relying on the symmetry of user locations [2] to reduce complexity.

Compared with classic single networks, HetNets have shown potential for significant EE improvement [9]. However, dedicated green scheduling schemes with cross-tier coordination have yet to be designed. Current green scheduling schemes considering HetNet are limited to a single-tier coordination [7, 10].

Case Study of Green Scheduling Schemes

Green scheduling techniques have been shown to be effective for saving energy in a single-tier mobile network, as reported in [2, 5, 8, 11, 13] for a classic cellular layout, or in [6, 7] for a small-cell only layout. However, the effectiveness of these techniques in a two-tier HetNet scenario remains to be characterized. In order to examine the effectiveness of green scheduling in a two-tier HetNet scenario, and identify the limitations of current existing schemes, we apply two of the various green scheduling schemes (i.e. "GS-Co" and "GS-NC") presented in this article. The two green scheduling schemes are chosen such that the impact of cross-tier coordination can be observed. We also apply two traditional SE-based schedulers (i.e. "SE-Co" and "SE-Or") as benchmarks to see how EE optimization compares with traditional scheduling schemes. Further details about the four compared schedulers are given in Table 1. We first consider a three-sector macro-only layout and then the two-tier HetNet layout as described in [3], where sectorized macro BSs and uniformly distributed small BSs coexists in the orange dodecagonal area, as depicted in Fig. 2. Our results are obtained through Monte-Carlo simulations by using MATLAB. Moreover, we have considered the power/system parameters in Table I of [2] and Tables 27 and 32 of [3] for plotting Figs. 3, 4, and 5, with N = 600 subcarriers and K = 20 uniformly distributed users in each sector of the dodecagonal area in Fig. 2. Regarding the power parameters for the small cells, we have used 0.13 Watt and 6.8 Watt for the maximum transmit power and circuit power, respectively, and set the RF dependent slope to 4.

In Figs. 3 and 4, we compare the EE performance of the four schemes with various ISD values for macro-only and HetNet with three randomly distributed small cells, respectively. For the macro-only scenario in Fig. 3, EE-based schedulers outperform SE-based schedulers. The "SE-Or" scheduler uses full power for transmission, and though there is no interference between different sectors, the available frequency bands for each sector was reduced by a third compared with other schedulers. It therefore has the worst EE performance [2]. The "SE-Co" scheduler can take advantage of available CSI to adjust its transmit power rather than transmitting at full power. However, since SE-based schedulers are not designed to improve EE, their performance is worse than "GS-NC" and "GS-Co" schedulers in terms of EE. The "GS-Co" scheduler has the best EE performance given that it uses CSI knowledge to optimize EE by enforcing coordination among BSs.

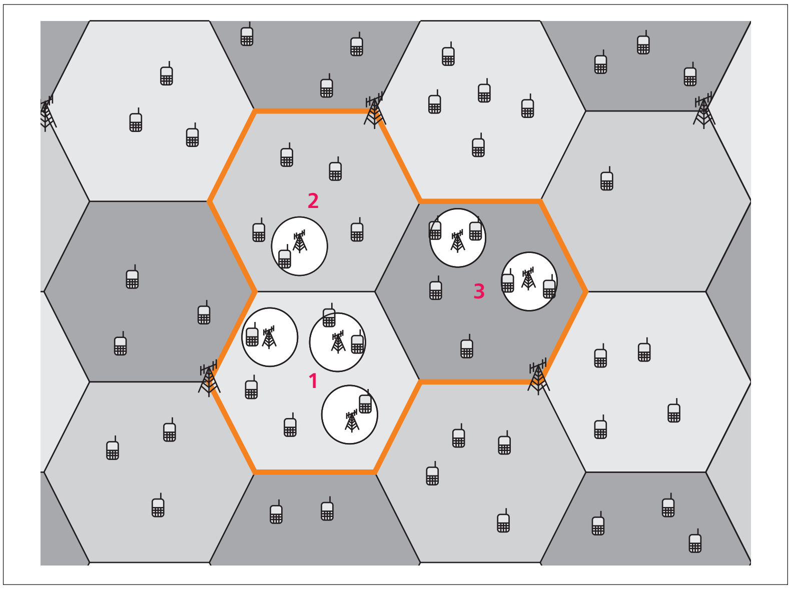


Figure 2. Sectorized planar cellular system layout.

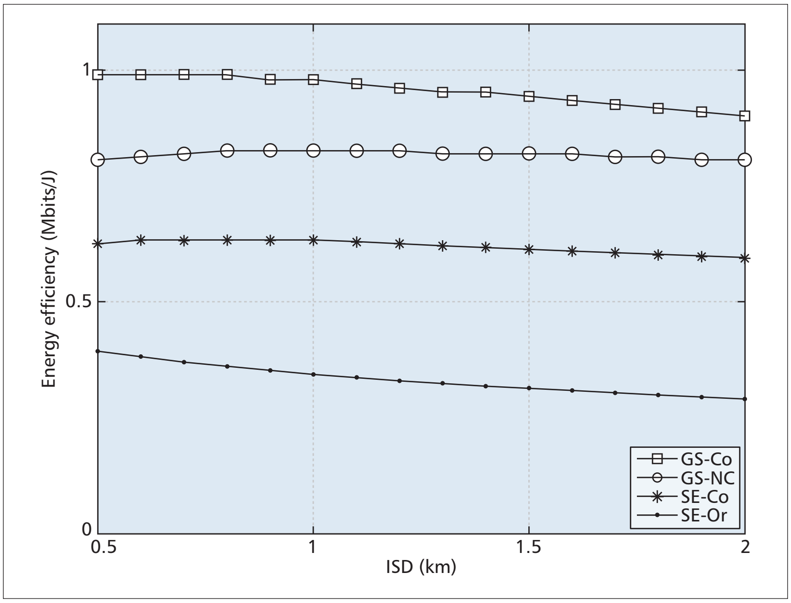


Figure 3. EE performance comparison of various scheduling methods vs. the inter-site distance in one-tier macro-only layout.

Figure 4 shows the EE performance of the four schemes and the fairness rate distribution between tiers in HetNet, in the upper and lower parts of Fig. 4, respectively. We observe that the "GS-Co" scheduler has the best EE performance. Interestingly, it can be remarked that SE coordination outperforms EE non-coordination in the HetNet scenario. The macro tier is not aware of the small cells for the "GS-NC" scheme, the transmit power of the macro BSs will increase with the ISD, creating more interference to the small cells. In turn, small cells will increase their transmit power in order to combat the interference. Such behaviors consume more power and result in rate degradation due to both increased intra-tier and inter-tier interference. In contrast, the "SE-Co" scheduler jointly considers the transit powers for both tiers such that interference can be minimized. This observation illustrates the importance of coordination among different tiers to improve EE. Our results confirm the benefit of coordination and sharing CSI between BSs, as discussed in [15].

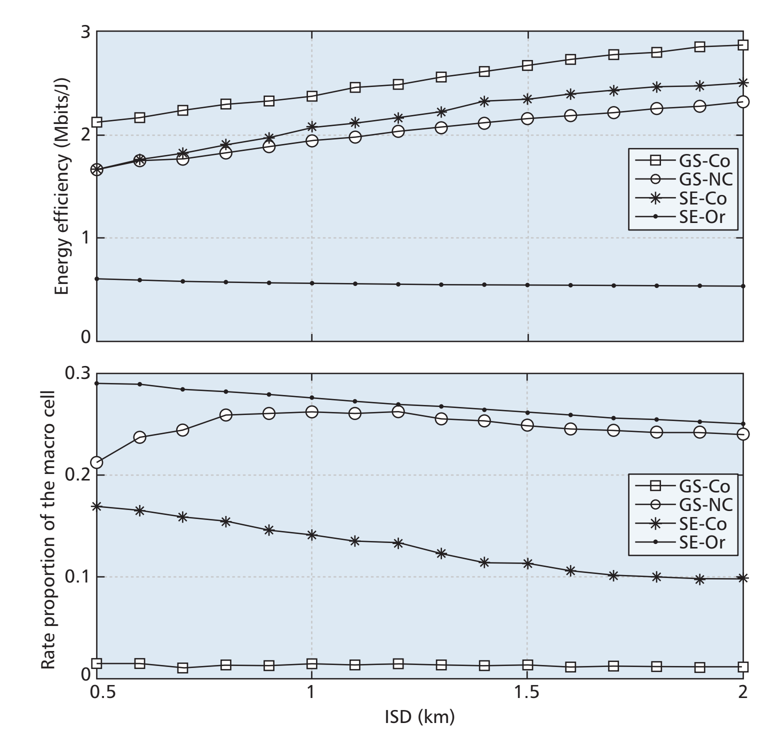


Figure 4. Performance comparison of various scheduling methods in terms of EE and rate fairness vs. the inter-site distance in two-tier HetNet layout.

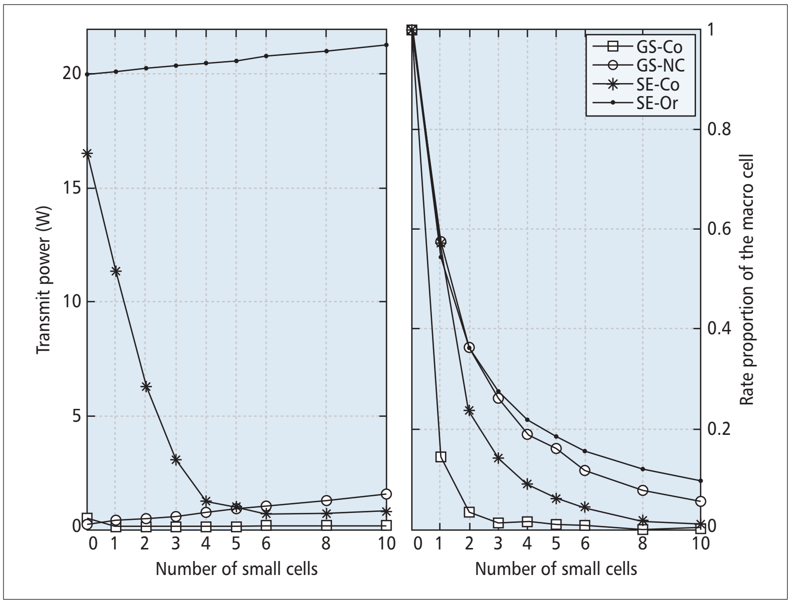


Figure 5. Performance comparison of various scheduling methods in terms of transmit power and rate fairness vs. number of small cells.

When comparing the results for the macro and HetNet scenarios in Figs. 3 and 4, it can be remarked that green schedulers provide an EE improvement of at least 100 percent in the HetNet scenario when compared to the macro-only scenario. This is because small cells in HetNet use lower transmission power to communicate with UEs while achieving a similar rate as users served by the macrocells. This can be observed in Eq. 1 where adding similar rate elements in the numerator while reducing the level of the power in the denominator contributes to better EE performance. In addition, as seen in Fig. 5, adding small cells has the effect of lowering the transmission power of coordinated schemes, and hence this reduces interference, which in turn improves the rate and EE.

The lower part of Fig. 4 compares the rate proportion of the macrocell for the four schemes, which is an indicator of how fair the rate distribution is between the two tiers. It can be observed that some of the existing macro-only green scheduling solutions [2, 5], when directly applied to the HetNet scenario, generate unfair load distribution between the two tiers, mainly due to the different power and propagation characteristics of BSs in different tiers. Indeed, the results presented in the lower part of Fig. 4 indicate that "GS-Co" achieves the best EE performance of the system by over-favoring the small cells at the expense of the macrocell rate. The same unfair load balancing problem can also be noticed in SE coordination, although the impact is not as severe. On the other hand, both "GS-NC" and "SE-Or" can maintain a fairer load distribution between tiers as they do not perform cross-tier coordination. In summary, this result clearly shows the existence of a trade-off between EE and load balancing in a multi-tier system. Even though some existing macro- only green schedulers can be used to improve EE in HetNet, they still suffer from the unfair rate allocation problem. This indicates that green scheduling with cross-tier coordination requires additional attention. Even more so, when mentioning that the existing green coordinated scheduling schemes do not take into account the on/off switching capabilities of small cells, this certainly represents an extra degree of freedom to improve EE, but also an extra challenge to make it work effectively.

Figure 5 illustrates the corresponding transmit power of the various schedulers as a function of the number of small cells within the dodecagonal area illustrated in Fig. 2. For non-coordinated scheduling schemes, that is, "SE-Or" and "GS-NC," the schedulers of each sector are not aware of other sectors. While for "SE-Or," BSs always transmit at full power, for "GS-NC," BSs obtain their transmit powers for each sector without considering the possible co-channel interference they cause to impact other sectors. Therefore, for the non-coordinated schedulers, the transmit power increases with the number of small cells to combat interference. As for the coordinated scheduling schemes such as "SE-Co" and "GS-Co," the schedulers oversee the entire network to manage rather than combat interference. In this case, the transmit power of one sector will be adjusted according to the power from others. When more small cells are present in the network, more interference occurs, and the transmit power is automatically reduced accordingly to meet the EE or SE performance.

Conclusion

This survey has shown that efforts have been made to address the EE problem in both single-tier macro networks and two-tier HetNets. With the presented results, we conclude that coordination between the BSs can achieve better EE. Apart from green scheduling in single-cell scenario, there have been recent works focusing on coordinated multicell scenarios [2, 5, 8]. The main challenge in the multicell scenario is the complexity while dealing with co-channel interference, such that finding an optimal green scheduling solution is likely to be very challenging, if not impossible. In the HetNet scenario, most research into green scheduling still focuses solely on the small-cell tier; as such, cross-tier coordination represents a worthwhile track of research. Our exploration of green scheduling has unveiled several interesting directions to be further researched:

- Developing a green scheduler with coordinated macro and small BSs.
- Addressing the unfairness problem among different tiers.
- Exploring other user selection/grouping schemes apart from greedy user selection, such as a graph-based approach.
- Reducing the complexity of scheduling algorithms. This is an important issue, especially for densely deployed networks.

In the HetNet scenario, most of the green scheduling investigations still focus solely on the small-cell tier; as such cross-tier coordination represents a worthwhile track of research. Our exploration of green scheduling has unveiled several interesting directions to be further researched.

References

[1] G. Fettweis and E. Zimmermann, "ICT Energy Consumption: Trends and Challenges," *Proc. 11th Intl. Symp. WPMC*, 2008.

[2] F. Héliot, M. Imran, and R. Tafazolli, "Low-Complexity Energy-Efficient Resource Allocation for the Downlink of Cellular Systems," *IEEE Trans. Commun.*, vol. 61, no. 6, 2013, pp. 2271–81.

[3] A. Ambrosy *et al.*, "D2. 2: Definition and Parameterization of Reference Systems and Scenarios," INFSOICT-247733 EARTH (Energy Aware Radio and NeTwork TecHnologies) tech. rep., 2010.

[4] D. Feng *et al.*, "A Survey of Energy-Efficient Wireless Communications," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 1, 2013, pp. 167–78.

[5] L. Venturino *et al.*, "Energy-Efficient Scheduling and Power Allocation in Downlink OFDMA Networks with Base Station Coordination," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, 2015, pp. 1–14.

[6] J. Jiang *et al.*, "Energy-Efficient Resource Allocation in Heterogeneous Network with Cross-Tier Interference Constraint," *Proc. IEEE 24th Int'l. Symp. PIMRC Workshops*, 2013.

[7] Z. Zhang *et al.*, "Low Complexity Energy-Efficient Resource Allocation in Down-Link Dense Femtocell Networks," *Proc. IEEE 24th Int'l. Symp. PIMRC*, 2013.

[8] D. Ng, E. Lo, and R. Schober, "Energy-Efficient Resource Allocation in Multi-Cell OFDMA Systems with Limited Backhaul Capacity," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, 2012, pp. 3618–31.

[9] X. Xiao *et al.*, "An Energy-Efficient Hybrid Structure with Resource Allocation in OFDMA Networks," *Proc. IEEE WCNC*, 2011.

[10] Z. Ren *et al.*, "Energy-Efficient Resource Allocation in Downlink OFDM Wireless Systems with Proportional Rate Constraints," *IEEE Trans. Vehic. Tech.*, vol. 63, no. 5, 2014, pp. 2139–50.

[11] X. Xiao, X. Tao, and J. Lu, "QoS-Aware Energy-Efficient Radio Resource Scheduling in Multi-User OFDMA Systems," *IEEE Commun. Lett.*, vol. 17, no. 1, 2013, pp. 75–78.

[12] Z. Zheng *et al.*, "Energy-Efficient Resource Allocation for Downlink OFDMA Systems," *Proc. IEEE ICC Workshops*, 2013.

[13] C. Xiong *et al.*, "Energy-Efficient Resource Allocation in OFDMA Networks," *IEEE Trans. Commun.*, vol. 60, no. 12, 2012, pp. 3767–78.

[14] L. Venturino, N. Prasad, and X. Wang, "Coordinated Scheduling and Power Allocation in Downlink Multicell OFDMA Networks," *IEEE Trans. Vehic. Tech.*, vol. 58, no. 6, 2009, pp. 2835–48.

[15] D. Gesbert *et al.*, "Multi-Cell MIMO Cooperative Networks: A New Look at Interference," *IEEE JSAC*, vol. 28, no. 9, 2010, pp. 1380–1408.

Biographies

Ting Yang received her M.Sc. degree in mobile communications systems from the University of Surrey, United Kingdom in 2013. She is currently working toward the Ph.D. degree at the Institute for Communication Systems (ICS), University of Surrey. Her research interests focus on energy efficiency, user scheduling, and radio resource management, especially for the densely deployed small cells.

Fabien Héliot [S'05, M'07] received the Ph.D. degree in mobile telecommunications from King's College London in 2006. He is currently a lecturer at the Institute for Communication Systems (ICS) at the University of Surrey. He has been actively involved in European Commission funded projects such as FIREWORKS, ROCKET, SMART-Net, EARTH, and LEXNET. His main research interests are EMF exposure, energy efficiency, cooperative communication, MIMO, and radio resource management.

Chuan Heng Foh received the M.Sc. degree from Monash University, Australia in 1999, and Ph.D. degree from the University of Melbourne, Australia in 2002. After his Ph.D., he spent six months as a lecturer at Monash University in Australia. In December 2002 he joined Nanyang Technological University, Singapore, as an assistant professor, a position he held until 2012. He is now a senior lecturer at the University of Surrey. His research interests include protocol design and performance analysis of various computer networks, including wireless local area and mesh networks, mobile ad hoc and sensor networks, 5G networks, and data center networks.

Assessing Network Energy Consumption of Mobile Applications

Chien Aun Chan, Wenwen Li, Sen Bian, Chih-Lin I, André F. Gygax, Christopher Leckie, Ming Yan, and Kerry Hinton

Chien Aun Chan, André F. Gygax, Christopher Leckie, Ming Yan, and Kerry Hinton are with The University of Melbourne.

Wenwen Li, Sen Bian, and Chih-Lin I are with China Mobile Research Institute.

ABSTRACT

Continuous growth in the energy consumption of mobile networks has become a major concern for mobile carriers. Since current mobile networks are dominated by mobile data traffic generated by over-the-top mobile applications, it is crucial for mobile carriers to understand how much network energy is used to deliver these applications. Here, we use real network and application measurements to comprehensively analyze the energy consumption of 12 common mobile applications by breaking down their total energy consumption into data and signaling energy components. The results provide insights into the different proportions of data and signaling energy (due to LTE signaling) for different mobile applications. They show that the energy consumption of a mobile application can vary at different base station cell sites due to different ratios of throughput to physical resource block utilization. We estimate the total monthly energy consumption of all 4G users of a major mobile carrier using conventional mobile services, such as voice and the short messaging service, and two over-the-top applications, i.e. a popular instant messaging application in China and an online video application. The results show that signaling energy consumption may become a major concern for mobile carriers, and that this issue will be exacerbated as the usage of over-the-top applications continues to grow.

Energy assessment of mobile applications provides valuable information to enable mobile carriers to improve the energy efficiency of their networks. An energy assessment tool that enables real-time network and service energy monitoring will also assist in developing energy-efficient network policies for diverse applications with different energy consumption profiles. Furthermore, given our signaling energy consumption findings for over-the-top applications, there may be benefits for mobile operators to introduce a surcharge for signaling traffic to mitigate the high signaling energy consumption of some over-the-top mobile applications.

INTRODUCTION

Global mobile data traffic has increased from 0.2 exabytes per month in 2010 to 2.5 exabytes per month in 2014, with a staggering growth of 1,150 percent over four years. Rapid growth in mobile data traffic is expected to continue into the future. As estimated by [1], global mobile traffic will increase by a factor of 10 between 2014 and 2020. Rapid growth in global mobile data traffic has been driven by the increase in the usage of smartphone applications such as social networking applications, gaming, and online video applications [1]. These mobile applications use mobile networks to deliver over-the-top (OTT) content to the end-users, and therefore are known as OTT applications.

With hundreds of millions of smartphone users of these OTT mobile applications, vast amounts of OTT content are generated in mobile access networks every day. For example, Deloitte estimated that in 2014 the total number of messages generated by OTT instant messaging (IM) applications surpassed 50 billion a day compared to 21 billion conventional short messaging service (SMS) messages [2]. The increase in popularity of OTT applications and the rapid growth of mobile data traffic are driving the massive deployment of mobile base station (BS) cell sites globally, and hence rapidly increasing the total energy consumption of mobile networks. For example, the number of BS sites operated by China Mobile has increased from 400,000 in 2005 to 1.5 million in 2014, which includes 700,000 time-division duplex long-term evolution (TDD-LTE) BSs [3]. As a result, the total energy consumption of China Mobile increased significantly from 2,000 GWh in 2005 to 17,110 GWh in 2014 [3].

Mobile BS cell sites consumed more than 60 percent of a company's total energy consumption for most mobile carriers [4, 5], therefore current research has a strong focus on improving the energy efficiency of mobile access networks [6-8]. The authors of [6] first established a foundational energy model to estimate the energy consumption of wireless networks. That network energy model was then widely used to quantify the energy efficiency of various LTE networks.

0163-6804/15/$25.00 © 2015 IEEE

The authors of [7] and [8] presented comprehensive tutorials and surveys on techniques and solutions to reduce mobile networks' energy consumption and improve the energy efficiency of mobile wireless networks.

Most existing research is focused on the energy consumption of mobile access networks as a whole. There is little existing work that focusses on the energy consumption and energy efficiency of mobile *applications*. The characteristics of OTT applications in terms of data traffic and signaling traffic overhead (due to LTE signaling) were first analyzed by [9]. However, determining the energy consumption of OTT applications with diverse characteristics is challenging due to the lack of an energy model that integrates real measurements from both the mobile access networks and OTT applications to quantify the energy consumed by the networks in delivering these applications. Since mobile data traffic has become dominated by OTT applications, it is important for mobile carriers to quantify the energy cost of delivering these applications via their mobile networks.

The remainder of this article is organized as follows. We first discuss a methodology to assess the energy consumption of various mobile applications in a TDD-LTE radio access network (RAN). We present the measurements required to assess the total energy consumption of mobile applications: power measurements of TDD-LTE BSs, throughput measurements from these TDD-LTE cell sites, and test results of 12 commonly used mobile applications. Next, we present the energy consumption breakdown of these 12 mobile applications by separating the total energy consumption of the application into downlink data energy, uplink data energy, and signaling energy (due to LTE signaling). The results show how the energy consumption of a mobile application can change when used in different cell sites. Then we discuss the significance of signaling energy overhead for OTT applications by providing an estimate of the total monthly signaling energy consumed by all 4G users of China Mobile at the end of 2013. The results show that signaling energy accounted for around 12 percent of the total energy consumption of conventional mobile services such as voice and short messaging services, but for the OTT instant messaging (IM) application, the signaling energy could account for over 90 percent of the total application energy consumption. Hence, we discuss several potential strategies that can be used to reduce the overall network energy consumption for diverse mobile applications with different energy consumption profiles. Finally, we provide the conclusion of the article.

Energy Assessment of Mobile Applications

Mobile BS cell sites currently consume 60 percent to 70 percent of mobile carriers' total corporate energy usage [4, 5]. This highlights the need for mobile carriers to ascertain the energy consumed by OTT applications delivered via their mobile access networks. Since mobile data traffic is dominated by OTT applications, energy assessment of mobile applications and services will be very valuable for mobile carriers in the following areas.

Network Design and Planning: Mobile access networks are often heterogeneous and require coordination between macro cells and small cells. Knowing the customer base and the popularity of the mobile applications, a mobile carrier could predict the network capacity requirements and the energy efficiency of delivering these mobile applications before the deployment of new cell sites. For example, assessing the energy consumption of mobile applications could help mobile carriers to evaluate whether it is energy-efficient to offload certain mobile applications to WiFi and small cells. Mobile carriers could also evaluate the energy efficiency of delivering specific mobile applications using new techniques such as the decoupling of signaling traffic and data traffic, and the decoupling of uplink and downlink data transmissions.

Energy-Efficient Service Provisioning: The typical duration over which a new mobile application gains popularity is extremely short. For example, WeChat had only 50 million users by the end of 2011, but the number of active users had grown to more than 450 million in 2014 [10]. Energy assessments of mobile applications could assist the mobile carrier to understand the energy impact of new services or applications that are rapidly gaining popularity in their networks. Hence, mobile operators could develop appropriate network policies to optimize the energy and network resources used by those services and applications.

Costs and Pricing: Different mobile applications generate different amounts of signaling and data traffic and therefore consume different proportions of data and signaling energy. For example, our results show that OTT IM applications consume a significant amount of signaling energy compared to the data energy. However, for OTT video applications, the data energy consumption is more than 97 percent of the total application energy. As energy prices increase and the acceptance of carbon accounting grows, there is a need to include these factors in the pricing of services. Therefore, assessing the energy consumption of mobile services and applications is expected to help mobile carriers to develop pricing schemes for OTT service providers that include these considerations.

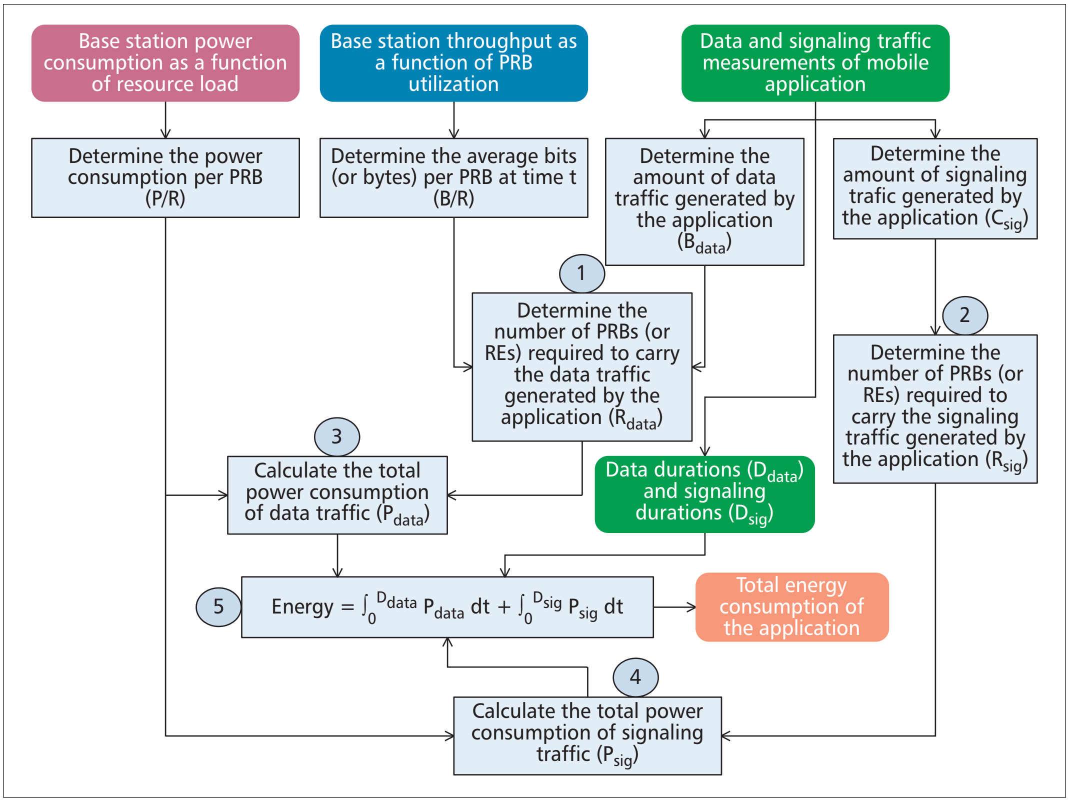
Real time Network Monitoring and Assessment: Every cell site is different due to different user behaviors, different coverage, different numbers of users, and different geographical areas. As demonstrated by the results in this article, using the same mobile application at a different cell site will incur different energy consumption levels. Therefore, the energy assessment methods for mobile applications and services presented in this article will enable real-time monitoring of energy consumption in wireless networks. This will be particularly useful in heterogeneous wireless networks, as a tool for improving the energy efficiency of resource allocation strategies.

Assessing the energy consumption of a mobile application requires measurements on:

- BS power consumption.
- BS throughput as a function of physical resource block (PRB) utilization.
- The data and signaling traffic generated by using the mobile application.

As energy prices increase and the acceptance of carbon accounting grows, there is a need to include these factors in the pricing of services. Therefore, assessing the energy consumption of mobile services and applications is expected to help mobile carriers to develop pricing schemes for OTT service providers that include these considerations.

At a low CQI, a lower modulation format and lower coding rate are used. As a result, the number of PRBs used to carry the information generated by a mobile application depends on the quality of the link and quality of service between the mobile device and the BS cell site, which is time and cell site dependent.



Base station power consumption as a function of resource load
Base station throughput as a function of PRB utilization
Data and signaling traffic measurements of mobile application
Determine the power consumption per PRB (P/R)
Determine the average bits (or bytes) per PRB at time t (B/R)
Determine the amount of data traffic generated by the application (B_{data})
Determine the amount of signaling trafic generated by the application (C_{sig})
1
Determine the number of PRBs (or REs) required to carry the data traffic generated by the application (R_{data})
2
Determine the number of PRBs (or REs) required to carry the signaling traffic generated by the application (R_{sig})
3
Calculate the total power consumption of data traffic (P_{data})
Data durations (D_{data}) and signaling durations (D_{sig})
5
Energy = $\int_0^{D_{data}} P_{data}\,dt + \int_0^{D_{sig}} P_{sig}\,dt$
Total energy consumption of the application
4
Calculate the total power consumption of signaling traffic (P_{sig})

Figure 1. The process of assessing the energy consumption of delivering a mobile application through a RAN.

Figure 1 shows the process of assessing the network energy consumption of a mobile application. The energy assessment methods could also be used for frequency division duplex (FDD)-LTE systems. However, it should be noted that the PRB allocations of FDD-LTE systems for uplink data, downlink data, and signaling channels will be different compared to TDD-LTE systems. Furthermore, the data and signaling traffic measurements of mobile applications could also be different in FDD-LTE networks. Therefore, one should follow the methods shown in this article to conduct similar studies using FDD-LTE systems. In addition, the energy assessment methods shown in Fig. 1 are designed to assess the cumulative energy consumed by a TDD-LTE RAN in serving one or more mobile applications to a mobile device as follows:

1) First, we determine how much BS input power is consumed by one PRB (which consists of 12 subcarriers and seven symbols in the LTE resource grid). However, such information may not be retrievable from current network measurement tools. Therefore, as an alternative, the input power of the BS is measured as a function of output radiation power in an offline calibration. Since there is a linear relationship between BS output power and resource utilization, as observed in a practical LTE system [6], the normalized output power of the BS (%) can be approximated to the normalized resource utilization (%).

2) Second, the BS's actual throughput as a function of PRB utilization information will help us to determine the average number of information bits that can be carried by one PRB at different times of the day at different cell sites. If the average signal-to-interference-noise ratio (SINR) of the cell site is poor at time t, it is expected that more PRBs are required to carry the same amount of data compared to the time when the average SINR is good.

3) Third, data and signaling traffic measurements need to be made for the mobile application. These measurements provide information on how much data and signaling traffic are generated by the application to perform a certain task (e.g. to send a text message or a picture). It should be noted that data traffic is generally measured in bits and the usage duration is in seconds. However, signaling traffic is measured in terms of the number of control channel elements (CCE). One CCE equals 36 resource elements (REs) and one RE represents one symbol and one subcarrier in the LTE resource grid. Since LTE physical channels for signaling use the same modulation format (i.e. quadrature phase shift keying) to transmit all signaling messages, one CCE also equals 72 bits (or 9 bytes).

By combining the information from 2 and 3, the number of PRBs required to carry the data and signaling traffic generated by the mobile application can be determined (indicated by "1" and "2" in Fig. 1). Then, given the information from 1, the data and signaling power consumption of the mobile application can be determined (indicated by "3" and "4" in Fig. 1). Finally, the total energy consumption is the sum of the integrals of the data and signaling power consumption over the duration of use (indicated as "5" in Fig. 1).

Power Consumption of TDD-LTE Base Station

The equipment power consumption of a TDD-LTE BS has been measured as shown in Fig. 2. As depicted in Fig. 2, the power consumption of the baseband unit (BBU) and the remote radio unit (RRU) are measured as a function of output radiation power (P_{out}). In the downlink, the BS load defined by $P_{\text{out}}/P_{\max}$ (in which $P_{\max}$ is the maximum output radiation power), is proportional to the level of utilized resources, comprising both data and control signals [6]. As a result, the normalized P_{out} in Fig. 2 is linearly proportional to the utilization of radio resources in a TDD-LTE BS cell site. Assuming that the normalized P_{out} in (%) equals the normalized downlink BS resource utilization in (%) [6], we can then determine the BS power consumption per PRB or resource element (RE). It should be noted that the number of REs for TDD-LTE depends on the sub-frame configuration according to 3GPP specifications.

The BS equipment power profile shown in Fig. 2 consists of a baseline power component (total power of 356 W) and a variable power component (total power of 356 W to 715 W). The baseline power of the BS equipment corresponds to the condition where the BS has no attached users and no service traffic. In this situation, small amounts of radio resources are used to transmit fixed signaling such as the cell reference signal (CRS), synchronization signal (PSS and SSS), and physical broadcast channel (PBCH). The variable power consumption of the BS equipment corresponds to the increase in radio resource utilization and hence an increase in the power consumption of the BS. As shown in Fig. 2, since the variation in input power of the BBU is almost negligible when the radio resource utilization varies, the entire variation of the BS's input power is primarily determined by the power consumption of the power amplifier in the RRU.

Base Station's Throughput as a Function of PRB Utilization

In an LTE system, a total of three different modulation formats can be used: quadrature phase shift keying (QPSK), 16 quadrature amplitude modulation (16 QAM), and 64 QAM, which can carry two bits, four bits, and six bits of information per RE, respectively. However, an LTE system also uses channel coding on top of the chosen modulation format depending on the link channel quality. The combination of different modulation formats and different channel coding rates is determined by the channel quality indication (CQI) indexes. Each user will have a specific CQI index when connecting to the BS. At a low CQI (which indicates low channel quality or low SINR), a lower modulation format (e.g. QPSK) and lower coding rate are used. As a result, the number of PRBs used to carry the information generated by a mobile application depends on the quality of the link and quality of service between the mobile device and the BS cell site, which is time and cell site dependent.

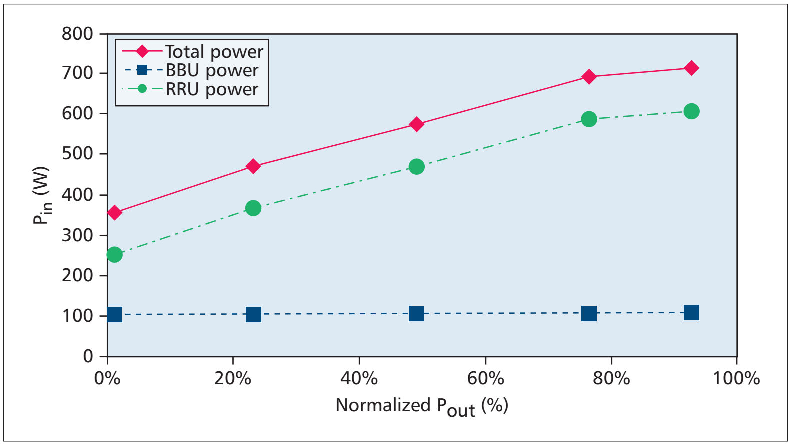


Figure 2. Total equipment power consumption of a three-sector TDD-LTE BS. The configuration of each cell sector is 2 × 2 multi-input multi-output (MIMO) and the bandwidth per sector is 20 MHz. The maximum output power of each transceiver is 40W. The power measurements were conducted in-house (off-line calibration).

The average number of information bits per RE at a specific time of day can be determined by analyzing the BS's information log. Figure 3 shows the actual throughput and the PRB utilization as a function of the time of day for three different cell sites. The ratio of actual throughput to PRB utilization gives the estimated maximum throughput, which is the average maximum bit/s that can be supported by the cell site at different times of the day. It should be noted that all three BS cell sites have the same configurations for fair comparison: TDD-LTE with 20 MHz bandwidth, 2 x 2 MIMO, downlink-uplink sub-frame configuration 2, and special sub-frame configuration 5.

Comparing cell sites A, B, and C in Fig. 3, cell sites A and B had higher actual throughput throughout the day compared to cell site C, but both had much lower estimated maximum throughput. This is due to several reasons:

- Different numbers of users per cell site.
- Different geographical areas (cell site A is in an urban area surrounded by university campuses and colleges, while cell sites B and C are in dense urban business districts).
- Different population densities (the average population density of cell sites A, B, and C are 4800/km^2, 6375/km^2, and 25470/km^2, respectively).
- Different user behaviors.
- Different popularities of mobile applications. As a result, different mobile BS cell sites consumed different amounts of energy to deliver the same mobile application to each user they serve.

Data and Signaling Traffic of Mobile Applications

The total energy consumption of a mobile application depends on the amount of data and signaling traffic generated by the application. Here we develop test cases for 12 different popular mobile applications to collect the downlink and uplink data throughput and the signaling traffic generated by each mobile application.

SMS and the multimedia messaging service (MMS) are conventional telecom-level mobile services, which are still providing significant profits for mobile carriers. Weibo is similar to Twitter and is one of the most popular OTT social networking applications in China. Cloud upload and download are OTT applications used to upload and download files to and from the cloud. File transfer protocol (FTP) download and upload are conventional services used to transfer files. Web browsing, email services, and video play are the most common OTT applications in most smartphones today (e.g. WeChat, Tmall, iQiYi video, rich communication services, etc.).

Table 1 shows the average downlink and uplink data rates (measured in kb/s), average signaling traffic (measured in number of RE/s), data duration, and signaling duration of each mobile application. We capture all the information listed in Table 1 using a commercial software package, which is capable of analyzing signaling and data packet information transmitted via the air interface. A smartphone was first connected to a PC and configured to connect to 4G networks only. Then test scripts and test information were loaded into the smartphone to automate the testing procedure. The same testing procedure was repeated six times and we found small variations in the parameters shown in Table 1. The values listed in Table 1 are the average values from six different measurements. It should be noted that only the signaling traffic generated by using the mobile application is measured. Control signaling traffic overheads that are not directly related to the service such as CRS, synchronization signals, position tracking, and handover are not taken into account. Furthermore, it should be noted that the signaling traffic duration is generally longer than the data duration because signaling channels are first used by the mobile device to request network

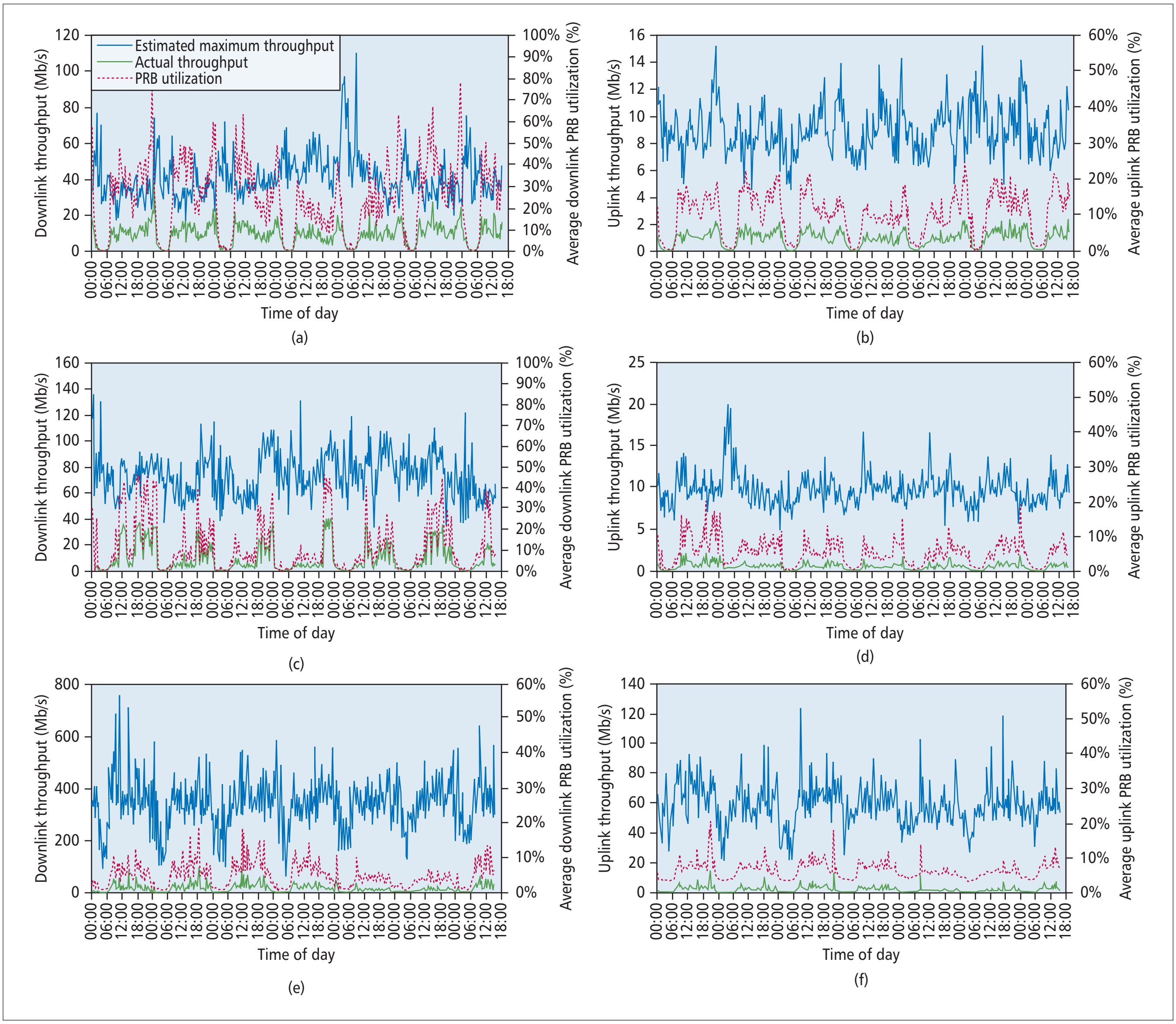


Figure 3. a), c), and e) are the downlink cell site throughputs as a function of downlink PRB utilization for cell sites A, B, and C in a week (data collected in October 2014), respectively; b), d), and f) are the uplink cell site throughputs as a function of uplink PRB utilization for cell sites A, B, and C in a week, respectively.

resources before the data transmission can begin. Upon completion of data transmission, it takes several seconds (which is known as the radio resource control timer) for the mobile device to switch from the radio resource control (RRC) "connected" mode to the RRC "idle" mode.

Due to the nature of the transmission control protocol (TCP), all applications will have both uplink and downlink data, as shown in Table 1. For example, uploading a file to the cloud is mainly dominated by uplink data. However, the server will send Acknowledgment messages to the mobile device to acknowledge the receipt of the file. Therefore, a small amount of downlink data will be received by the mobile device when uploading a file.

Energy Consumption of Mobile Applications

Next we assess the energy consumption of the 12 mobile applications (see Table 1) applying the process shown in Fig. 1. The power usage effectiveness (PUE) (i.e. for cooling, power supply efficiency, etc.) of the BS cell sites are assumed to be 1.5 according to China Mobile's cell site measurements. Figure 4 shows the energy consumption to deliver the 12 mobile applications from the three different BS cell sites (see Fig. 3) during high-peak hours (8am to 1am). Three observations can be made from Fig. 4.

First, different cell sites consume different amounts of energy to deliver the same mobile application, because different cell sites have different average throughput to PRB utilization ratios, as discussed previously in the article. Referring to Fig. 3, cell site A has the lowest throughput to PRB utilization ratio compared to the other sites, while cell site C has the highest throughput to PRB utilization ratio. This means that the average users in cell site A will require more PRBs to transfer the same amount of data compared to the users in cell site C. As more PRBs are required, the energy usage of the BS will also increase. As a result, delivering the same application at cell site C consumes the lowest amount of BS energy among the three cell sites, as shown in Fig. 4.

Second, it is also expected that for a given cell site, the energy consumption of delivering the same application varies throughout the day. This is because the cell site has different average throughput to PRB utilization ratios throughout the day, as depicted in Fig. 3. The tables in Fig. 4 show the average minimum and maximum energy consumption of delivering each mobile application from the three different cell sites during peak hours.

Third, the bar charts in Fig. 4 show the energy breakdown of each mobile application. For mobile applications that generate bursty traffic with a small amount of data, the proportion of signaling energy consumption is significant (e.g. conventional mobile services such as SMS and MMS, and OTT applications such as receiving emails and uploading small files to the cloud). For mobile applications that generate a large amount of continuous data, the data energy tends to dominate the total application energy consumption and therefore signaling energy is less important. Furthermore, the energy consumption for signaling has the same profile for all applications: the modulation format (which is using QPSK) and the signaling exchanges are independent of the link conditions and application. Therefore, the signaling energy depends only on the duration of the application session. In contrast, the link conditions, the application data rate, and the data duration determine the energy consumption for data transfer. Therefore, with good link quality, the BS consumes less energy to deliver the application because fewer PRBs are required to carry the same amount of data. As the data energy consumption decreases, the ratio of signaling energy to total application energy increases. This phenomena can be seen by comparing the signaling energy (%) for SMS, MMS, cloud, and email services in BS cell sites A, B, and C.

Significance of Signaling Energy Overhead

Observing the results presented in Fig. 4, the significance of signaling energy consumption is unclear. For example, when receiving an SMS, the BS site C will consume around 0.0013 joules of data energy and 0.067 joules of signaling energy. However, watching 60 seconds of Internet video will consume around 16 joules of data energy. Although signaling energy dominates the total energy consumption of SMS, the energy consumption of video viewing is significantly greater than that of SMS. Therefore, why should mobile carriers worry about signaling energy of mobile applications?

The usage of OTT applications has been rapidly increasing since the introduction of smartphones in 2007. Unlike conventional mobile services, OTT applications require frequent connection reestablishment to check for updates (to maintain the "always-on" characteristic). This process is known as "heartbeat" or "keep-alive" [9, 11]. For example, a social networking application client will frequently ping the server, and the server will push update information to the OTT client when necessary. Every heartbeat will trigger a significant amount of signaling traffic between the smartphone and the BS. If there is no update, the signaling traffic generated from this heartbeat will be wasted. With hundreds of millions of smartphones sending frequent heartbeats, the signaling resources of mobile access networks are constantly highly utilized. Once network congestion occurs, the OTT applications will retry connections and cause a so-called signaling storm [9, 11]. Various solutions have been proposed by 3GPP as well as Apple iOS and Android to prevent this occurrence. However, this issue has not been mitigated due to the rapid increase in the number of users of OTT applications and the number of OTT applications per smartphone.

Next we compare the estimated total monthly energy consumed by the TDD-LTE mobile access networks in delivering:

- Conventional mobile services (i.e. voice and SMS).

Due to the nature of the TCP, all applications will have uplink and downlink data. For example, uploading a file to the cloud is mainly dominated by uplink data. However, the server will send Acknowledgment messages to the mobile device to acknowledge the receipt of the file. Therefore, a small amount of downlink data will be received by the mobile device when uploading a file.

No.	Services	Test information	Average DL data (kb/s)	Average UL data (kb/s)	Average data duration (s)	Average signaling traffic (REs/s)	Average signaling duration (s)
1	Send SMS	Send a 38-Chinese character text message	1.26	1.60	1	192.00	14
2	Receive SMS	Receive a 38-Chinese character text message	1.43	0.62	1	144.00	12
3	MMS	Send a picture (size = 112 kB) with 19-character text message	1.61	139.28	19	144.00	35
4	Weibo	Combining the actions below: • Log-in as the blogger and upload a picture (size = 1.59 MB) • Log-in as the fan and repost the blogger's post	27.38	65.05	220	136.52	236
5	Cloud upload	Upload a file (size = 162 kB) to the cloud	10.19	259.53	5	145.62	18
6	Cloud download	Download a file (size = 162 kB) from Baidu cloud	191.79	6.54	7	160.28	19
7	FTP upload	Upload a large file (size = 54 MB) to a public FTP address	95.87	4678.38	101	124.05	114
8	FTP download	Download a large file (size = 54 MB) from a public FTP address	1562.90	27.28	299	138.65	311
9	Web browsing	Browse www.taobao.com without refresh and click on the page	511.13	31.43	27	138.09	39
10	Send email	Send an email with one attachment (picture size = 1.59MB)	16.31	1136.59	17	129.87	31
11	Receive email	Read the email sent from (10) without downloading the attachment	86.56	5.21	3	164.16	17
12	Video play	Open a video link from v.youku.com (length = 60 seconds, size = 1.84 MB)	267.06	6.15	62	144.00	78

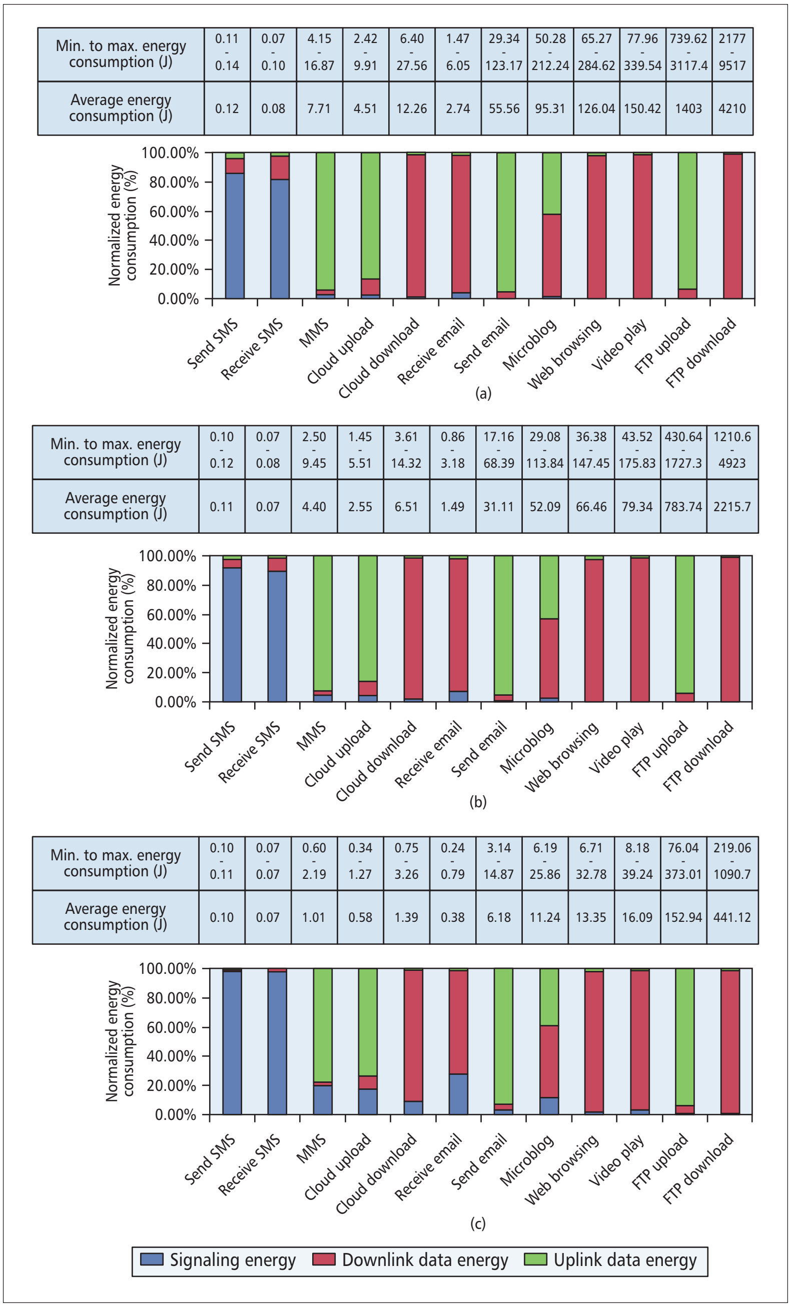
Table 1. Measured parameters for OTT applications (Weibo, cloud upload and download, FTP upload and download, web browsing, send and receive email and video play) and conventional mobile services such as SMS and MMS.

- An OTT instant messaging (IM) application.
- An OTT video application to all 4G users of China Mobile.

The estimates are based on BS cell site C in Fig. 3 and Fig. 4. The results are shown in Table 2. The monthly usage of each application and service were derived from China Mobile's network statistics and [3, 9, 12, 13]. It should be noted that the statistics were collected around the end of 2013. Therefore, it is expected that in 2014 these monthly usages and energy consumption of OTT applications did increase substantially due to the massive growth in the usage of OTT mobile applications.

For voice services, the total monthly energy consumed by all 4G users connected to China Mobile's networks was approximately 56,459 mega joules, with signaling energy representing 28 percent of the total monthly service energy. For SMS, the total monthly network energy consumption was approximately 422 mega joules, with signaling energy representing a significant 97 percent of the total monthly service energy consumption. For OTT video applications, China Mobile's 4G users consumed 218,425 mega joules of monthly mobile network energy, and the data energy represents 97 percent of the total application energy. To reduce the energy consumption of data dominated applications, network operators could offload heavy data users to small cells because the power consumption of small cells is much lower compared to macro cells [6] when used to deliver the same amount of data to the user.

For IM text messages, the total monthly network energy consumed by China Mobile's 4G users was approximately 712 mega joules, with signaling energy consumption representing 89 percent of the total monthly application energy. For IM pictures, the total monthly network energy consumed by 4G users was approximately 89 to 763 mega joules, depending on the size of the pictures. The corresponding signaling energy accounts for around 47 percent to 7 percent of the total monthly application energy for compressed and full size pictures, respectively. For IM voice messages, depending on the message length, the total monthly energy was approximately 256 to 431 mega joules, and the signaling energy varied between 73 percent and 47 percent of the total application energy. As discussed previously, OTT applications will trigger a frequent heartbeat to maintain "always-on" connections with the server as a stay-alive mechanism. Assuming a heartbeat frequency of five minutes per application [9, 12], and the average usage hours of mobile users in China is less than five hours [13], there will be a total idle period of 19 hours each day, which corresponds to 228 heartbeats a day from a single OTT application per user. This generates a staggering 577 billion heartbeats, which consumed 35,614 mega joules (98 percent signaling and 2 percent data) of China Mobile's TDD-LTE RAN energy per month solely from one OTT IM application. To reduce the energy consumption due to LTE signaling overheads for IM-type applications, the applications should be more network-friendly by considering the following two approaches:



Min. to max. energy consumption (J)	0.11 - 0.14	0.07 - 0.10	4.15 - 16.87	2.42 - 9.91	6.40 - 27.56	1.47 - 6.05	29.34 - 123.17	50.28 - 212.24	65.27 - 284.62	77.96 - 339.54	739.62 - 3117.4	2177 - 9517
Average energy consumption (J)	0.12	0.08	7.71	4.51	12.26	2.74	55.56	95.31	126.04	150.42	1403	4210

Min. to max. energy consumption (J)	0.10 - 0.12	0.07 - 0.08	2.50 - 9.45	1.45 - 5.51	3.61 - 14.32	0.86 - 3.18	17.16 - 68.39	29.08 - 113.84	36.38 - 147.45	43.52 - 175.83	430.64 - 1727.3	1210.6 - 4923
Average energy consumption (J)	0.11	0.07	4.40	2.55	6.51	1.49	31.11	52.09	66.46	79.34	783.74	2215.7

Min. to max. energy consumption (J)	0.10 - 0.11	0.07 - 0.07	0.60 - 2.19	0.34 - 1.27	0.75 - 3.26	0.24 - 0.79	3.14 - 14.87	6.19 - 25.86	6.71 - 32.78	8.18 - 39.24	76.04 - 373.01	219.06 - 1090.7
Average energy consumption (J)	0.10	0.07	1.01	0.58	1.39	0.38	6.18	11.24	13.35	16.09	152.94	441.12

Figure 4. Energy consumption of delivering the mobile applications from BS cell sites A, B, and C (refer to Fig. 3). The bar charts show the energy breakdown of each mobile application and the tables show the minimum, maximum and average energy consumed by the BS cell sites in delivering the mobile applications: a) base station site A; b) base station site B; c) base station site C.

- Using a network socket request manager (NSRM) designed to bundle the heartbeat requests of all applications installed in a smartphone to reduce the number of heartbeats generated by a mobile device [14].
- Further reduce the number of heartbeats per mobile device by re-designing the heartbeat mechanism in such a way that the heartbeat durations are adapted to the user's daily active time (i.e. when the user is active). During inactive periods, the heartbeat durations could be increased (instead of having a fixed duration) to reduce the

number of heartbeats per day per device and hence reducing the overall signaling energy consumption.

Conclusion

Massive deployments of mobile networks are driven by the rapid increase in both mobile data traffic and the usage of OTT applications. As a result, the continuing increase in the energy consumption of mobile carriers has become a major concern. Therefore, it is crucial for mobile carriers to understand the energy consumed by their network in delivering different mobile applications and services to their users. In this article we presented a method to assess the energy consumption of mobile applications and services. We also presented the three stages of measurement information required to conduct the assessment: BS power measurement, BS throughput as a function of PRB utilization, and the data and signaling traffic measurements of mobile applications. The results in this article show that using the same mobile application at different times of the day at different BS cell sites could result in different energy consumption levels due to different ratios of BS throughput to PRB utilization. The results also show that for mobile applications that generate small amounts of data, the proportion of signaling energy is significant. In contrast, as the data traffic of the mobile application increases, the significance of signaling energy consumption decreases. However, with the number of active users of small data OTT applications increasing rapidly, signaling energy consumption will become a major concern. Using statistics from China Mobile and other sources, we estimate the signaling traffic generated by an OTT IM application in a TDD-LTE mobile access network could consume a total monthly energy of 36,106 to 36,134 mega joules compared to data energy of 565 to 1,386 mega joules. As both the number of users for OTT applications and the number of OTT applications installed per mobile device continue to increase, the data and signaling energy consump-

	Average session rate (kb/s)	Average session length (s)	Data size (kBytes)	Per service energy (joules)		Monthly usage (4G users)	Total monthly energy (mega joules)	
				Data	Signaling		Data	Signaling
Conventional mobile service								
Voice	13.3	60		0.9882	0.3884	41 billion minutes[a]	40,530	15,929
SMS			0.25	0.0017	0.0610	6.7 billion messages[a]	11	411
OTT instant messaging application								
Text message (average 1 kB)			0.94	0.0071	0.0610	10.45 billion messages[b]	74	638
Picture • Compressed (10 kB) • Full (150 kB)			10 150	0.0747 1.1207	0.0666 0.0832	633.72 million pictures[b]	47 to 710	42 to 53
Voice • Short – 3 s (3 kB) • Long – 10 s (10 kB)			3.07 10	0.0230 0.0747	0.0610 0.0666	3.05 billion voice messages[b]	70 to 228	186 to 203
Idle (heartbeat)			0.082	0.0006	0.0611	577 billion heartbeats[c]	374	35,240
OTT video application								
Video	200	60	1500	11.2050	0.3884	18.84 billion minutes[d]	211,108	7,317

Table 2. Estimated total energy consumed by conventional mobile services and OTT applications in 2013.

Notes

[a] the monthly usage of conventional voice and SMS services are based on China Mobile's annual report.

[b] the monthly usage of short text messages, pictures, and voice messages are extrapolated from the China Mobile Hubei network operations report 2013.

[c] assuming 5 minutes of heartbeat frequency during an idle period [12], an average idle period of 4G users in China of 19 hours/day [13], and China Mobile's 4G customer base of around 84.4 million users [3].

[d] the monthly usage of video is based on China Mobile's network data. The estimates are based on BS cell site C in Fig. 3 and Fig. 4.

tion results presented in this article are expected to grow significantly. The energy assessment methods presented in this article will help mobile carriers better quantify the energy consumption required to provide various mobile applications through their networks.

Acknowledgment

We thank Ampalavanapillai Nirmalathas (CEET), Elaine Wong (CEET), Robert Ayre (CEET), Jeff Cheong (CEET), Leith Campbell (CEET), Zecai Shao (CMCC), Jinwei He (CMCC), and Ke Li (CMCC) for valuable discussions. We thank China Mobile network operation centers for providing the base station log information.

References

[1] S. Korotky, "Semi-Empirical Description and Projections of Internet Traffic Trends Using a Hyperbolic Compound Annual Growth Rate," *Bell Labs Tech. J.*, vol. 18, no. 3, Dec. 2013, pp. 5–21.

[2] R. Karimiyazdi and M. Mokhber, "Improving Viral Marketing Campaign via Mobile Instant Messaging (MIM) Applications," *Proc. 1st World Virtual Conf. Social and Behavioural Sciences*, Jun. 2015, pp. 1–13.

[3] "China Mobile Limited Annual Report 2014," China Mobile Limited Annual Report, 2015.

[4] Vodafone Group PLC, "Sustainability Report 2013/14," Vodafone Sustainability Report, 2014.

[5] C. I, "Green Evolution of Mobile Communications (CMCC Perspective)," GreenTouch Members Meeting, Nov. 2012.

[6] 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.

[7] Y. Wu *et al.*, "Green Transmission Technologies for Balancing the Energy Efficiency and Spectrum Efficiency Trade-off," *IEEE Commun. Mag.*, vol. 52, no. 11, Nov. 2014, pp. 112–20.

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

[9] Y. Chen *et al.*, "Small Data Optimized Radio Access Network Signaling/Control Design," *Proc. IEEE Int'l. Conf. Commun. Workshop*, Jun. 2014, pp. 49–54.

[10] "Tencent Holdings Limited Annual Report 2014," Tencent Holdings Limited Annual Report, 2015.

[11] Y. Choi, "The Impact of Application Signaling Traffic on Public Land Mobile Networks," *IEEE Commun. Mag.*, vol. 52, no. 1, Jan. 2014, pp. 166–172.

[12] C. I *et al.*, "Toward Green and Soft: A 5G Perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 66–73.

[13] R. Shi, "Where did Mobile Traffic Go? China's First Mobile Phone Traffic Usage Report Released," *Global Business*, vol. 6, 2014, pp. 18–19.

[14] S. Tarkoma *et al.*, *Smartphone Energy Consumption: Modelling and Optimization*, Cambridge University Press, 1st ed., 2014.

Biographies

Chien Aun Chan (chienac@unimelb.edu.au) received his Ph.D. degree in electrical engineering from the University of Melbourne (UoM) in 2010. Since 2011 he has been a research fellow with the Centre for Energy-Efficient Telecommunications, UoM, where his research focuses on developing new modelling techniques and energy-efficient techniques for Internet and mobile services. His research interests include energy efficiency of content distribution networks, optical access networks, and mobile wireless networks.

Wenwen Li (liwenwen@chinamobile.com) received her M.S. degree in telecommunications engineering from Xidian University in 2008. After graduation she joined the Terminal Technology Department of the China Mobile Research Institute as a project manager. In 2013 she joined the Green Communications Research Center to continue her research in the fields of energy-efficient communications and wireless interface protocol of TD-SCDMA/TD-LTE/WLAN for network end-to-end green ecosystems. She holds five patents and has published extensively in these fields.

Sen Bian (biansen@chinamobile.com) is currently a senior project manager in the Green Communications Research Center of China Mobile Research Institute (CMRI). His research interests are mainly focused on green technologies in wireless communication systems, green energy in mobile networks, and energy harvesting technology for low power systems. Prior to joining CMRI he was a senior system engineer at Motorola China Ltd., responsible for the research in mobile network performance.

Chih Lin I (icl@chinamobile.com) received her Ph.D. degree in electrical engineering from Stanford University, and has almost 30 years experience in the wireless communication area. She has worked in various world-class companies and research institutes, including the wireless communication fundamental research department of AT&T Bell Labs; Headquarters of AT&T, as Director of Wireless Communications Infrastructure and Access Technology; ITRI of Taiwan, as Director of Wireless Communication Technology; Hong Kong ASTRI, as VP and the Founding GD of the Communications Technology Domain. She received the *IEEE Transactions on Communications* Stephen Rice Best Paper Award, and is a winner of the CCCP "National 1000 Talent" program. Currently she is the China Mobile Chief Scientist of Wireless Technologies, in charge of advanced wireless communication R&D efforts of the China Mobile Research Institute (CMRI). She established the Green Communications Research Center of China Mobile, spearheading major initiatives including 5G key technologies R&D; high energy efficiency system architecture, technologies, and devices; green applications; and C-RAN and soft base station. She was an elected board member of the IEEE Communications Society, Chair of the Communications Society Meeting and Conferences Board, and the Founding Chair of the IEEE WCNC Steering Committee. She is currently the Chair of FuTURE Forum 5G SIG, an executive board member of GreenTouch, and a network operator council member of ETSI NFV.

André F. Gygax (agygax@unimelb.edu.au) received his Ph.D. degree in finance from the University of Melbourne, where he is currently on the faculty of the Department of Finance. He is a research associate at the Centre for Energy-Efficient Telecommunications at the Melbourne School of Engineering, and a fellow at the Centre for Business Analytics at Melbourne Business School. His research and teaching focus on the life cycle of industries, firms, technologies, and networks. He is a member of Beta Gamma Sigma.

Christopher Leckie (caleckie@unimelb.edu.au) received the B.Sc. degree in 1985, the B.E. degree in electrical and computer systems engineering (with first class honors) in 1987, and the Ph.D. degree in computer science in 1992, all from Monash University, Australia. He joined Telstra Research Laboratories in 1988, where he conducted research and development into artificial intelligence techniques for various telecommunication applications. In 2000 he joined the University of Melbourne, Australia, where he is currently a professor in the Department of Computing and Information Systems. His research interests include using artificial intelligence for network management and intrusion detection, and data mining techniques such as clustering.

Ming Yan (yanm@cuc.edu.cn) received his Ph.D. in communication and information systems from the Communication University of China (CUC) in 2012. He then joined the Faculty of Science and Technology at CUC as a research assistant. From April 2014 to March 2015 he was a visiting research scholar at the Centre for Energy-Efficient Telecommunications (CEET), University of Melbourne, where his research focused on developing new energy models for mobile services. His research interests include green technologies in wireless communication systems and mobile multimedia broadcast technologies.

Kerry Hinton (k.hinton@unimelb.edu.au) received a B.E. (hons) in 1978 from the University of Adelaide, and a Ph.D. in theoretical physics from the University of Newcastle Upon Tyne in 1984. In 1984 he joined Telstra Research Laboratories, working on modelling of optical systems and components. In 2006 he joined the University of Melbourne, at the Centre for Ultra-Broadband Information Networks. In 2011 he joined the Centre for Energy-Efficient Telecommunications (CEET), researching the energy efficiency of the Internet. He is now Director of the CEET.

As both the number of users for OTT applications and the number of OTT applications installed per mobile device continue to increase, the data and signaling energy consumption results presented in this article are expected to grow significantly.

Toward Green Data Centers as an Interruptible Load for Grid Stabilization in Singapore

Wenfeng Xia, Yonggang Wen, Kok-Chuan Toh, and Yew-Wah Wong

The authors are with Nanyang Technological University.

ABSTRACT

For sustainability and environmental friendliness, renewable energy (RE) and distributed generation (DG), for example, photovoltaic, are being integrated in electrical systems in many countries. RE and DG, however, can be unstable for the power grid. As the power grid integrates an increasing amount of RE and DG, we present technical solutions along with an economic incentive model to enable data centers to serve as a novel "interruptible" load (i.e. a power load that can be scaled down temporally) to stabilize the power grid. We propose a novel real-time power analytics framework called embedded software as sensors, where software hooks are embedded into a range of data center subsystems, from chip to system to application level, to monitor ICT activities and power usage in a fine grained, real-time manner. Data from these virtual sensors are then mined to construct energy consumption models, which in turn are used to develop optimal algorithms for energy aware operation of computing, power distribution, and cooling systems in the data center. This holistic monitoring and optimization framework can reduce the overall power consumption of the data center, and furthermore enables time shifting of workloads in the data center in response to power fluctuations arising from the integration of RE and DG to the power grid.

INTRODUCTION

The world is in the midst of transforming its electrical system to integrate more renewable energy (RE) and distributed generation (DG), such as wind power, solar energy, biomass, and geothermal energy. In Singapore, owing to the availability of intense equatorial sunshine, solar photovoltaic (PV) represents a key energy source of the future. It has been reported that renewable energy would meet 7–10 percent of 2012's demand in Singapore by 2025, and PV would contribute 5.6 percent of the total electricity generated. However, distributed generation and solar power suffer from variable yields during nighttime and on days when the skies are cloudy. These changing yields would result in potential instability in the power grid, possibly leading to catastrophic effects. Multiple approaches have emerged from both the supply and demand sides to address this problem, with varying costs and complexities [1]. Among them, one of the simplest yet most effective methods is to identify so called "interruptible loads," where large scale but non-critical electricity consumers can scale down their usage in times of low electrical supply.

In this research, we argue that the data center can be an ideal candidate for interruptible loads for grid stabilization in Singapore. Lately the data center has evolved as one of the largest electricity consumers. In Singapore, a typical data center occupies approximately 1,300 m^2 of floor space and consumes approximately 500W of power per m^2. Comparatively, power consumption of a single data center could be approximately 1,500 times more than an average household (about 400W for a three-room HDB). Moreover, given its relatively small territory, Singapore is hosting a large number of data centers (approximately 50), and commercial data center space in Singapore is forecast to grow by 50 percent from 2010 to 2015 (reaching 336,99m^2 in 2015). Finally, technological advancements, originally developed to curb energy consumption in data centers, have made it possible for data centers to scale up and down their power consumption temporally [2, 3]. For example, server virtualization technology makes it feasible to scale up and down data center operations in real-time, resulting in energy savings up to 40 percent of its peak load [4, 5]. This reconfigurable scaling leads to controllable electricity demand, which can be used to absorb short term fluctuations in the power grid from RE and DG. In addition, energy-aware scheduling techniques allow time shifting of computation tasks to periods when the grid has sufficient power, further stabilizing its load.

Our conceptual paradigm of green data center as an interruptible load for grid stabilization,

0163-6804/15/$25.00 © 2015 IEEE

is illustrated in Fig. 1. This envisioned paradigm involves two stakeholders: data center operators and power grid operators. Specifically, data center operators adopt technical solutions to measure and manage their power consumption, matching the changing yields offered by power grid operators. These two parties mutually benefit from a well defined inter-operability policy and a well designed economic model.

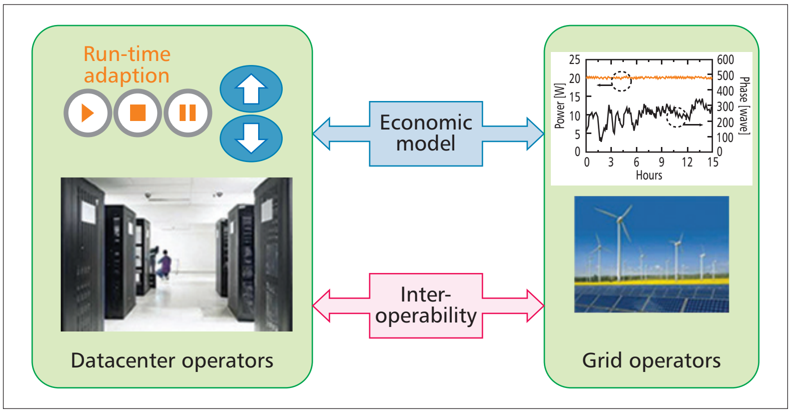


Figure 1. Conceptual paradigm: green data center can be leveraged as an "interruptible" load for grid stabilization in Singapore.

Compared to other interruptible loads (e.g. household, factory, and mill), our proposed solution stands out for its high adoption readiness and effective impact on power grid management. First, due to the enormous electricity consumption of a typical data center (equivalent to 1,500 households, as mentioned before), even a small change in its energy efficiency has a significant effect on the total electricity consumption across the grid. Second, this paradigm would yield a tremendous cost saving in Singapore. In fact, even if the proposed solution is adopted by 50 percent of data center operators and the power saving efficiency is 1/3, the total annual electricity cost savings for data center operators would be more than SGD 61 million.[1] Finally, data center operators are incentivized to adopt this paradigm to convert their cost into a potential profit opportunity. In practice, electricity consumption accounts for approximately 50 percent of the costs of data center operations, and data center operators have strong incentives to reduce their energy consumption and shift loads to times when electricity prices are low. Furthermore, it is also vastly easier to enlist the participation of a few dozen data center operators than to coordinate the actions of an equivalent of hundreds of thousands of households.

The rest of this article is organized as follows. In the following section we outline the design principles of our proposed framework. Then we describe the architecture of embedded software as sensors (ESaS), and elaborate on its key enabling subsystems. We then address technical challenges in analyzing the collected data center status data and designing data center power management strategies. Following that, we describe promising techniques for power management in data centers. Then we highlight the integration of data centers into the power grid, serving as an interruptible load. We then present business models to bring this design into reality. We conclude and summarize in the final section.

System Design Principles

Green data centers and grid stabilization are real-world issues that might not be solved without real-world practices. In this research we follow the principle that *combines practice with theory*.

In practice, we build an integrated framework to establish energy intensive data centers as an interruptible load for power grid stabilization in Singapore. Our framework is data centric, consisting of three stages:

- Data collection. We build the embedded software as sensors (ESaS) framework to collect running status data from hardware and software components in a data center.
- Data analytics. We employ various techniques on the data collected, trying to gain insights. In particular, we focus on power consumption related analytics.
- Operations adaptation. According to the data processing results, intra data center ICT optimization and inter domain optimization can be adopted for green data centers and grid load stabilization. Working with our industrial partners, Toshiba and SP PowerGrid, we are implementing and testing our proposed framework on a full-scale container-based data center, as illustrated in Fig. 2.

From practice to theory, we employ various theoretic methods for system optimization. In data collection, we investigate issues such as sensor placement, sensing duty cycle, and in-network data processes. In data analytics, we build power consumption models based on multi-factor data using theoretical models, e.g. machine learning. In operations adaptation, we design system operation strategies to minimize data center power consumption with guaranteed quality of service, and economic models encouraging all the stakeholders to work together for grid load stabilization.

Embedded Software as Sensors

For real-time data collection in data centers, we propose a new technical framework, embedded software as sensors (ESaS). In the ESaS framework, software "hooks" (e.g. open-source tools such as Net-SNMP, libvirt, libgtop, and IPMI) are embedded in ICT subsystems. These software hooks work as virtual sensors to monitor ICT activities (e.g. CPU usage, I/O statistics, memory footprint, heap statistics, and thread information) and power usage in real-time. Figure 3 illustrates a typical data center map with our proposed framework. Specifically, we have implemented a prototype of the ESaS framework, consisting of various "hooks" to source data from individual components (e.g. server, storage, network, and VM), an XMPP based information bus for data transfer, and data stored in a MongoDB database for subsequent data analytics.

To further optimize the ESaS framework, the overhead from real-time monitoring should be reduced, while maximizing information collec-

[1] We calculate the savings based on the total estimated data center floorspace in Singapore by 2015 (336,990m^2), the current electrical consumption per square meter (~500W) for a data center, the estimated energy reduction (1/3), and the current price of electricity in Singapore (25 cents/KWh). If 50 percent of data center operators adopt the solution, the savings would be (336,99m2 ¥ 500W/m^2) × (50 percent × 1/3) × (365 × 24) × *S*$0.25/KWh ≈ *S*$61,500,000.

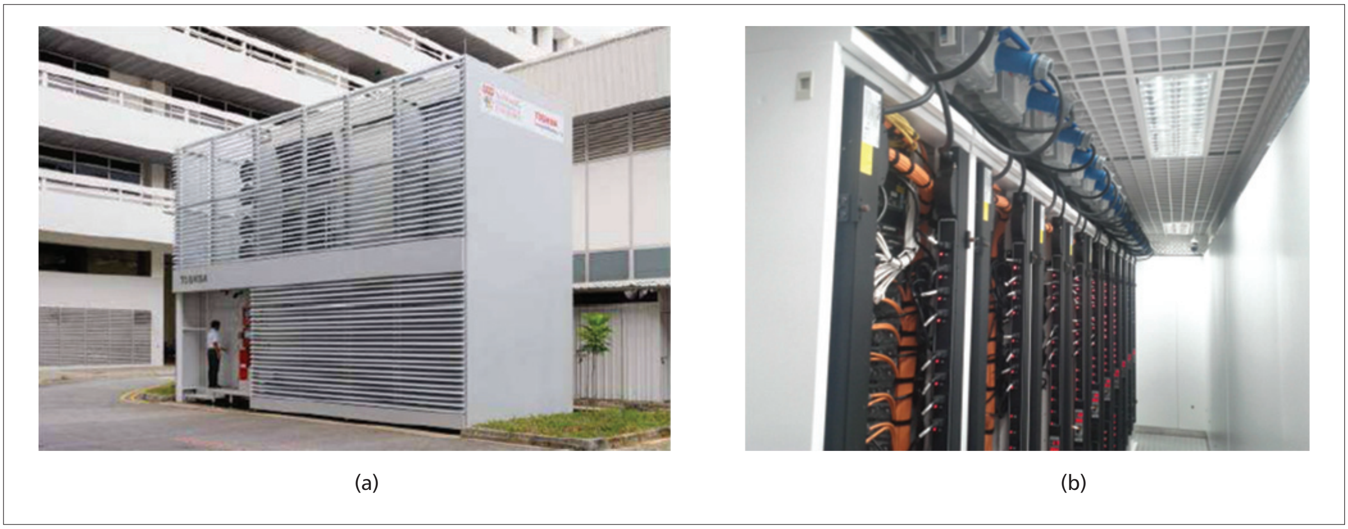


Figure 2. Modular data center testbed: a full-scale container-based data center as the testbed. a) External view; b) Internal view.

tion capability. Moreover, the virtual sensing framework is designed to be non-intrusive, with minimum impact on data center operations. To achieve these objectives, the following optimization techniques can be applied.

Sensor Placement: The sensors should be well positioned in the system to balance the trade-off between monitoring accuracy and deployment cost. In the framework of ESaS, a sensor can collect data from multiple devices. For example, an SNMP based sensor can collect data from multiple switches and servers. The amount and placement of the sensors can affect the deployment cost.

Sensing Duty Cycle: It is not cost effective to monitor all the events in a data center continuously. Hence, the duty cycle of sensors should be designed to minimize monitoring costs while providing information with high confidence. Compressed-sensing techniques can be adopted to exploit the inherent data sparsity of the underlying processes to design optimized sampling algorithms.

In-Network Data Processing and Routing: Data collected by the ESaS framework exhibits temporal and spatial redundancy; directly transferring all the data would incur unnecessary costs. In-network features can be introduced to pre-process collected data traces along data migration paths. For example, if the maximum of a set of measurements is of interest, one would strategically take the maximum of a subset of measurements along the data collection tree [6].

Fault Diagnosis: Sensor failures can result in the disruption of monitoring and subsequent data analytics. It can be difficult to detect, localize, and repair. A network fault diagnosis framework, e.g. group testing over graphs [7] for network tomography, can be adopted to diagnosis sensor failures.

Real-Time Data Center Power Analytics

In the data analytics stage, the collected data traces are analyzed via data mining [8] algorithms to construct energy consumption models for different parts of the data center infrastructure. We develop energy consumption models of data center operations by mining the relationship between ICT activities (e.g. CPU cycles, IO reads/writes, network throughput, and application delay) and measured/estimated power usage.

Models for power consumption in data centers have been a subject of extensive research. General frameworks are based on a strong affine relationship between CPU utilization and the total power consumption of a server. However, this model is not sufficient to determine energy usage in a data center for the following reasons. First, the energy consumed by servers in a data center is only a part of total ICT energy consumption. Second, ICT components such as network and storage have different power dynamics than servers. Third, the energy consumption of ICT systems indirectly dictates the energy consumed by the infrastructure such as power distribution and cooling. For these reasons, a multi-factor model must be developed for data center energy consumption.

Our power analytics approach will estimate the power consumption of a data center based on multiple features extracted from ICT logs gathered through the ESaS framework, as well as the data center's architecture. The framework then generates regression models that predict the power consumed by various subsystems in the data center in real and hypothetical optimized operations. Specifically, we will refine existing CPU-based server energy models by incorporating multi-factor data including CPU, RAM, and storage usage. We will also develop an ICT energy model based on the architecture of ICT components including computing, storage, and network. Our modified framework (Fig. 4) could be based on theoretic models, e.g. deep learning models and Gaussian mixture models.

Finally, the energy models are used to predict energy consumption for optimizing data center operations in computing, power distribution, and cooling systems, in order to achieve the following two objectives:

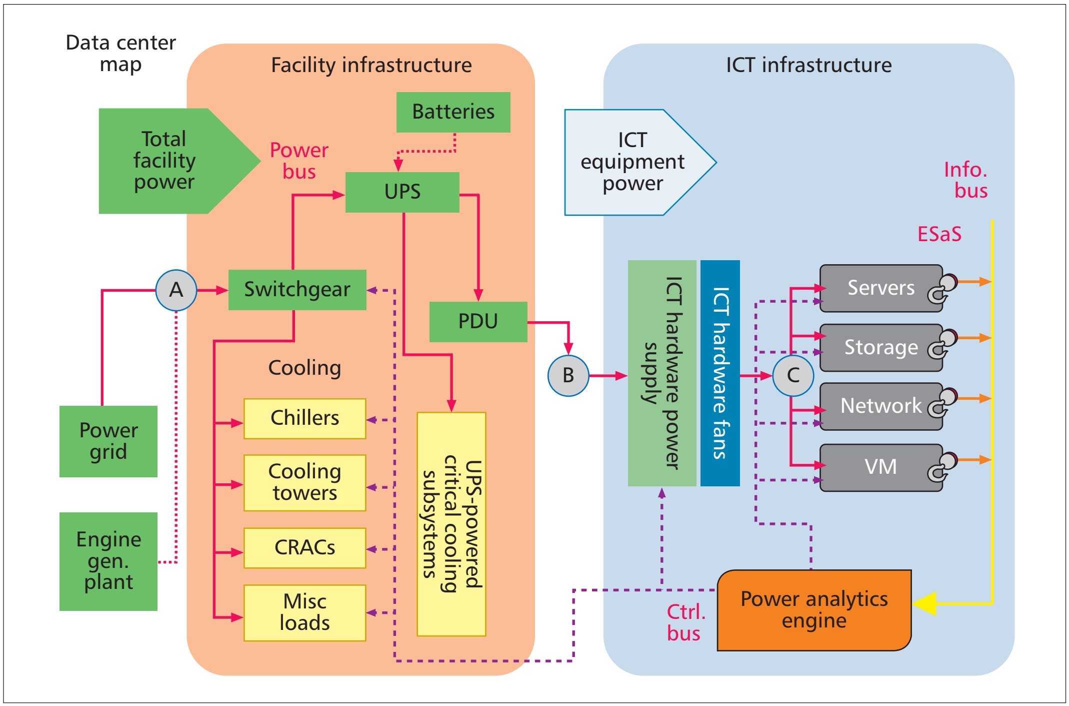


Figure 3. A typical data center map with our proposed ESaS framework.

- To reduce data center energy consumption by scaling operations up and down to match demand in real-time.
- To stabilize grid load by time shifting workloads in data centers in response to power grid fluctuations.

Dynamic Power Management in Data Centers

Existing dynamic power management strategies fall into four categories: computer system optimizations, cloud system optimizations, better software engineering practices, and inter-domain operations optimizations. Our work seeks to incorporate these existing techniques into a unified system, as well as to develop novel power saving methods.

Computer System Optimization Strategy

Computer system optimization strategy is based on control of hardware and firmware through the operating system (OS). Various features of these systems, such as their ON/OFF state or clock frequency, can be changed dynamically to control their energy usage. The adjustments are divided into two categories: dynamic component deactivation (DCD) and dynamic performance scaling (DPS). DCD judiciously deactivates components of the computer system during periods of inactivity, while DPS adjusts the performance of the components (e.g. the clock frequency) to suit a desired power consumption level.

Dynamic Component Deactivation (DCD): Which components to deactivate and when has a huge impact on power consumption. Deactivation and subsequent reactivation of a component, e.g. spinning a hard disk down and up, incurs significant time and energy costs. These costs are only worthwhile if they can be amortized by energy savings from a substantial period of inactivity. As these inactive periods change dynamically during an execution, the problem of which components to deactivate and when can be formulated as a stochastic optimization problem, which takes an appropriate probabilistic model of the system as input and generates an online schedule of energy optimal state transitions.

Dynamic Performance Scaling (DPS): DPS is exemplified by the widely adopted dynamic voltage and frequency scaling (DVFS) technique. In a server, DVFS can be applied to dynamically adjust its operating frequency to meet QoS demands while minimizing energy usage. As energy consumption varies superlinearly with a component's clock frequency, it is advantageous to scale down a component as long as it can meet its service level agreement (SLA) deadlines [9].

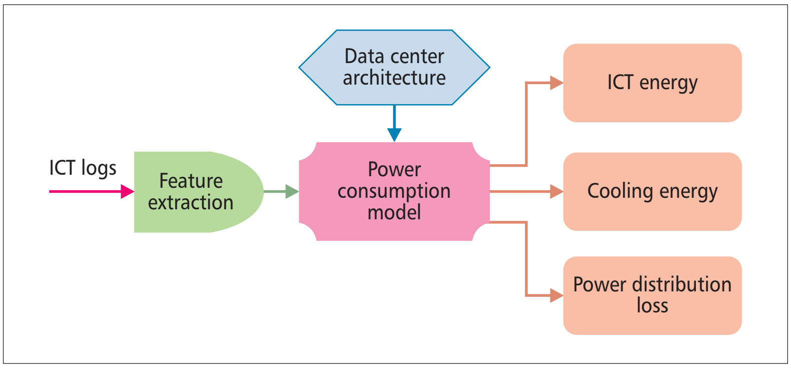


Figure 4. A multi-factor power analytic framework based on features extracted from ICT system and data center architecture.

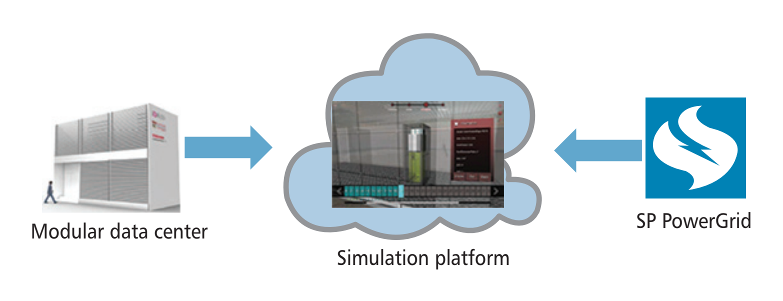


Figure 5. Simulation platform takes data input from power grid and the MDC.

Cloud System Optimization Strategy

A key recent technique to improve energy efficiency is server virtualization in cloud systems. This involves running multiple virtual machines on each physical server, and when possible consolidating workloads to a small number of servers and shutting down the rest, in order to minimize the number of active servers.

Consolidation can occur at a more refined level than switching off an entire server. As in DCD, components in a server can be selectively deactivated, and the corresponding load moved to other active components. For example, during periods of light load, several servers can share the storage unit on one server, turning off their own storage. In a similar vein, by proactively monitoring traffic load, re-adjusting channel bonding groups, and shutting down underutilized upstream/downstream ports, energy proportional networking can be realized in cable modem termination system (CMTS) networks [10].

While workload consolidation comes with clear benefits, there are costs to consider as well. One is the cost to run a hypervisor to manage the global pool of resources. A second cost is the time and energy expenditure needed to reallocate workloads. Finally, an important limitation exists on the degree to which we can effectively consolidate. Indeed, excessively loading a few machines can significantly impair overall performance. One example is the memory thrashing that occurs when multiple processes compete for the limited memory of a server on which they have been consolidated. In fact, it is observed that the energy consumption per transaction upon consolidation follows a "U"-shaped curve, in which energy is wasted with either too little or too much consolidation, i.e. underutilization or overloading [11]. Several techniques can be utilized to determine the optimal level of consolidation, including:

- Joint VM placement and migration strategy in energy aware data centers.
- Leveraging the geographical distribution of data centers.
- Balancing loads across multiple servers.

Energy Aware Software Design

Energy efficient hardware techniques need to be coupled with energy conscious software to realize maximum savings. Energy efficiency should be elevated to a first-order consideration in software design and development. To this end, energy conscious design principles as guidelines for software design, and high performance energy minimizing software libraries targeting specific high consumption tasks in data centers for software development, will be significant.

Typical applications where these techniques can be applied include media streaming applications and big data analytics. These are two of the most energy intensive tasks in data center computing, but they are also sufficiently structured and flexible so as to enable a range of optimizations to reduce or shift their peak energy consumption. In particular, we observe a large diversity in the energy requirements and dynamic ranges of ICT components such as CPU, RAM, disk, and networking. In many instances we can obtain power savings by increasing loads of certain components in order to reduce loads on others. For example, a popular video file should be cached in memory while less popular ones should be directly read from disk. When the transition from "popular" to "not" occurs, and which files to cache in limited amounts of memory, becomes an optimization problem that needs to simultaneously consider the application, architecture, and instantaneous energy availability and cost. Another important class of optimizations is to adjust the quality of computations in accordance with current energy costs. For example, one can reduce the quality of a video stream or reschedule a big data computation when energy costs are high. Again, determining the right trade-off between quality and cost is an optimization problem in the underlying economics.

Inter-Domain Operations Optimization

With our unified power analytics substrate, various inter-domain operational strategies can be sought to build interaction between ICT infrastructure and other infrastructures (e.g. power supply and cooling), and further reduce the energy consumption of data centers. Two areas of optimization will be sought in our research.

Power Distribution: Smart power distribution reduces cable and converter losses by channelling power flows to servers optimally in response to shifting server loads. This strategy can leverage the ESaS framework to formulate a localized "smart grid," providing intelligent energy management across systems and facilities.

Cooling Technologies: Advanced cooling technologies that react proportionally to the amount of load at the rack level can save energy by only providing required cooling. For current air cooled systems, an air distribution network is required to react in a timely manner with the right quantity of air to the workload in each rack. The emphasis is on understanding the transient response of the system to sensor information so as to avoid over-provisioning that is typical to cater to uncertainties in the air distribution. To bring cooling closer to the ICT equipment, and hence to be able to follow ICT loads more intimately, close coupled liquid cooling technologies can be introduced. However, higher penetration of such cooling arrangements requires resolution of design issues related to operating such systems with mainstream server equipment and long-term reliability concerns. A critical challenge is how to integrate the cooling structures into the server system to provide seamless operation and maintenance. A major shift in design philosophy is required from an add-on cooling system design to one which is part of the server and rack architecture.

Interoperability between Data Center and Smart Grid

Green data centers as interruptible loads for grid stabilization is an innovative idea. The interoperability issues between data centers and the power grid, and the impact on operators of data centers and the power grid, should be investigated before real-world practices. We focus on identifying operational changes and supporting infrastructures for both operators, and understanding potential risks of these changes in terms of reliability, security, and cost.

Simulation Platform

Given its potential risk, we conduct this part of our research via emulation and simulation. We set up a simulation platform in our lab. The simulation platform, as illustrated in Fig. 5, has three parts, including:

- Power Grid: We obtain a sanitized data set from SP PowerGrid to emulate a power distribution network. As illustrated in Fig. 6, the data set mainly consists of two parts: grid load that follows daily and weekly patterns (Fig. 6a), and an electricity tariff that is adjusted quarterly (Fig. 6b).
- Modular Data Center: We install smart meters in our modular center testbed to measure its current status (and feed it into our simulation platform.
- Simulation Platform: We use the energy consumption models from the real-time power analytics to emulate the operations of a typical data center. This platform also provides a 3D view of the data center [12].

To study the effectiveness of using data centers as an interruptible load for grid stabilization, two suites of test cases are to be investigated on this emulation platform. The first suite deals with data centers as an independent load for the power grid, by investigating its impact on grid load stabilization during fluctuations. The second suite addresses the case when the data center coexists with other loads (e.g. households, factories, and office buildings). In both cases, we aim to gain insights into the scale in which data centers can shape power grid demand, providing information for the economic model research.

Economic Model

The success of our proposed technical solution requires an economic incentive model that puts the interests of all stakeholders (IT users, data center operators, and power grid operators) in balance. The payment exchange between IT users and data center operators is often unidirectional and inflexible. In Singapore, power price is adjusted quarterly by the government, as shown in Fig. 6b. Specifically, when the SLA is delivered to IT users, data center operators will receive payments accordingly. However, the payment exchange between data center operators and power grid operators could be bidirectional. In one direction, data center operators pay power grid operators for electricity usage. In the other direction, when the power grid suffers from a supply shortage due to stochastic yields, the data center operators can voluntarily scale down energy consumption, while satisfying the SLA with IT users and receiving some credits or discounted prices from power grid operators.

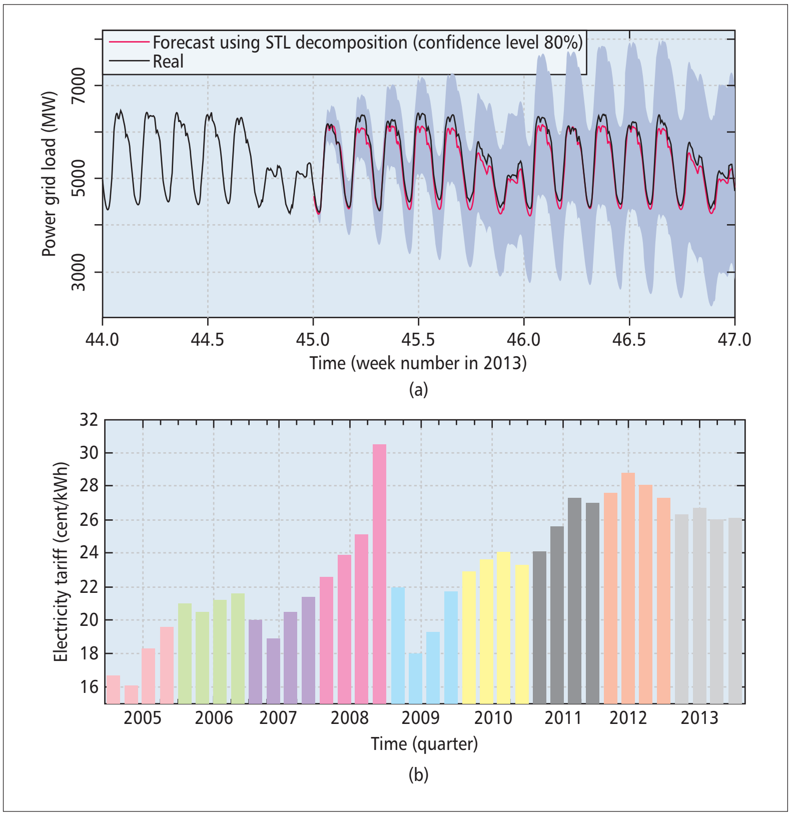


Figure 6. Singapore power grid load and price. a) Power grid load prediction using historical data in Singapore; b) Quarterly electricity tariff in Singapore from 2005 to 2013.

On this subject, a well designed payment scheme between data center operators and power grid operators should be in order. The payment scheme should be designed with the following two objectives:

- No Arbitrage Scenario: None of the stakeholders can exploit another party for free service.
- Marginal Utility Equilibrium: No one has an incentive to deviate from the resulting pricing scheme.

A game theoretic utility model, for example, can be adopted to price the electricity provided by the power grid operator, accounting for the flexibility in power usage provided by the data center operator.

Conclusions

Renewable energy (RE) and distributed generation (DG) are being widely adopted, which can be unstable. We proposed using green data centers as an interruptible load for grid stabilization. The novelty of our proposed solution is multi-fold. First, to our knowledge we probably are the first in the world to propose data centers as an interruptible load for grid stabilization. Our solution is not only technically feasible, but has greater practicality than other methods, such as controlling the energy usage of vast numbers of households. As the first of its kind, our idea has the potential to be game changing. Second, the ESaS framework transforms data center monitoring, traditionally based on wireless sensor networks (WSNs) and power meters, with a new paradigm based on embedded software as virtual sensors. Our solution is inherently green, low cost, and provides fine grained information collection. Third, our proposed real-time power analytics provides the capability of unifying operation optimizations across ICT, power distribution, and cooling. It takes a step toward a unified IP-based control system for future data centers. Finally, our integrated approach to addressing interoperability issues and business models renders our methods more practical and meaningful than purely technical solutions. Currently, we are verifying green data centers as an interruptible load for grid stabilization in Singapore. In the future, this solution may be applied in other countries with proper modifications.

Our solution is not only technically feasible, but has greater practicality than other methods, such as controlling the energy usage of vast numbers of households. As the first of its kind, our idea has the potential to be game changing.

References

[1] X. Fang *et al.*, "Smart Grid — The New and Improved Power Grid: A Survey," *IEEE Commun. Surveys Tutorials*, vol. 14, no. 4, Fourth Qtr. 2012, pp. 944–80.

[2] J. Yang *et al.*, "Dynamic Cluster Reconfiguration for Energy Conservation in Computation Intensive Service," *IEEE Trans. Comput.*, vol. 61, no. 10, Oct. 2012, pp. 1401–16.

[3] H. Yang *et al.*, "CSO: Cross Stratum Optimization for Optical as a Service," *IEEE Commun. Mag.*, vol. 53, no. 8, Aug. 2015, pp. 130–39.

[4] Y. Jin, Y. Wen, and Q. Chen, "Energy Efficiency and Server Virtualization in Data Centers: An Empirical Investigation," *Proc. 2012 IEEE Conf. Computer Communications Workshops* (INFOCOM WKSHPS), 2012, pp. 133–38.

[5] Y. Jin *et al.*, "An Empirical Investigation of the Impact of Server Virtualization on Energy Efficiency for Green Data Center," *The Computer Journal*, 2013.

[6] Z. Lu and Y. Wen, "Distributed Algorithm for Tree-Structured Data Aggregation Service Placement in Smart Grid," *IEEE Syst. J.*, vol. 8, no. 2, 2014, pp. 553–61.

[7] N. J. A. Harvey *et al.*, "NonAdaptive Fault Diagnosis for All-Optical Networks via Combinatorial Group Testing on Graphs," *Proc. IEEE INFOCOM*, 2007, pp. 697–705.

[8] H. Hu *et al.*, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," *IEEE Access*, vol. 2, 2014, pp. 652–87.

[9] Y. Wen, W. Zhang, and H. Luo, "Energy-Optimal Mobile Application Execution: Taming Resource-Poor Mobile Devices with Cloud Clones," *Proc. 2012 INFOCOM*, 2012, pp. 2716–20.

[10] Z. Zhu and W. Y. G., "Architecting Green Broadband Cable Access Network: Energy-Delay Trade-off," *Proc. IEEE OFC/NFOEC*, Los Angeles, Mar. 2011.

[11] S. Srikantaiah, A. Kansal, and F. Zhao, "Energy Aware Consolidation for Cloud Computing," *Proc. 2008 Conf. Power Aware Computing and Systems*, vol. 10, USENIX Association, 2008.

[12] J. Yin *et al.*, "Cloud3DView: An Interactive Tool for Cloud Data Center Operations," *Proc. ACM SIGCOMM 2013 Conf. SIGCOMM*, ser. SIGCOMM '13, New York, NY, USA: ACM, 2013, pp. 499–500.

Biographies

WENFENG XIA (wenfeng_xia@ntu.edu.sg) received his B.S. and M.S. degrees in computer science from the University of Science and Technology of China (USTC), Hefei, Anhui, China, in 2011 and 2014. He is currently a research assistant at Nanyang Technological University (NTU) in Singapore. His research interests include data center networks and software defined networking.

YONGGANG WEN (S'99–M'08–SM'14) (ygwen@ntu.edu.sg) received the Ph.D. degree in electrical engineering and computer science (minor in western literature) from Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. He is currently an assistant professor at the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include cloud computing, green data centers, big data analytics, multimedia networks, and mobile computing.

KOK-CHUAN TOH has a B.E. (mech) from the University of Auckland, a Dip.Bus.Admin from the National University of Singapore, and an M.S.(mechanical engineering) from Stanford University. He is a principal research scientist with Temasek Laboratories at Nanyang Technological University in Singapore. His current areas of interest are in thermal management of electronic systems and energy efficiency of data centers.

YEW-WAH WONG obtained his B.Eng. (hons) from the then University of Singapore in 1976, and an M.Sc. in mechanical engineering from the National University of Singapore in 1985. He is senior research fellow at the NTU School of Mechanical and Aerospace Engineering and Energy Research Institute@NTU (ERI@N). His research interest is in efficient energy use in buildings and data centers.

Green Sensing and Access: Energy-Throughput Trade-offs in Cognitive Networking

Hossein Shokri-Ghadikolaei, Ioannis Glaropoulos, Viktoria Fodor, Carlo Fischione, and Antony Ephremides

ABSTRACT

Limited spectrum resources and the dramatic growth of high data rate applications have motivated opportunistic spectrum access exploiting the promising concept of cognitive networks. Although this concept has emerged primarily to enhance spectrum utilization and to allow the coexistence of heterogeneous network technologies, the importance of energy consumption imposes additional challenges, because energy consumption and communication performance can be at odds. In this article the approaches for energy efficient spectrum sensing and spectrum handoff, fundamental building blocks of cognitive networks, are investigated. The trade-offs between energy consumption and throughput, under local as well as under cooperative sensing, are characterized. We also discuss the additional factors that need to be investigated to achieve energy efficient cognitive operation under various application requirements.

INTRODUCTION

The popularity of devices such as smart phones, tablets, and laptops all wirelessly connected to the Internet, as well as the recent development of the Internet of Things paradigm, has led to an ever-growing demand for spectrum along with the need for heterogeneous networking technologies that suit various networked applications. Cognitive networks and opportunistic spectrum access in licensed frequency bands may become a key technology to address those demands by increasing spectral efficiency, providing sufficient resources to realize massive machine-type communication for billions of interconnected devices forecast for 2020,[1] and facilitating coexistence among diverse networks and integration into future cellular networks.

There has been substantial effort devoted to the design of cognitive networks, with a focus on throughput maximization. At the same time, the importance of reduced energy consumption, to reduce operating costs and to support battery operated devices, has imposed new challenges. As reducing energy consumption may reduce communication performance, energy consumption optimization must be considered with the target application quality requirements in mind. One of the main existing efforts to address this challenge is GreenTouch, a consortium of academia, vendors, and operators, launched in 2010, with the mission of decreasing per bit energy consumption to one-thousandth of that in 2010 by 2015.[2]

The investigation of energy efficient cognitive radio technology as a means to increase the spectral efficiency of future wireless networks requires the understanding of the energy cost imposed by the functionalities related to the cognitive operation. Compared to traditional wireless networks, opportunistic spectrum access in a cognitive network requires appropriate spectrum sensing and spectrum handoff mechanisms, which may be a substantial source of energy consumption in a network with a large number of cognitive devices. In general, more accurate sensing and handoff control demands higher energy consumption, which can be justifiable if it leads to a significant gain in spectrum utilization. Thus, it introduces a trade-off between energy consumption and throughput enhancement. Our objective is to characterize this trade-off and evaluate which parameters need to be considered to optimize cognitive network operation in different networking environments. Based on the discussion of existing proposals, we evaluate the additional parameters, such as cooperative sensing incentives, that should be considered to allow energy efficient operation in large networks, where users may have different transmission needs and possibly conflicting interests.

FUNDAMENTALS

COGNITIVE NETWORKS FOR OPPORTUNISTIC SPECTRUM ACCESS

Under opportunistic spectrum sharing, two or more networks share some part of the spectrum. The primary network, with several primary users

Hossein Shokri-Ghadikolaei, Ioannis Glaropoulos, Carlo Fischione, and Viktoria Fodor are with the KTH Royal Institute of Technology.

Antony Ephremides is with the University of Maryland.

This work was supported by the FP7 EU project Hydrobionets.

[1] See http://machinaresearch.com/forecasts/.

[2] Detailed information is available on http://www.greentouch.org.

0163-6804/15/$25.00 © 2015 IEEE

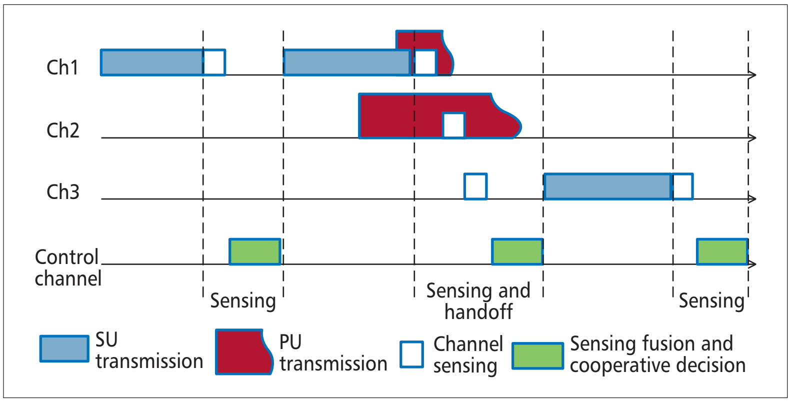


Figure 1. The SU interrupts transmission to evaluate the availability of the channel using cooperative sensing. If the channel is busy, it starts a spectrum handoff procedure to find an idle channel. Spectrum sensing and handoff consume both energy and time.

(PUs), owns the spectrum. The secondary network(s) of secondary users (SUs) can access the spectrum if this action does not cause significant degradation of the primary communication, in terms of interference level, throughput, or delay. As secondary communication needs to take the state of the primary network into account, cognitive network operation is necessary.

To find opportunities for spectrum access, the cognitive secondary network learns the wireless environment and adapts to it. The learning is often based on sensing the spectrum, while the adaptation includes the tuning of various parameters of the communication stack protocols of the secondary network. As shown in Fig. 1, to effectively find the transmission opportunities and to protect the PUs from harmful interference, the SUs need to sense the channels regularly using local or cooperative sensing, and to start a spectrum handoff procedure, if the current channel is busy.

Spectrum Sensing

The most important parameters affecting the performance of spectrum sensing are the time available to sense the transmission channels and the strength of the primary signals. A-priori information on the primary technology may determine which spectrum sensing method should be applied, ranging from energy sensing to feature based detection schemes. Under all schemes, however, noise and channel impairments such as shadowing and fading lead to decision errors, quantified in terms of false alarm and misdetection probabilities. A false alarm occurs when a free channel is mistakenly sensed busy, while a misdetection happens whenever an occupied channel is sensed free. To improve sensing performance, cooperative sensing may be introduced, where a group of SUs together decide about the availability of the channel, increasing the robustness of spectrum sensing by utilizing the spatial diversity of the individual links.

Multichannel Spectrum Sensing

As typically there are more than one primary channels available for secondary access, spectrum sensing methods are generally classified into wideband and narrowband sensing. Under wideband spectrum sensing, an SU senses multiple channels simultaneously. Although this may allow short sensing duration, it requires complex hardware implementation. Under narrowband sensing, only one channel can be sensed at a time, which allows simple sensing hardware and decision mechanisms, and therefore this is the preferred solution for most of the proposed cognitive systems. In this case, sensing time and sensing energy consumption may increase linearly with the number of sensed channels.

Spectrum Handoff Strategy

The spectrum handoff strategy answers the questions: When should an SU vacate the current channel? Should the SU wait on this channel or start searching for an available channel? Which channels should be sensed and in what order?

The strategies can be categorized as reactive, proactive, or as a combination of these, hybrid. Under reactive spectrum handoff, the SU recognizes that a PU has started to use the channel, and therefore it needs to vacate the channel. The SU then initiates searching among the channels to find transmission opportunities and pursue its unfinished transmission. Although a larger delay becomes inevitable, reactive spectrum handoff builds on up-to-date channel status estimation. Proactive schemes, on the other hand, exploit the long term traffic statistics of the channels to establish a proper policy for future spectrum handoffs. To allow detailed channel occupancy statistics, these schemes may require two radios, one for transmission and one for continuously scanning the channels. Hybrid strategies combine the advantages of the two basic schemes, that is, prepare the sensing order of the channels in advance based on available statistics, but perform reactive channel sensing at handoff triggers to find an idle channel.

Energy Efficiency

Energy efficiency is generally defined as the number of information bits transmitted per unit of energy, measured in *bit-per-Joule*. Alternatively, it is reflected by the energy cost, that is, the energy required to transmit a unit of information, measured in *Joule-per-bit*. The energy consumed in the secondary network consists of consumption for data transmission and reception, spectrum sensing, and the communication protocol to support the cognitive operation, including, for example, information exchange for cooperative spectrum sensing and for coordinating secondary transmissions. Finally, minor components are the circuit powers and the power consumed for tuning to a target channel [1, 2].

By Shannon's capacity formula, it is known that in a dedicated spectrum, linearly increased transmission power leads to a logarithmic increase in the achievable transmission rate. Consequently, energy efficiency, as the ratio of the rate and the invested energy has an optimum value. The trade-off between energy consumption and throughput becomes more complex in cognitive networks, as sensing consumes energy, valuable time when the primary channel is idle, and also communication resources for sensing cooperation.

In Table 1 we summarize the solutions pro-

Design parameters	Applicability	Characteristics	Reference
Channel sensing time	Necessary in all networking scenarios; gain under loose PU interference constraints	Closed form based optimization, or quasi-concave optimization under non-backlogged traffic	[3, 4]
Waiting or handoff	Necessary to avoid unnecessary handoffs in multi-channel systems; gains under loose SU delay constraints and heterogeneous channel occupancy	Convex optimization under homogeneous primary channels; suboptimal algorithms with polynomial complexity for heterogeneous channels; building on the channel occupancy statistics	[2, 5]
Sensing order	Gains under heterogeneous channel occupancy in multi-channel systems	Greedy or polynomial suboptimal algorithms; need learning of the channel occupancy statistics	[1, 6]
Handoff maximization	Significant gain for non-saturated SUs	Closed form based optimization	[3]
Combined sensing and channel access	Important for uncoordinated SUs; efficient if SU load is not too high	Local optimization based on Markovian system model, suboptimal iterative algorithms	[7, 8]
Cooperation with the PU	Important under spectrum shortage; significant gain if PU delay requirements are not strict	Local; convex optimization	[9]
Cooperative sensing, resource allocation	Necessary for cooperative sensing; significant gain under known SU density	Local numerical optimization based on analytic models	[10, 11]
Cooperative sensing, user selection	Necessary in cooperative sensing, significant gain under diverse and correlated local sensing performance	Efficient suboptimal greedy algorithms; possible distributed realizations; integrated with the correlation estimation	[12, 13]
Sensing report forwarding	Important if reporting costs are significant; improved efficiency if channel occupancy statistics are available	Efficient greedy algorithms if channel occupancy statistics are available, otherwise, suboptimal iterative node selection algorithms and local learning	[14]
Decision combining	Optimal combining rule leads to significant gain if reporting links are unreliable	Centralized, numerical analysis based optimization	[15]

Table 1. Main design parameters to achieve energy efficient spectrum sensing and handoff strategies.

posed to improve the energy efficiency of sensing and channel access both under local and under cooperative sensing. In the next sections, we discuss in detail the involved parameters, the effect of their optimization, and the implementation challenges. The presented results are based on a variety of primary technologies (DTV, LTE, IEEE 802.11, etc.), and therefore we discuss trends instead of quantitative results.

Local Spectrum Sensing and Handoff Strategies

Local spectrum sensing can provide adequate sensing performance if the primary signals are sufficiently strong. In the following, we discuss how key design decisions, such as per channel sensing time, number and order of handoffs, and sensing and channel access coordination, affect the energy efficiency of the cognitive network. The results we discuss typically consider energy detection based sensing, albeit more advanced sensing methodologies present similar trade-offs.

Channel Sensing Time

Sensing time is a key parameter that affects energy efficiency. Increasing the time spent to sense a single channel improves the performance of spectrum sensing at the expense of increased energy consumption and possibly decreased transmission time of an SU. In multi-channel systems, accurate sensing with long sensing times might still be beneficial, as this can avoid unnecessary handoffs, leading to a reduction of the energy consumption of the overall sensing process as well as an increase in the time available for transmission.

The interplay between sensing time, achievable throughput, and energy consumption for a multi-channel system is evaluated in [3]. As shown in Fig. 2a, energy efficiency first increases with the sensing time, due to more accurate spectrum sensing, and reaches a maximum value. After this point, energy efficiency falls, as the increased sensing performance cannot compensate for the increased energy consumption and for the decreased time available for transmission. The optimal sensing time for maximizing energy efficiency is higher than that for maximizing throughput, as it becomes more important to avoid false alarms and unnecessary additional sensing.

Secondary access without spectrum sensing (that is, zero sensing time) may improve the secondary throughput, if the primary system can tolerate some interference and the channel between the SU transmitter and PU receiver is

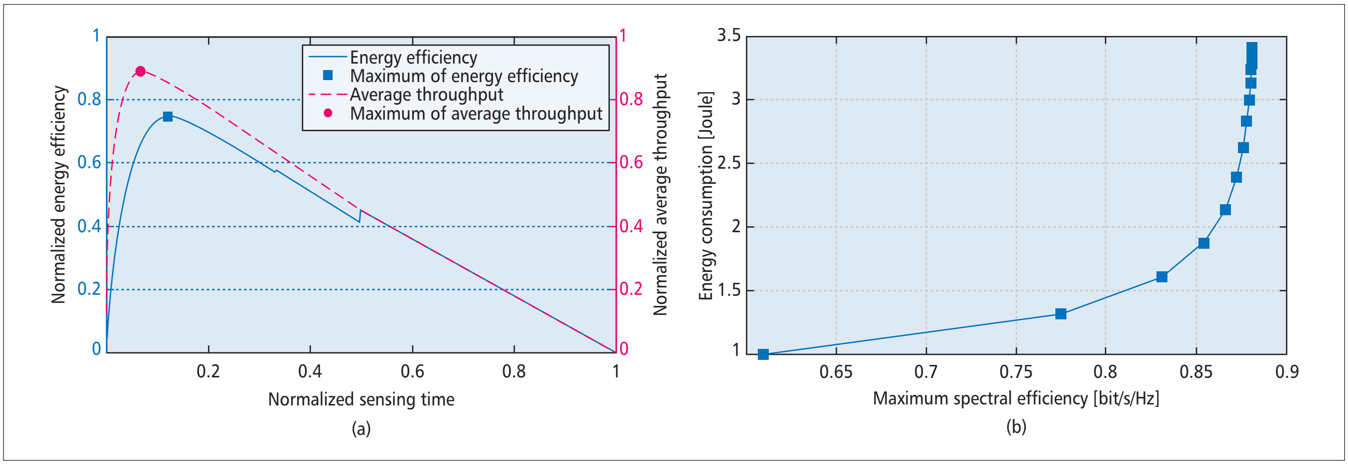


Figure 2. a) Energy efficiency versus sensing time in a multi-channel cognitive network with a single SU. The energy consumption, the denominator of the energy efficiency, is a non-continuous function, which causes the jump in the energy efficiency curve. b) Throughput-energy trade-off by increasing the maximum number of handoffs. Each dot corresponds to an increase of the number of handoffs. The energy consumption needs to be increased significantly to utilize all the available spectrum resources [3].

expected to be weak, as shown in [4]. As this scheme at the same time introduces higher packet loss in the secondary network with multiple uncoordinated SUs, its energy efficiency needs to be evaluated.

Waiting or Handoff

Once a PU returns to its channel, the SU may decide to wait until the channel becomes idle again, or invest some time and energy to start the spectrum handoff procedure and migrate to an idle channel. As [2] suggests, the decision should be based on the throughput and delay requirements of the SU. Unless the secondary quality requirements are very strict, optimizing the probability of waiting instead of migrating can decrease the energy consumption of the SU by 20 percent. Clearly, the gain decreases as the throughput or delay requirements get strict, and the SU cannot afford to simply wait for the new transmission opportunity in the current channel.

Given that a waiting SU needs to discover that the channel becomes idle again, the authors in [5] investigate how often the channel should be sensed. More frequent sensing allows the SU to start to transmit with lower discovery delay and thus achieve higher throughput, at the cost of higher sensing energy consumption. However, sensing does not need to be periodic. As shown in [5], the adaptation of the sensing interval, based on some knowledge of the PU busy time distribution, can reduce discovery delay by half, and thus increase secondary throughput, under the same sensing energy budget as periodic sensing.

Sensing Order

Under narrowband sensing, an SU sequentially senses the channels until an idle channel is found. The order of the channels to be sensed affects throughput and energy consumption. As a result of an improper sensing order, an SU may sense several channels to find a transmission opportunity, and thereby may suffer from higher energy consumption and shorter remaining transmission time. Therefore, hybrid spectrum handoff strategies are considered in [1], where the SU learns both the channel occupancy and the transmission channel quality statistics, and defines the sensing order accordingly. It is shown that optimizing based only on one of the above parameters can be highly sub-optimal, with a loss of energy efficiency in the range of 5 percent to 20 percent for the considered scenarios.

For example, primary traffic shaping as a consequence of applied network coding can increase the performance of learning the channel occupancy statistics, and can decrease the number of channels sensed until a transmission opportunity is found by as much as 50 percent, leading to significantly improved energy efficiency, as shown in [6].

Maximum Number of Handoffs

The performance of narrowband sensing depends not only on the sensing time of a single channel and on the sensing order, but also on the number of channels that should be sensed before the SU stops searching for a while. Clearly, allowing an SU to investigate more primary channels increases the chances of finding an empty channel, leading to throughput enhancement. However, as we see in Fig. 2b, the energy consumption cost of this increase can be significant, once the system is close to the throughput limit. For instance, to increase the throughput above 0.85, only 3 percent transmission rate enhancement is achieved by 81 percent more energy consumption, which devastates energy efficiency. This suggests that the maximum number of possible handoffs needs to be limited and the SU forced to wait, to improve the energy-throughput trade-off.

Sensing and Channel Access

In a secondary network with several uncoordinated SUs, finding an idle channel does not guarantee successful transmission. Here, all SUs may sense the popular primary channels (like the ones with low load and good transmission quality), and then compete for accessing the same channel, while other channels might be

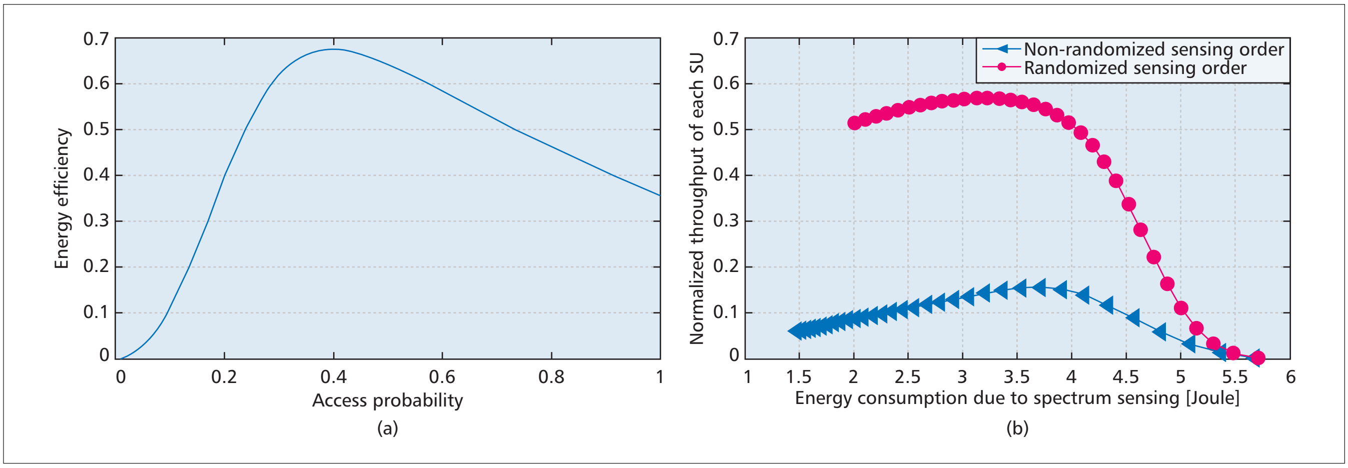


Figure 3. a) Energy efficiency as a function of secondary access probability. Optimizing the access probability can improve energy efficiency significantly; b) Secondary throughput as a function of invested sensing energy. The joint access probability and randomization of the channel sensing order achieves significant throughput gain [7].

idle. To solve this problem, [7] suggests to coupling sensing and channel access control, by introducing a randomized scheme whereby the SUs sense and then access the channels with some access probability. As shown in Fig. 3a, access probability has a significant effect on energy efficiency, due to the trade-off between throughput enhancement at more intentions to access the channels and the consequent increase of the contention level. The optimum access probability depends on the size of the secondary and primary networks. Significant further gain can be achieved by randomizing the order of the channels to be sensed, as shown in Fig. 3b, as it avoids potential constant scheduling conflicts among SUs. Hence, the joint optimization of sensing order and access strategy, based also on the channel occupancy statistics, is a logical next step [8]. However, this approach requires precise SU synchronization and extensive signalling, challenging its applicability in ad hoc settings.

QoS Control and Cooperation

If the primary throughput or delay requirements are not strict, some controlled secondary interference can be accepted at the primary receivers. In this case secondary sensing and channel access control solutions can be parameterized such that the primary packet loss is kept at an acceptable level. As shown in [4], such controlled interference can benefit the secondary network. Further gains can be achieved if interference and the consequent unsuccessful primary transmissions are compensated by cooperative relaying from the SUs. Therefore, the authors in [9] propose cooperation on the network layer, which imposes only low signaling overhead, resulting in up to 50 percent energy efficiency gain.

Cooperative Spectrum Sensing

In the case of strong primary signals, local sensing may be sufficient to ensure adequate sensing performance. However, the cooperation of several spectrum sensing devices, i.e. SUs in the area, is needed if the primary signal is weak, or if the radio propagation environment is harsh. Under cooperative sensing, the spatial diversity among the SUs mitigates the effect of link impairments due to fading and shadowing, and the SUs together can more effectively discover spectrum access opportunities. At the same time cooperative sensing introduces additional energy cost as local sensing results are reported to a central node or shared among the SUs in the area.

The design factors discussed for local spectrum sensing can also be optimized in cooperative sensing scenarios, considering the wireless environment of the individual SUs. However, there are additional open questions affecting energy efficiency for cooperating users, which we discuss in the following subsections.

Sensing Resource Allocation

Under cooperative sensing, the sensing resource is not only the sensing time but also the set of SUs that cooperate to discover a spectrum access opportunity. Increasing the number of cooperating SUs may decrease the required contribution of each of them, but may increase the overall energy consumption, or decrease the number of channels that can be sensed. As the discovered spectrum access opportunities are used by the SUs themselves that are discovering spectrum opportunities, the SUs now need to decide how large a share of the spectrum space, dedicated for secondary access, they want to utilize. On one side they may want to increase the number of sensed channels, so that there are more transmission opportunities to share. On the other side, this either requires increased sensing efforts from each SU or results in a decreased per channel sensing accuracy under a constraint on the sensing cost of an SU. Consequently, there is an optimal value of the sensed channels that maximizes per SU throughput or minimizes the average SU energy cost to achieve a transmission of a unit of data for each SU [10]. As shown in Fig. 4a, this optimal value depends on the network density. Moreover, as we see in Fig. 4b, the energy cost, even if minimized, strongly depends on the primary network quality requirements as well as the density of the secondary network. Networks

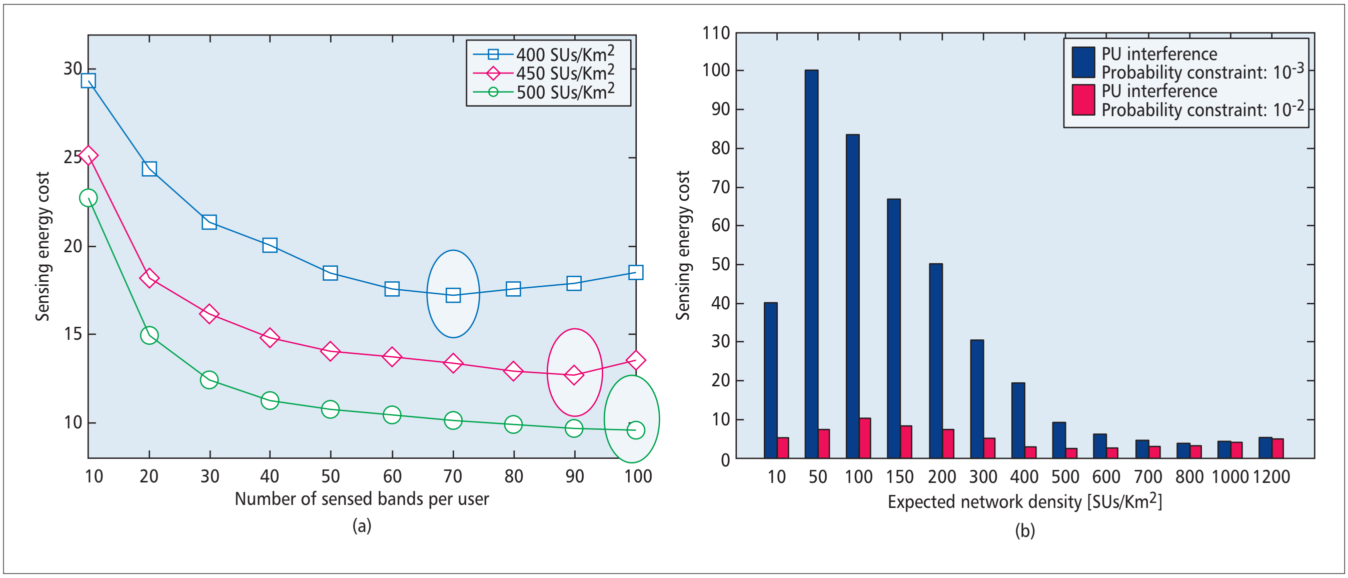


Figure 4. a) Average SU sensing energy cost per unit of data transmitted for each SU. The energy cost is minimized when the number of sensed channels is optimal. The optimum — indicated by a circle — depends on the SU density. A small number of bands results in a lower achievable SU throughput, while exceeding the optimal spectrum space results in lower sensing efficiency per band, thus higher cost per achievable SU throughput; b) The energy cost is lowest at optimal cognitive network density, above which the sensing performance improvement does not compensate for the increased demand for cognitive capacity [10].

with moderate density are worse off, where the cooperative sensing performance is moderate, but the gained access opportunities must be shared by a relatively large set of nodes. Increased network density improves energy efficiency significantly. Under very high densities the sensing energy cost increases again, as too many SUs need to share the low marginal sensing gain.

Sensing User Selection

As the number of users participating in cooperative sensing needs to be carefully selected, the remaining issue is which particular users should cooperate.

The authors in [11] propose an iterative solution to involve SUs in sensing, until the desired overall sensing performance in terms of misdetection and false alarm probabilities is met for all channels, with the objective of balancing sensing energy consumption. Clearly, the gain of optimized SU selection increases together with the number of available SUs, and therefore is important in dense secondary networks.

Given that the main reason for cooperative sensing is to mitigate fading and shadowing, the authors in [12] suggest that users experiencing uncorrelated link attenuation should be selected to cooperate. As shown in Fig. 5, the efficiency of this correlation-aware policy depends on the spatial distribution of the SUs. It can decrease the number of SUs required to sense the primary channel, and consequently the sensing energy cost, by more than 50 percent without affecting sensing accuracy.

Sensing Report Forwarding

Under cooperative sensing the local sensing results need to be reported, if a centralized architecture is implemented, or shared among the neighboring SUs in a distributed fashion. The reporting or sharing of local sensing results might require a significant amount of energy and perhaps time, particularly if high transmission power or multihop transmission is required. Therefore, the authors in [13] compare different approaches to determining the cooperating SUs, considering only the sensing time (TXT), the local sensing performance (SEM), the sensing result transmission cost (REM), or all of these (EE), with the objective to minimize the total required sensing energy cost for maintaining an overall sensing quality. As shown in Fig. 6, the gain of joint optimization is significant, if the sensing itself does not require much energy, i.e. in the high SNR regime.

As reporting the sensing results may have significant cost, the authors in [14] suggest that the SUs, even if included in cooperative sensing, should choose not to report the sensing results, if that might have little impact on the cooperative decision, while it would raise the overall reporting cost. The authors show that if the primary channel utilization statistics are a-priori known to the SUs, then the individual SUs can have a good estimate on the validity of their sensing results. In this case, censored reporting can drastically reduce the total sensing energy overhead by up to 40 percent, while the desired sensing performance is maintained.

In addition to introducing significant overhead to the overall energy cost of collaborative sensing, the reporting of the individual sensing results may impose a threat to the cooperative sensing performance due to the inherent lack of reliability of the wireless links used for reporting. As the authors in [15] demonstrate, the quality of cooperative decisions based on the individual decisions of the SUs (defined as hard decision) can degrade by up to 60 percent if the reporting links are unreliable. Instead, using cooperative decisions based on quantized raw sensing results (that is, soft decision combining) the overall

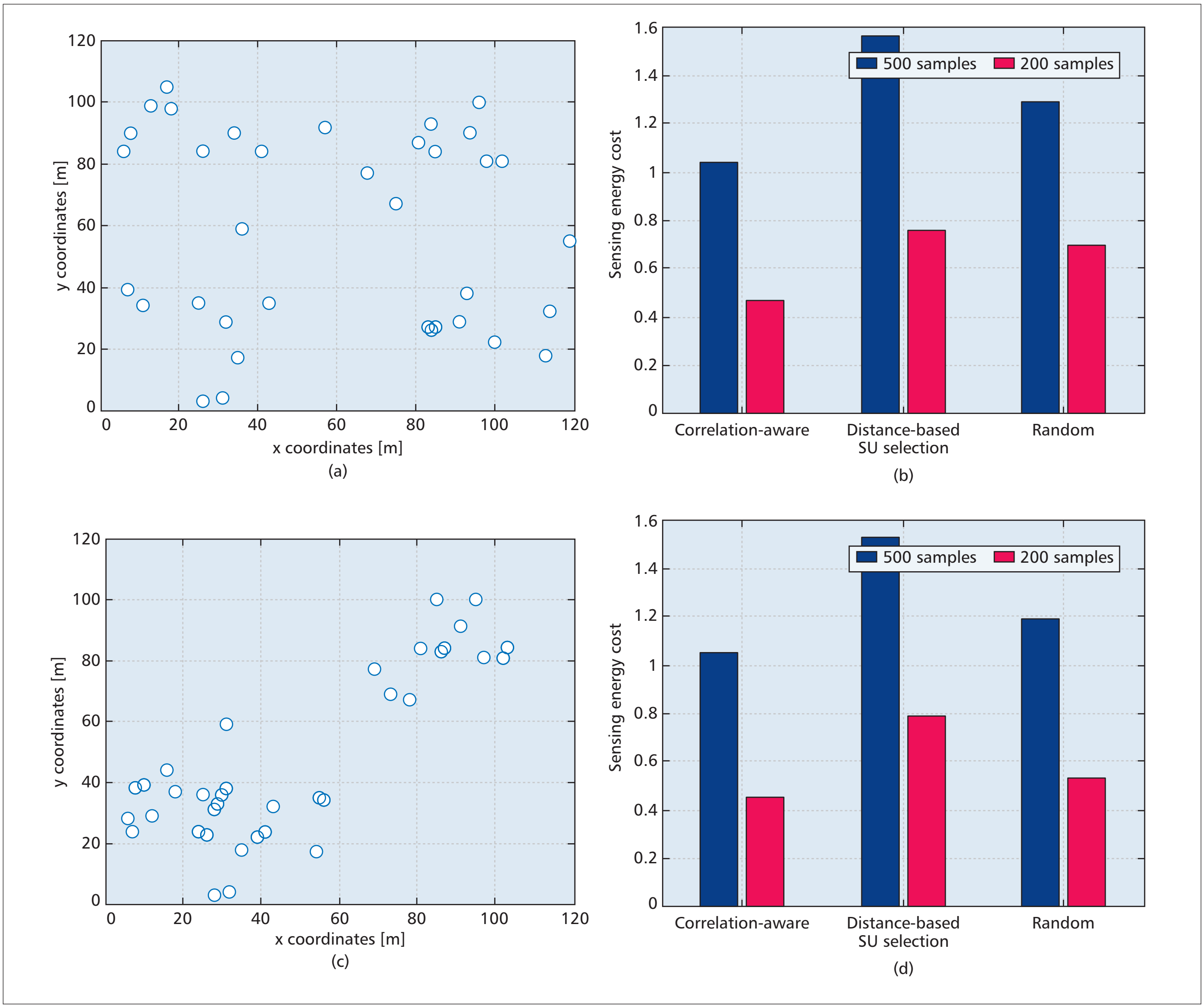


Figure 5. The energy cost per unit of SU throughput decreases when the correlation between SU channel measurements is taken into consideration in the iteration-based user selection algorithm. The improvement compared to random selection is smaller in d) since the nodes are located in disjoint geographical areas, as shown in c). The higher sensing accuracy, as a result of the increase in the sensing time per channel (500 samples), does not compensate for the linear increase in the sensing energy overhead [12].

sensing performance can be maintained at a relatively high level. The granularity of the reported sensing results needs to be tuned carefully to trade-off the energy cost and delay of reporting and the throughput gain due to correct spectrum decisions.

Open Issues

We have provided an overview of the most prominent mechanisms that aim to maximize the energy efficiency of spectrum sensing and handoff under local and cooperative sensing. Until now the main focus of these various works is to characterize the achievable gains of these mechanism, under different networking scenarios, as summarized in Table 1. However, to realize the predicted gains, several issues need to be addressed by the research community.

Energy Harvesting: Emerging architectures with energy harvesting from interfering wireless signals change the general assumption of homogeneous energy resources at the nodes. To utilize energy harvesting, both local and coordinated sensing schemes need to be extended to consider the temporally and spatially varying harvested energy.

Local Sensing under Dynamic Traffic: Most existing works consider SUs with saturated traffic and ideal wireless channel models (see [7] and references therein). However, real network traffic is bursty, which makes it challenging to achieve the benefits of learning based system optimization, due to the under-sampled or sparse network state information [8, 6, 5].

Coexistence of SUs under Local Sensing: As SUs performing local sensing may belong to different networks, they may have no means or incentive to coordinate, and may have significantly different traffic demands and performance

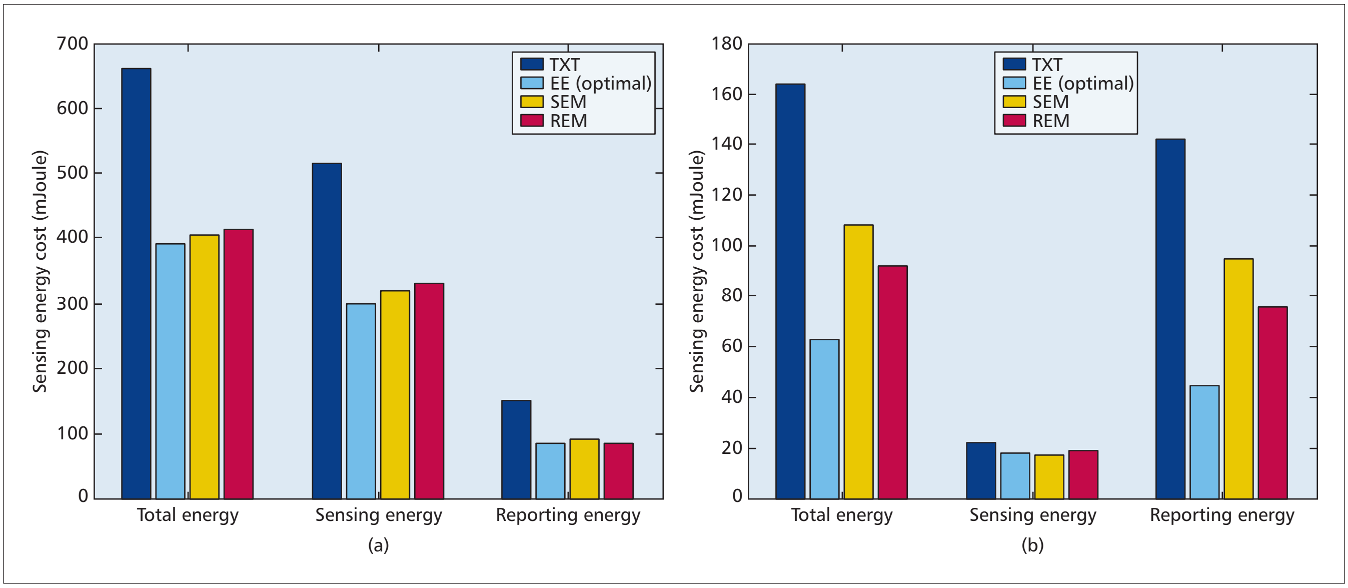


Figure 6. a) In the low SNR regime the sensing energy dominates the total energy consumption; b) It drops significantly for high SNR. The optimal SU selection scheme (EE) outperforms the heuristic solutions SEM and REM, which consider sensing energy cost and transmission costs respectively, and also TXT that minimizes the sensing time. The relative gain is more significant in high SNR regime, when sensing itself costs little energy [13].

objectives. To take this heterogeneity into account, sensing and channel access optimization [7] needs to be extended with learning, fairness, and incentive mechanisms.

Fair Cooperative Sensing: The optimization of the set of cooperating SUs, based on the sensing quality they can provide or the cost of communication [12, 13, 14], may inherently lead to unfair allocation of sensing burdens in cooperative systems. Future research is needed to evaluate whether this unfairness can be significant in fixed and mobile environments, and how the performance of the proposed schemes changes if fairness is enforced, e.g. considering a uniform sensing energy budget at the nodes, or contributions that are proportional to the needs of the individual SUs.

Cooperative Sensing Incentives: Incentives are necessary to avoid free-riders, and if possible achieve a social optimum. Incentive schemes need to be discussed considering short and long term objectives. On the short term, an SU may have incentive to cooperate if it has traffic to send and needs free spectrum. However, under dynamic traffic, long term incentives need to be considered to ensure that nodes cooperate, even if they do not have immediate gain.

Conclusions

Improving the energy-throughput trade-offs in spectrum sensing and access requires proper designs of the maximum number and the order of the primary channels sensed by an SU, the frequency of spectrum sensing, and the selection of the per-channel sensing time, in order to avoid wasting energy resources for a marginally higher throughput. In multi-channel scenarios, the selection of the number and order of the channels to be sensed becomes even more important. In cooperative sensing scenarios, allocation of sensing tasks to SUs with relatively good individual sensing and uncorrelated channel conditions substantially reduces energy consumption with negligible penalty in network throughput. Carefully reporting and combining individual sensing results, along with allocating sensing tasks to SUs with low-cost reporting links, increases overall sensing and thereby energy efficiency.

References

[1] Y. Pei *et al.*, "Energy-Efficient Design of Sequential Channel Sensing in Cognitive Radio Networks: Optimal Sensing Strategy, Power Allocation, and Sensing Order," *IEEE JSAC*, vol. 29, no. 8, Sept. 2011, pp. 1648–59.

[2] S. Wang *et al*., "Energy-Efficient Spectrum Sensing and Access for Cognitive Radio Networks," *IEEE Trans. Veh. Technol*., vol. 61, no. 2, Feb. 2012, pp. 906–12.

[3] H. Shokri-Ghadikolaei, Y. Abdi, and M. Nasiri-Kenari, "Analytical and Learning-Based Spectrum Sensing Time Optimization in Cognitive Radio Systems," *IET Commun*., vol. 7, no. 5, Mar. 2013, pp. 480–89.

[4] A. Fanous and A. Ephremides, "Access Schemes for Mitigating the Effects of Sensing Errors in Cognitive Wireless Networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 6, Jun. 2014, pp. 3343–52.

[5] C. Zhang and K. Shin, "What Should Secondary Users Do Upon Incumbents' Return?" *IEEE JSAC*, vol. 31, no. 3, Mar. 2013, pp. 417–28.

[6] A. Fanous, Y. Sagduyu, and A. Ephremides, "Reliable Spectrum Sensing and Opportunistic Access in Network-Coded Communications," *IEEE JSAC*, vol. 32, no. 3, Mar. 2014, pp. 400–10.

[7] H. Shokri-Ghadikolaei and C. Fischione, "Analysis and Optimization of Random Sensing Order in Cognitive Radio Networks," *IEEE JSAC*, vol. 33, no. 5, May 2015, pp. 803–19.

[8] Y. Song and J. Xie, "Prospect: A Proactive Spectrum Handoff Framework for Cognitive Radio Ad Hoc Networks Without Common Control Channel," *IEEE Trans. Mobile Comput*., vol. 11, no. 7, Jul. 2012, pp. 1127–39.

[9] M. Costa and A. Ephremides, "Energy Efficiency in Cooperative Cognitive Wireless Networks," *Proc. IEEE Conf. Information Sciences and Systems (CISS)*, Mar. 2014, pp. 1–6.

[10] I. Glaropoulos and V. Fodor, "Spectrum Sharing with Low Power Primary Networks," *Proc. IEEE Int'l. Symposium on Dynamic Spectrum Access Networks (DYSPAN)*, Apr. 2014, pp. 315–26.

[11] S. Althunibat, R. Palacios, and F. Granelli, "Energy-Efficient Spectrum Sensing in Cognitive Radio Networks by Coordinated Reduction of the Sensing Users," *Proc. IEEE Int'l Conf. Communications (ICC)*, Jun. 2012, pp. 1399–1404.

[12] A. Cacciapuoti, I. Akyildiz, and L. Paura, "Correlation-Aware User Selection for Cooperative Spectrum Sensing in Cognitive Radio Ad Hoc Networks," *IEEE J. Select. Topics Signal Process.*, vol. 30, no. 2, Feb. 2012, pp. 297–306.

[13] S. Eryigit, S. Bayhan, and T. Tugcu, "Energy-Efficient Multichannel Cooperative Sensing Scheduling with Heterogeneous Channel Conditions for Cognitive Radio Networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 6, Jul. 2013, pp. 2690–99.

[14] S. Maleki, A. Pandharipande, and G. Leus, "Energy-Efficient Distributed Spectrum Sensing for Cognitive Sensor Networks," *IEEE Sensors J.*, vol. 11, no. 3, Mar. 2011, pp. 565–73.

[15] S. Chaudhari *et al.*, "Cooperative Sensing with Imperfect Reporting Channels: Hard Decisions or Soft Decisions?" *IEEE Trans. Signal Process.*, vol. 60, no. 1, Jan. 2012, pp. 18–28.

Biographies

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

Ioannis Glaropoulos is an embedded software developer with Yanzi Networks AB, Sweden. He received the diploma in electrical and computer engineering from Aristotle University of Thessaloniki, and the M.Sc. and Ph.D. degrees in electrical engineering from KTH Royal Institute of Technology, Sweden, in 2005, 2008, and 2015, respectively. His current research interests include protocol design for low-power wireless networks and network protocol optimization for application scenarios within the Internet of Things.

Viktoria Fodor received the M.Sc. and Ph.D. degrees in computer engineering from the Budapest University of Technology and Economics, Hungary, in 1992 and 1999, respectively. She has held research positions at Polytechnic University of Turin (1994) and Boston University (1995). Since 1999 she has been with KTH Royal Institute of Technology, Sweden, where she is now an associate professor with the Laboratory for Communication Networks. Her research interests include performance evaluation, cognitive and cooperative communication, and protocol design.

Carlo Fischione is an tenured associate professor at KTH Royal Institute of Technology, Sweden. He received the Ph.D. degree in electrical and information engineering and the laurea degree in electronic engineering from the University of L'Aquila, Italy. He has held research positions at Massachusetts Institute of Technology (2015), Harvard University (2015), and the University of California at Berkeley (2004 and 2007-2008). His research interests include optimization with applications to wireless networks and cyber-physical systems.

Anthony Ephremides has been with the University of Maryland since 1971. He is interested in problems relating to wireless communications, networking, optimization, and all aspects of communication and control systems. He has mentored more than 40 doctoral students and has had extensive consulting and collaborative activities in research worldwide. His motto is "never tire or retire."

Advertisers' Index

COMPANY	PAGE
IEEE ComSoc BlackSeaCom	5
IEICE	23
IEEE Member Digital Library	Cover 4
IEEE Sales & Marketing	Cover 3
Keysight	Cover 2
Keysight Tutorial	3
National Instruments	1

ADVERTISING SALES OFFICES

Closing date for space reservation: 15th of the month prior to date of issue

National Sales Office
James A. Vick
Sr. Director Advertising Business, IEEE Media
Email: jv.ieeemedia@ieee.org

Marion Delaney
Sales Director, IEEE Media
Email: md.ieeemedia@ieee.org

Mark David
Sr. Manager Advertising & Business Development
Email: m.david@ieee.org

Mindy Belfer
Advertising Sales Coordinator
Email: m.belfer@ieee.org

Northern California
George Roman
Tel: (702) 515-7247
Fax: (702) 515-7248
Email: George@George.RomanMedia.com

Southern California
Marshall Rubin
Tel: (818) 888 2407
Fax:(818) 888-4907
Email: mr.ieeemedia@ieee.org

Mid-Atlantic
Dawn Becker
Tel: (732) 772-0160
Fax: (732) 772-0164
Email: db.ieeemedia@ieee.org

Northeast
Merrie Lynch
Tel: (617) 357-8190
Fax: (617) 357-8194
Email: Merrie.Lynch@celassociates2.com

Jody Estabrook
Tel: (77) 283-4528
Fax: (774) 283-4527
Email: je.ieeemedia@ieee.org

Southeast
Scott Rickles
Tel: (770) 664-4567
Fax: (770) 740-1399
Email: srickles@aol.com

Midwest/Central Canada
Dave Jones
Tel: (708) 442-5633
Fax: (708) 442-7620
Email: dj.ieeemedia@ieee.org

Midwest/Ontario, Canada
Will Hamilton
Tel: (269) 381-2156
Fax: (269) 381-2556
Email: wh.ieeemedia@ieee.org

Texas
Ben Skidmore
Tel: (972) 587-9064
Fax: (972) 692-8138
Email: ben@partnerspr.com

Europe
Christian Hoelscher
Tel: +49 (0) 89 95002778
Fax: +49 (0) 89 95002779
Email: Christian.Hoelscher@husonmedia.com