October 2015, Vol. 53, No. 10

IEEE Communications MAGAZINE

www.comsoc.org

Military Communications

Also in This Issue

- Future Railway Communications
- Social Networks Meet Next Generation Mobile Multimedia Internet
- Integrated Circuits for Communications

IEEE COMMUNICATIONS SOCIETY

A Publication of the IEEE Communications Society

IEEE Communications MAGAZINE

THANKS OUR CORPORATE SUPPORTERS

SAMSUNG

Anritsu
Test and Measurement Solutions

BEEcube

CISCO™

ROHDE&SCHWARZ

KEYSIGHT TECHNOLOGIES

October 2015, Vol. 53, No. 10

IEEE Communications MAGAZINE

www.comsoc.org

Military Communications

Also in This Issue

- Future Railway Communications
- Social Networks Meet Next Generation Mobile Multimedia Internet
- Integrated Circuits for Communications

IEEE COMMUNICATIONS SOCIETY

A Publication of the IEEE Communications Society

Your 5G Eureka moment will happen sooner or later.

We'll help make it sooner.

The fifth generation of wireless communications may seem years away. But if you want to be on the leading edge, we'll help you gain a big head start. We offer unparalleled expertise in wideband mmWave, 5G waveforms, and Massive MIMO. We also offer the industry's most comprehensive portfolio of 5G solutions. Whether you need advanced antenna and radio test hardware or early simulation software, we'll help you with every stage of 5G.

HARDWARE + SOFTWARE + PEOPLE = 5G INSIGHTS

PEOPLE

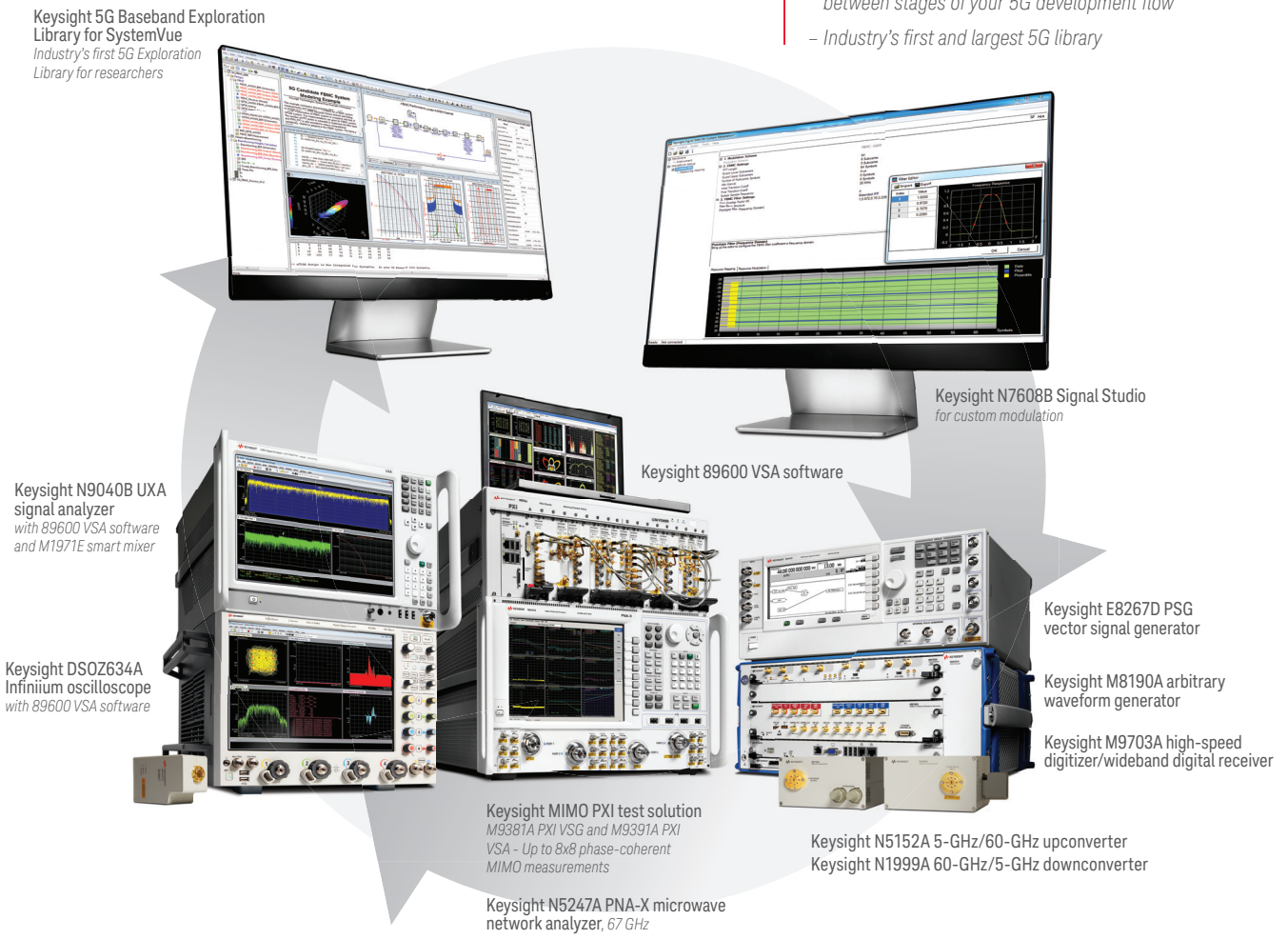
- *Keysight engineers are active in the leading 5G forums and consortia*
- *Keysight engineers are keynote speakers at 5G conferences and key contributors in top technical journals*
- *Applications engineers are in more than 100 countries around the world*

Download our white paper *Implementing a Flexible Testbed for 5G Waveform Generation and Analysis* at www.keysight.com/find/5G-Insight

USA: 800 829 4444 CAN: 877 894 4414

HARDWARE + SOFTWARE

- *Designed for testing 5G simulation to verification*
- *Software platforms and applications that work seamlessly across our 5G instruments*
- *Incorporate iterative design and rapidly move between stages of your 5G development flow*
- *Industry's first and largest 5G library*



KEYSIGHT TECHNOLOGIES

Unlocking Measurement Insights

IEEE Communications MAGAZINE

OCTOBER 2015, Vol. 53, No. 10

www.comsoc.org/commag

BEEcube

A National Instruments Company

FPGA Based Rapid Prototyping Platforms for Military Communications

BEE7

Off-the-shelf communicators platform for high level milcom research with full speed network I/O up to 1.1 Tbps throughput

RF Expansion Cards

	FMC 104	FMC 105	FMC 106	FMC 107	FMC 108	FMC 109	FMC 111
Applications	E-band radio		WLAN		LTE Advanced		SDR
Sampling	Direct RF or IF		MIMO Baseband & IF		Direct RF		Tunable RF
Frequency Range (MHz)	5 - 2500		0.4 - 500	4.5 - 1500	4.5-3000 1st Nyquist 650-4000 2nd Nyquist		400 - 6000
Function	ADC	DAC	ADC	DAC	ADC	DAC	RF
Channels	1, 2 or 4	Single	Quad		1 or 2	Dual	2x2 MIMO
Rate (Gsps)	5 Single 2.5 Dual 1.25 Quad	5	0.5	1.5	4 Single 2 Dual	5.6 with 2x Interpolation	0.25 or 0.125 ADC 1 DAC
Resolution	10	12	14	16	12	14	14 or 16
Bandwidth	2.5 GHz		250 MHz	750 MHz	1.0/2.0 GHz	1.4 GHz	80 or 40 MHz

Join us at MILCOM in Tampa, Florida for a live demonstration of our platforms

www.BEEcube.com

BEEcube

A National Instruments Company

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

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

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

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

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

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

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

CURRENTLY SCHEDULED TOPIC

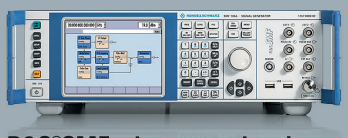
	PUBLICATION DATE	MANUSCRIPT DUE DATE
WIRELESS COMMUNICATIONS, NETWORKING, AND POSITIONING WITH UNMANNED AERIAL VEHICLES	MAY 2016	NOVEMBER 1, 2015
BIO-INSPIRED CYBER SECURITY FOR COMMUNICATIONS AND NETWORKING	JUNE 2016	NOVEMBER 1, 2015
LTE EVOLUTION	JUNE 2016	NOVEMBER 30, 2015
WIRELESS TECHNOLOGIES FOR 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

Explore the limits.
T&M solutions for aerospace and defense.

Today's aerospace and defense technologies demand ever more sophisticated test and measurement solutions to stretch the limits of what is feasible. As a full-range supplier, Rohde&Schwarz offers a broad portfolio that proves its capabilities in even the most demanding applications. Our leading-edge expertise in microwave, RF and EMC helps customers assess performance, optimize platforms and get the most out of systems. Convince yourself.

www.rohde-schwarz.com/ad/sat/smf



R&S®SMF microwave signal generator

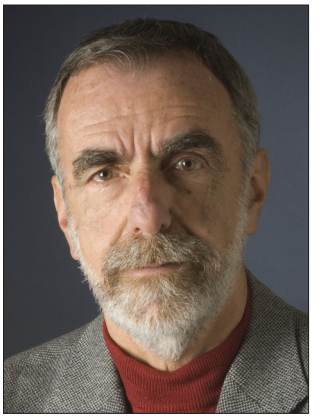
- Frequency range (generator) up to 43.5 GHz
- Frequency multipliers up to 110 GHz with adjustable output levels
- Excellent spectral purity, e.g. typ. –120 dBc (1 Hz) at 10 GHz, 10 kHz offset
- High output power, e.g. typ. +25 dBm at 20 GHz
- Flexible pulse generation for radar applications
- Easy replacement of legacy instruments

ROHDE&SCHWARZ

A REVIEW OF ACTIVITIES IN MEMBER RELATIONS AND PUBLICATIONS

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 second page, October 2015, is coauthored by Katie Wilson, Stefano Bregni, and myself, and summarizes the activities in the areas of Publications and Member Relations.



SERGIO BENEDETTO

STEFANO BREGNI

KATIE WILSON

THE MEMBER RELATIONS COUNCIL (TO BE RENAMED THE MEMBER AND GLOBAL ACTIVITIES COUNCIL IN 2016)

According to the current Bylaws of the IEEE Communications Society, the Member Relations Council (MRC) addresses all activities and programs related to members, chapters, membership development, marketing, industry relations, sister and related societies, and Society regions.

Chaired by the Vice-President for Member Relations, it also includes seven Directors, each chairing his/her own Board, viz. Marketing and Industry Relations, Membership Programs Development, Sister and Related Societies, and the four Regions (Asia Pacific; Europe, Africa, Middle East; Latin America; and North America).

The seven Directors who have been serving in this term 2014-15, now approaching its end, are:

- Director M&IR: Ashutosh Dutta, Columbia Univ., USA
- Director MPD: Koichi Asatani, Kogakuin Univ., Japan
- Director SRS: Curtis Siller, USA
- AP Director: Wanjiun Liao, National Taiwan Univ., ROC
- EAME Director: Hanna Bogucka, Poznan Univ. of Technology, Poland
- LA Director: Pedro Aguilera, Switch Comunicaciones Ltda, Chile
- NA Director: Merrily Hartmann, USA

In 2015 the ComSoc Strategic Committee discussed and agreed on a thorough revision of the leadership organization of our Society, which was approved by the Board of Governors last June. Effective from 2016, the MRC has been renamed as the Member and Global Activities Council (MGAC) and reorganized along a slightly different structure.

The focus of the MGA-C will be more specifically centered on member needs and interests. It will be chaired by the Vice President for Member and Global Activities (VP-MGA). The following appointed officers will report to the VP-MGA: Director for Member Services; Director for Sister and Related Societies; Director of Asia/Pacific Region; Director of Europe, Middle-East and Africa Region, Director of Latin America Region; Director of North America Region; Chair of the Women in Communications Engineering Standing Committee.

Strategies of the Member Relations Council

Beginning my term as VP-MR in 2014, I indicated five strategic directions for the ComSoc Member Relation Council, which I envisioned as the Golden Pentagon: globalization, academia, industry, women, and students (see figure next page). During these last two years, several initiatives have been launched or revamped along those five directions.

Globalization means incorporating global culture and values in ComSoc, on the one hand by strengthening balanced participation of members from all regions, on the other by disseminating knowledge and technical skills in all countries worldwide, paying extra attention to disadvantaged areas.

Aiming at both industry and academia implies creating multidisciplinary linkages between ComSoc, universities, and industry. As a matter of fact, currently the participation from universities in ComSoc technical activities is largely prevailing over industry. Restoring a balanced participation from both industry and academia within ComSoc is a priority, while not disappointing the expectations of university students and professors, who rely on ComSoc publications and conferences as their primary reference.

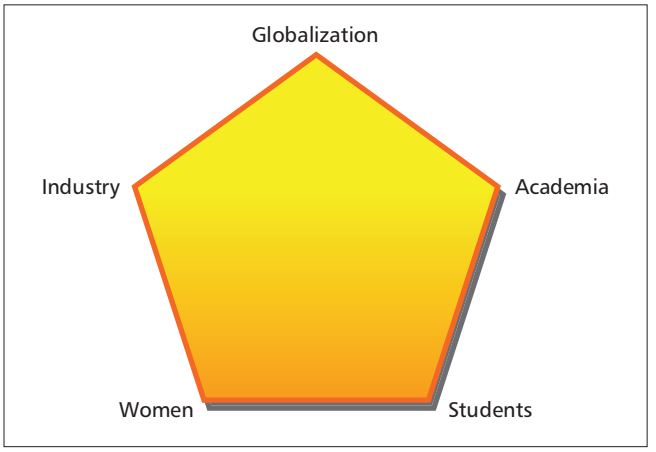
Due to a number of historical reasons, women are significantly underrepresented in most engineering disciplines and senior leadership positions, although to varying degrees in different countries. Communications engineering is not an exception. In our Society, significant effort has been made to grant true equal opportunities to genders.

Students truly represent the future of our Society. Therefore, one of our most important strategic directions has been toward promoting initiatives to involve young students in ComSoc activities. Online education, special programs dedicated to students (e.g. the Student Competition Award, now running its third edition), ad-hoc activities organized for students at major ComSoc conferences, and Student Travel Grants, have been pursued to encourage participation by the young generation and to improve the retention of student members. Special attention should be given to young students living in disadvantaged areas of the world, who have fewer opportunities to meet international experts at major conferences. Online webinars and the Distinguished Lecturer Program are two key programs to also serve this important social mission of ComSoc.

During 2014-15, the Member Relations Council worked cohesively along these five strategic directions. After having been re-elected for the next term 2016-17, I can ensure that we will continue to pursue these goals in the next two years. I wish to take this opportunity to thank all ComSoc members who trust in me. Your support is essential.

Chapters and Regional Activities

There are currently 211 ComSoc Chapters worldwide! Specifically, 41 are in the AP Region, 52 in the EMEA Region, 25 in the LA Region, and 93 in the NA Region. The



The Golden Pentagon of ComSoc's Member Relations.

oldest Chapter, according to IEEE records, is the Mohawk Valley Section (NA), formed in 1955. In conrast, the newest Chapter as of this writing is the Windsor Joint Section in Canada, just formed in August 2015. Welcome!

Some Chapters are very active and succeed in organizing high-impact initiatives reaching out to both industry and academia. Others struggle with a small base of local members. In response, we started an effort aimed at identifying and contacting less active Chapters, in order to support Chairs to analyze the situation.

The Regional Chapter Chair Congresses (RCCC) are the chief opportunity for Chapter Chairs to meet, compare experiences, and express their wishes to ComSoc leaders. The North-America RCCC was organized at IEEE GLOBECOM 2014 in Austin. I had the pleasant opportunity to meet 20 Chapter Chairs, who are enthusiastic and strongly motivated volunteers who spend a significant share of their time to serving ComSoc.

Regional Conferences: Key to Pursue Globalization

Our priority is to encourage participation of members from all regions, striving to reach out to all countries and paying extra attention to disadvantaged areas. ComSoc Regional Conferences are instrumental in this strategy. They may be the only opportunity for members in disadvantaged areas to attend first-level international events, saving the cost of travelling overseas.

Certainly, the key is to ensure that Regional Conferences meet the same quality requirements that are established for the global portfolio conferences of ComSoc. The ComSoc Conference Council is well focused on this issue and ensures, among other tasks, that the Technical Program Committees of Regional Conferences include qualified members from around the world and cover all technical areas of ComSoc (in particular, nominated by the Technical Committees).

Among ComSoc Regional Conferences, I wish to mention in particular the case of the IEEE Latin-American Conference on Communications (IEEE LATINCOM). The sixth edition was held successfully in Cartagena de Indias, Colombia in November 2014, attracting authors and attendees from around the world, thanks to a high-quality technical program and world-class keynote speakers (Maurizio Decina, Hikmet Sari, Neeli Prasad, Edward Knightly). The seventh edition will be held in Arequipa, Peru. IEEE LATINCOM has now become a regular meeting venue for students and engineers in Latin America.


GCN GLOBAL COMMUNICATIONS NEWSLETTER

June 2015
ISSN 2374-1082

REGIONAL REPORT

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

By Sattar Bader Sadkhan, Chair of IEEE Iraq Section

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

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

AN OVERVIEW OF IEEE IRAQ COMSOC CHAPYTER ACTIVITIES IN IRAQ

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

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

(Continued on Newsletter page 4)

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

June 2015 Global Communications Newsletter 1

The new *Global Communications Newsletter*: The Voice of the Chapters.

Other ComSoc Regional Conferences are the IEEE Asia-Pacific Conference on Communications (APCC), whose 21st edition was held in Kyoto, Japan in October 2015, and the most recent IEEE BlackSeaCom, whose third edition was held in Constanta, Romania, in May 2015.

The *IEEE Global Communications Newsletter*: The Voice of the Chapters

In our Member Relations strategy, the *IEEE Global Communications Newsletter* definitely has a central role. After many years of highs and lows and various fortunes, we succeeded in revamping it, making it more appealing and better known by ComSoc members. My aim, since the beginning of this term, has been to truly make it the "Voice of the Chapters." Four lines of action have been pursued:

1) Each GCN monthly issue is distributed by email to all Chapter Chairs as soon as it is published (since July 2014), in addition to being printed in *IEEE Communications Magazine* and posted to the GCN web site for public access, which has been done since its beginning. Chapter Chairs are encouraged to distribute it further to local members. This ensures a thorough and pervasive distribution to all membership, as well as fostering a closer relationship between Chapter Chairs and "their" members.

2) Significant effort was made to make its contents more interesting, by selecting events worldwide and inviting specific articles. A new series of eight monthly interviews by the VP with MR Directors and Committee Chairs was published from September 2014 to April 2015. A second such series will be launched in 2016.

3) The graphic layout was fully redesigned, giving the GCN a fresh and contemporary look (since October 2014). The new layout follows the IEEE Visual Identity Guidelines and is definitely more appealing for readers.

4) A new Online Edition of the GCN is under development and will be launched soon. The Online GCN will feature a striking contemporary style. The contents of the Online GCN will include:

- The same articles published in the monthly printed issue.
- A blog where Chapter Chairs will post their announcements any time they have updates.
- A forum restricted to Chapter Officers, i.e. a confidential venue where Officers can discuss freely and exchange informally their experiences and ideas. Links to social media will facilitate networking among members.

The Distinguished Lecturer Program: A Member Favorite

Among ComSoc's programs for Chapters and Members, the Distinguished Lecturer/Speaker Program is definitely one of the most acclaimed by members! Personally, I have a special fondness for this program, because I served as a Distinguished Lecturer for seven years, visiting 14 countries and 29 IEEE Sections/Chapters in all Regions, some more than once. It was one of the most rewarding experiences of my life, and certainly my ComSoc service I am most proud of.

In 2014 the Member Relations Council made an important decision. We extended the travel cost limit refundable by ComSoc from US$2000 to US$3000 (with the same total budget per year) thus facilitating longer tours, ultimately looking at those areas that, because of their geographic location, are seldom visited by lecturers. Our motto has been "Even less DLTs, but better DLTs!."

The Student Competition Award

In 2015 we are conducting the third ComSoc Student Competition, entitled "Communications Technology Changing the World," addressed to graduate and undergraduate students. I am the Co-Chair of this program with the past VP-MR Nelson Fonseca. This Competition is now a well established program of the Comsoc Member Relations Council and will be continued in future years.

The Competition has been established as a great success. In 2014 more than 70 projects were submitted and evaluated by an international committee made up of 40 ComSoc experts. The evaluation process takes into account social impact, technical content, originality, practical applicability and results, and quality of presentation. The winners receive their Award at GLOBECOM in a Plenary Session.

Women in Communications Engineering (WICE)

The mission of the WICE Standing Committee, chaired by Octavia Dobre, Memorial University of Newfoundland, Canada, is to promote the visibility and roles of women communications engineers. Among its major initiatives in 2014-15, I make note of: the institution of WICE Awards; the organization of a Women's Workshop and of social events at ComSoc flagship conferences with speakers and panels; a new web site and Facebook and LinkeIn pages to encourage networking; and finally, a mail list with more than 4000 subscribers.

I strongly support these efforts and I cordially invite all women communications engineers to follow Octavia and stay tuned for news!

Industry Outreach Programs

Industry is one of the five vertices of the Pentagon referenced earlier. Restoring a balanced participation from both industry and academia within ComSoc is our priority, for example by regaining the focus of membership program development on industry, or by designing new programs and services specifically oriented to industry.

An important area of high potential interest for industry is education and professional training. ComSoc already explored this area with the Wireless Communications Engineering Technology (WCET) education program, started a few years ago.

A significant initiative aimed at industry outreach is the organization of high-impact one-day summits focused on industry interests. The First IEEE International 5G Summit was held at Princeton University in 2015 (www.5gsummit.org), chaired by Ashutosh Dutta, Director of Marketing and Industry Relations. Approximately 300 engineers attended, with 17 talks given by industry, academic, and government leaders. More 5G Summits are being planned and organized to consolidate this success and ComSoc's leading presence in this area.

Sister and Related Societies

Sister Societies are national or international professional associations focused on communications engineering or equivalent disciplines, with charters similar to ComSoc, and with similar technical scopes. Examples of ComSoc Sister Societies are the IEICE in Japan, the Popov Society in Russia, the KIKS in South Korea, the VDE in Germany, and many others too numerous to list.

Related Societies are national or international professional associations with charters similar to ComSoc's, but complementary in technical scope. Examples of Related Societies are other IEEE Societies, the Association for Computing Machinery (ACM), the International Telecommunication Union (ITU), the Optical Society of America (OSA), and others.

In 2014-15, the SRS Board made significant progress in furthering relations between Sister and Related Societies and Comsoc. The templates for new "Memoranda of Understanding" were refined. Approximately 15 agreements were renewed. Several new initiatives were launched, including co-sponsoring events, inviting SRS to publicize their events in the *IEEE Communication Magazine* Conference Calendar, inviting SRS members to register for our portfolio conference events, and providing local SRS officers with a complimentary one-day conference registration.

Biography of Stefano Bregni

Stefano Bregni is an associate professor of telecommunications at Politecnico di Milano, Italy. He graduated in electronics engineering in 1990. After nine years in industry he joined Politecnico in 1999.

Stefano is an IEEE Senior Member (1999). He was a IEEE Distinguished Lecturer for seven years (2003-2009), visiting 14 countries and 29 IEEE Sections/Chapters worldwide. In ComSoc he has served as: Vice-President Member Relations (two terms: 2014-15 and 2016-17); Member at Large on the Board of Governors (two terms: 2010-12 and 2013); Director of Education (2008-11); Chair of the Transmission, Access and Optical Systems TC (2008-09; Secretary/Vice Chair 2002-07); Member at Large of the GLOBECOM/ICC Technical Content (GITC) Committee (2006-09). He received the 2014 ComSoc Hal Sobol Award for Exemplary Service to Meetings & Conferences.

He is or has been Technical Program Co-Chair of ICC 2016, Technical Program Vice-Chair of GLOBECOM 2012, Symposia Chair of GLOBECOM 2009, Symposium Co-Chair in nine other ICC/GLOBECOMs, and Technical Program Co-Chair of IEEE LATINCOM 2011. He is the Editor-in-Chief of the *IEEE Global Communications Newsletter*, published

monthly in *IEEE Communications Magazine*, and an associate editor of *IEEE Communications Surveys and Tutorials*. He has contributed to ETSI/ITUT standards on network synchronization. He is the author of more than 80 refereed papers and of the book *Synchronization of Digital Telecommunications Networks* (Wiley, 2002).

PUBLICATIONS

The Publications team comprises the VP of Publications, the Director of Journals (Len Cimini), the Director of Magazines (Steve Gorshe), and the Director of Online Content (Elena Neira). It has been a busy term and we are delighted to share our work with you in this President's Page.

Every five years IEEE publications are reviewed by the IEEE Periodical Review and Advisory Committee (PRAC). All wholly owned ComSoc journals and magazines were reviewed in November 2014. The journals under review were: *IEEE Communications Letters* (George Karagiannides, Editor-in-Chief), *IEEE Communications Surveys and Tutorials* (Ekram Hossain, Editor-in-Chief), *IEEE Journal on Selected Areas in Communications* (Muriel Medard, Editor-in-Chief), *IEEE Transactions on Communications* (Robert Schober, Editor-in-Chief), and *IEEE Transactions on Network and Service Management* (Rolf Stadler, Editor-in-Chief). The magazines under review were: *IEEE Communications Magazine* (Sean Moore, Editor-in-Chief), *IEEE Network Magazine* (Sherman Shen, Editor-in-Chief), and *IEEE Wireless Communications Magazine* (H.H. Chen, Editor-in-Chief).

The PRAC review is an opportunity for publications to reflect on the health and future of the journals and magazines. We had the opportunity to receive helpful suggestions from the PRAC, e.g. updating official scopes that still have twentieth century research terms such as fax and dial-up modems. In general, the PRAC found that ComSoc publications are not only helpful but exemplary. We are very proud of the work of the volunteers, editors, and authors who produce the excellent publications that are a jewel in ComSoc's crown.

In the past two years we also initiated the incubation of two new publications: *IEEE Journal on Selected Areas in Communications: Green Communications and Networking Series* (Editor-in-Chief, Ender Ayonoglu), and the *IEEE Communcations Standards Supplement* (Editor-in-Chief, Glenn Parsons). The *Green Communications and Networking Series* has been very healthy, generating more than 100 submissions per issue. Given the continued success of this series, we plan to propose a stand-alone journal on Green Communications. The *IEEE Communications Standards Supplement* is a joint effort with the VP of Standards, Rob Fish, and the Director of Standards Development, Alex Gelman. A new *Communications Standards Magazine* based on this supplement is now being proposed. In addition, two new publications (co-sponsored with other societies) were launched: *IEEE Transactions on Cognitive Communications and Networking* (Michele Zorzi, Editor-in-Chief), and *IEEE Transactions on Molecular, Biological and Multiscale Communications* (Urbashi Mitra, Editor-in-Chief). If you have not checked out these publications, please do.

Online Content has blossomed in the past two years. Best Readings (http://www.comsoc.org/best-readings) was started in 2010. Sergio Benedetto and Len Cimini were the driving force behind it, and currently Matthew Valenti is responsible for the content and delivery of Best Readings. Best Readings is a collection of articles and books by experts in the field and is the first place one should go when researching a new topic. Len, Matt, and Elena Neira have worked hard to improve the content and look of Best Readings and ensure a good set of topics for our community. *Communications Technology News* (http://www.comsoc.org/ctn) presents a set of high-quality articles on relevant areas in communications with a rotating set of Editors-in-Chief. In 2014 Steve Weber did an excellent job of collecting and editing a set of high-quality and intriguing articles on topics such as data pricing and network neutrality. Recently Alan Gatherer has curated an interesting and must-read series of articles on 5G. Elena Neira also introduced a series of fascinating interviews online called ComSoc Beats (http://beats.comsoc.org). ComSoc Beats has compelling interviews with communications colleagues on topics such as future research, future directions of the society, and what it's like to be a communications engineer. These videos are well-done and thought-provoking.

IEEE Communications Society publications continue to thrive, grow, and expand into new areas and media. I am confident that they will continue to thrive under the stewardship of the incoming VP-Publications, Nelson Fonseca, and I look forward to their future.

Biography of Sarah Kate Wilson

Sarah Kate Wilson earned her A.B. from Bryn Mawr College with honors in mathematics in 1979, and her Ph.D. from Stanford University in electrical engineering in 1994. She has worked in both industry and academia and has been a visiting professor at Lulea University of Technology, the Royal Institute of Technology in Stockholm, Stanford University, and Northeastern University. She is an associate professor at Santa Clara University. She has served as an editor for *IEEE Transactions on Wireless Communications*, *IEEE Communications Letters*, and *IEEE Transactions on Communications*, and as the Editor-in-Chief of *IEEE Communications Letters*. She is a Fellow of the IEEE and the Vice-President for Publications of the IEEE Communications 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)

SAN DIEGO, CALIFORNIA TO HOST IEEE GLOBECOM 2015

LEADING GLOBAL COMMUNICATIONS CONFERENCE TO SHOWCASE NEXT GENERATION TECHNOLOGIES AND INNOVATIONS FROM DECEMBER 6–10 IN "AMERICA'S FINEST CITY"

IEEE GLOBECOM 2015 will hold its 58th annual international communications conference from December 6 – 10 at the Hilton San Diego Bayfront Hotel in San Diego, California. Known as "America's Finest City" due to its economic, cultural and scenic diversity, San Diego's thriving information technology marketplace also makes it the ideal setting for this year's premier event showcasing the entire communications spectrum ranging from mobile cloud computing and green ICT to 5G cellular and Internet of Things (IoT) networking services and applications.

"Themed "Connecting All Through Communications," IEEE GlOBECOM 2015 represents the essence of San Diego's technology growth. We are dedicated to fostering advancements and stimulating interests that drive breakthroughs not only in research, but in practical, real-world solutions that make lives better. Thousands of researchers, professionals and academics are expected to join us in this quest, while attending hundreds of sessions presented by the world's leading scientific authorities," says Ed Tiedemann, IEEE GLOBECOM 2015 General Chair and Senior Vice President of Engineering at Qualcomm Technologies, Inc.

IEEE GLOBECOM 2015 will begin Sunday, December 6 with the first of two full days of workshops and tutorials exploring topics such as "Fog Network and Internet of Things (IoT) in Wireless 5G Environments," "Designing Next Generation Energy Efficient Wireless Networks," "Towards 5G Internet of Things," "SmartGrid Resilience" and "Wireless Networking, Control & Positioning for Unmanned Autonomous Vehicles."

On the following day, Monday, December 7, the conference will initiate three-days of executive forums, industrial panel discussions and technical symposia highlighting the entire spectrum of broadband, wireless, multimedia, data, image and voice communications. The conference's keynote agenda will begin on Monday morning with Mark Dankberg, Co-Founder, CEO and Chairman of the Board of ViaSat, a San Diego technology firm selected to the Inc. 500 list of fastest growing private companies on three separate occasions. He will be followed by Eric Starkloff, Executive Vice President of Global Sales and Marketing of National Instruments, who will speak on "Transforming Traditional Design Paradigms in 5G Wireless Communications" and overcoming complex system challenges with software defined radio and new graphical approaches. Other leading speakers will include:

- Matt Grob, Executive Vice President, Qualcomm Technologies, Inc. and Chief Technology Officer, who will talk about "From 4G to 5G: The Evolution of Mobile Communication" and the arrival of LTE in unlicensed spectrum, expanded connectivity needs and new connectivity paradigms
- Ron Nersesian, President and CEO, Keysight, who will address "The Future of Test and Measurement for Commercial Communications" including the drive to further simulation, measurement, and validation dimensions with an unprecedented emphasis on software and applications relating to network performance
- Seizo Onoe, CTO, EVP, Member of Board of Directors, and Managing Director of R&D Innovation Division of NTT DOCOMO, INC., who will cover "Evolution toward 5G and beyond" as well as the current status of LTE, LTE-Advanced and the latest technology trends
- Kenneth Stewart, Intel Fellow and Chief Wireless Technologist at Intel, who will talk about the "Future of Wireless Technologies – From 5G to IoT/MTC" and the development of new radio access technologies (RAT(s) focused on flexible and efficient physical layer frameworks: low power, low overhead and highly scalable multiple-access designs supporting massive IoT access; and efficient and flexible time and frequency domain multiplexing providing the optimal tradeoff between reliability, latency and efficiency
- Sachin Katti, Assistant Professor of Electrical Engineering and Computer Science at Stanford University, who will speak on "Full Duplex Radios: From Impossibility to Practice," including issues related to self-interference cancellation and the cross-disciplinary nature of the research enabling the design and build of world LTE phones, spectrum slicing, WiFi channel aggregation, mesh networks and novel backscatter RF imaging applications

Monday through Wednesday, IEEE GLOBECOM 2015 will host a comprehensive technical symposia program offering oral and poster presentations with 1,000+ scientific papers, grouped into 12 thematic symposia, and more than 15 parallel sessions. For example, specific presentations will target next generation research in device-to-device communications, self-organizing networks, green communications and computing, millimeter wave communications, content centric network design, vehicular networks, Internet security, video streaming, data storage, game theory, routing and reliability, and big data networking, among hundreds of other topics.

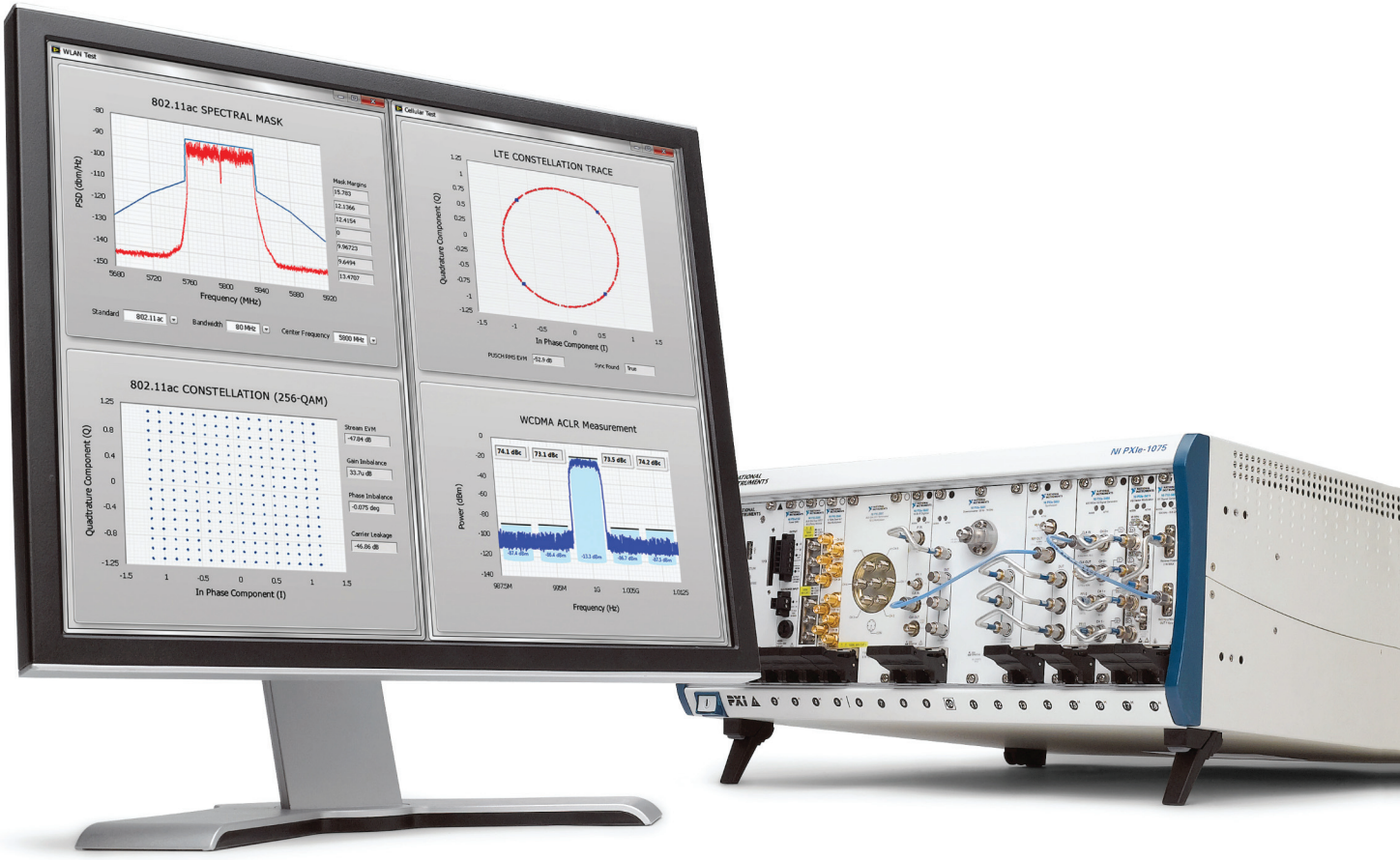
Other highlights include the conference's Industry Forum & Exposition (IF&E) composed of moderated business panels, demonstrations and poster presentations designed to promote new ideas, trends and product innovations, while facilitating peer networking opportunities. For instance, Wednesday afternoon's Lightning Talks session moderated by Leonard Reder of JPL offers a lively, informal format for 20 separate presenters to address timely and pressing technical industry topics in brief, five-minute talks. Another prominent feature is the interactive demonstrations of leading communications corporations and researchers exploring areas such as 10Gbps E-band Radio Links, Dense Cooperative Wireless Cloud Networking, Intelligent Electric Vehicle Charging Systems (IEVCS) and Radio-as-a-Service 4G LTE Networks.

IEEE GLOBECOM 2015 will then conclude on Thursday, December 10 with another full day of tutorials and workshops targeting subjects like "Tactile Internet: Application, Challenges and First Solutions," "On the 'Cloudification' of Mobile Core Networks," "Smart Cities and the Vehicular Cloud," "LTE to Unlicensed Bands," and "Green Standardization and Industry Issues for ICT and Relevant Technologies."

For more information about IEEE GLOBECOM 2015 including program updates and registration information, please visit http://globecom2015.ieee-globecom.org. All website visitors are also invited to network with colleagues and peers, share their professional experiences through the conference's Facebook, LinkedIn and Twitter pages.

Redefining RF and Microwave Instrumentation

with open software and modular hardware



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

WIRELESS TECHNOLOGIES

National Instruments supports a broad range of wireless standards including:

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

>> Learn more at ni.com/redefine

800 813 5078

NATIONAL INSTRUMENTS™

CONFERENCE CALENDAR

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

2015

NOVEMBER

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

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

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

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

AINL-ISMW FRUCT 2015 — Artificial Intelligence and Natural Language & Information Extraction, Social Media and Web Search FRUCT Conference, 9–14 Nov.
St. Petersburg, Russia
http://fruct.org/node/364339

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

IEEE 5G Toronto Summit, 14 Nov.
Toronto, Canada
http://www.5gsummit.org/toronto/

IEEE 5G Silicon Valley Summit, 16 Nov.
San Francisco, CA
http://www.5gsummit.org/sf/

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

DECEMBER

NETGAMES 2015 — Int'l. Workshop on Network and Systems Support for Games, 3–4 Dec.
Zagreb, Croatia
http://netgames2015.fer.hr/

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

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

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

ICSPCS 2015 — Int'l.ConferenceSignal Processing and Communication Systems, 14–16 Dec.
Cairns, Australia.
http://www.dspcs-witsp.com/icspcs_2015/index.html

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

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

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

2016

JANUARY

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/

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/

–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.

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/

CONFERENCE CALENDAR

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/

CALL FOR PAPERS
IEEE WIRELESS COMMUNICATIONS MAGAZINE
LTE IN UNLICENSED SPECTRUM

BACKGROUND

Recently, the Federal Communications Commission (FCC) has released an additional 295 MHz bandwidth in the 5G Unlicensed National Information Infrastructure (U-NII) bands. These unlicensed bands attract great attention for cellular operators to expand their network capacity. Many leading companies have addressed the standardizing of LTE in unlicensed spectrum (LTE-U) and the 3rd Generation Partnership Project (3GPP) has formally started working on Rel-13 Licensed-Assisted Access (LAA) technique. There are many challenges for LTE working on unlicensed spectrum. The most important issue is how to design harmonious coexistence mechanisms between LTE and WiFi users. Moreover, since unlicensed spectrum is open accessed by different cellular operators, fair and efficient spectrum sharing technique is required. Finally, how to provide quality-of-service for LTE traffic on the unreliable unlicensed bands is also a challenging issue. This feature topic will focus on various design issues for LTE in unlicensed spectrum, and aim at bringing together the state-of-art research results and industrial applications. Comprehensive surveys and original contributions, which are previously unpublished and not currently under review by another journal, are solicited.

SUBMISSIONS

For manuscript preparation and submission, please follow the guidelines in the Author Guidelines and Paper Submission Guidelines section at the IEEE Wireless Communications web page,http://www.comsoc.org/wirelessmag. A paper should have no more than 4500 words, no more than 6 tables/figures, and its abstract should have no more than 250 words. Any submission that fails to comply with the guidelines will be rejected without review. Papers must be submitted in PDF format to the Manuscript Centralhttp://mc.manuscriptcentral.com/ieee-wcm. The timetable is as follows:

SCHEDULE FOR SUBMISSIONS

- Manuscript Submission Deadline: February 1, 2016
- First Round of Review Due: May 1, 2016
- Revised Manuscript Due: June 15, 2016
- Final Acceptance Notification: July 15, 2016
- Final Manuscript Due: August 15, 2016
- Publication Date: October 2016

GUEST EDITORS

Prof. Guanding Yu
Zhejiang University, China
yuguanding@zju.edu.cn

Prof. Geoffrey Ye Li
Georgia Institute of Technology, USA

Prof. Li-Chun Wang
National Chiao Tung University, Taiwan

Dr. Amine Maaref
Huawei Technologies Canada, Ottawa, CA

Dr. Jemin Lee
Singapore University of Technology and Design, Singapore

Dr. David Lopez-Perez
Bell Labs, Alcatel-Lucent, Dublin, Ireland

OCTOBER 2015

free Tutorials Now™

Enabling SDN based High Performance Enterprise WiFi Systems with Zynq All Programmable SoCs

Designing secure, high performance Enterprise WiFi networks while serving a rapidly exploding number of wireless devices is a real challenge. NPUs with their inflexible feature sets increasingly struggle to handle the expanding universe of wireless devices and ever-increasing data rates; they stumble over new security protocols; and they just weren't designed for the brave new world of SDN (Software Defined Networking) and the closely associated OpenFlow protocol.

Xilinx Zynq® All Programmable SoCs, which meld ARM® Cortex-A9 processors with programmable logic, have sufficient flexibility to meet these design challenges head on. Even a smaller family member, the Zynq 7015 has sufficient processing horsepower and programmable hardware to easily tackle the complexities of routing and security in Enterprise WiFi networks. This webinar supplies you with the information and tools you'll need to develop SDN-based, Enterprise-class WiFi routers and access points based on Xilinx Zynq SoCs.

Sponsor content provided by: XILINX ALL PROGRAMMABLE

Limited Time Only at >> ww.comsoc.org/freetutorials

IEEE COMMUNICATIONS SOCIETY

For this and other sponsor opportunities,
please contact Mark David // 732-465-6473 // m.david@ieee.org

GCN GLOBAL COMMUNICATIONS NEWSLETTER

October 2015
ISSN 2374-1082

CHAPTER NEWS

IEEE ComSoc Central Texas Austin Chapter

Winner of the 2015 Chapter Achievement Award, North America Region; Winner of the 2015 Chapter of the Year Award

By Fawzi Behman, Central Texas Austin Chapter Chair, USA

As a senior member of IEEE, it has been a great privilege to work with a talented team of officers to serve the chapter professionals and community here in Austin, which has made it a rewarding year for all.

As a brief introduction: Austin is the capital of the state of Texas and home to the University of Texas (UT) and many high tech companies. Austin is known as the Live Music Capital of the World, and is one of America's fastest growing and most desirable cities. Over the years, the chapter has reached out and strengthened relationships with UT, as well as several of the local high tech companies including AT&T, IBM, NI, Freescale, HP, and others. As a result, we are grateful to be able to host our regular meetings at the nice facilities of AT&T. In addition, when the chapter brings a Distinguished Lecturer, it becomes a customary that not only does the DL addresses the IEEE but also visit and present at local companies such as IBM and AT&T.

2014 Planning: It all started with careful planning, articulating on the mission of delivering quality events to the membership/community and providing opportunities for education and networking. The theme for 2014 was Wireless & Networking Architectural Evolution and Services. Among the key objectives were: conducting more technical meetings, engaging with DLT/DSP, increasing membership by reaching out and participating in local activities, collaborating with other societies and young professionals, supporting GLOBECOM 2014, ensuring the chapter's financial health and ensuring that the chapter's website is updated (http://sites.ieee.org/ct-comsp/) promoting meetings and special events, and growing social media content such as Twitter, Facebook and LinkedIn.

Left to right: DL Dr. Lin, Dr. Bob Dailey, Dr. Jason, Fawzi Behmann.

December 12, 2014: DL Prof. Lin presenting to ComSoc/SP Austin Chapter and meeting with AT&T technical staff.

Technical Meetings/Topics:

•In 2014 the Austin chapter conducted 34 meetings, of which 21 were technical meetings representing the highest in CTS for 2014. All meetings were reported in vtools and L31.

•Among the key topics covered in 2014 included "Harnessing Electrical Demand Flexibility for a Sustainable Energy Future," "Heterogeneous Computing: Promises and Challenges," "ZPEG Video Compression," "Video Surveillance," "Geospatial Technologies," "Traffic Forensics," "Big Data, Cyber-Biological Systems," "UHD Technology Evaluation at AT&T," "IOE – Internet of Everything," "Engineering and Government," "Aging, Brain Science and the Future of Healthcare," among others.

Distinguished Lecturer/Speaker Program: In 2014 the chapter hosted Prof. Ying-Dar Lin, IEEE Fellow, National Chiao Tung University, Taiwan. During his visit to Austin, Dr. Lin delivered presentations to IBM, AT&T ,and COMSOC/SP and other joint chapters. Dr. Lin spoke on "Traffic Forensics: Capture, Replay, Classification, Detection, and Analysis," as well as "Software Defined Networking: Why, When, Where, and How."

Speaker Appreciation Award: The chapter values the time and sacrifice put in by the speaker in delivering quality content. As a token of appreciation, the chapter prepared and delivered a personalized plaque to each speaker.

Membership Development: The chapter took pride in participating in select events that helped recruit new members and generate awareness among the community. Among the events were National Instrument NI Week, Innotech, Texas Wireless Summit, IEEE Gold Event at SxSW, and IEEE Cybersecurity day.

As a chair of the chapter, I collaborated with the Chinese American Semiconductor Professional Association (CASPA) chapter chair in Austin and held a few joint sessions. CASPA has more than 100 engineers, most of whom are not IEEE members. Joint sessions provided a forum for a good exchange and to provide information about IEEE membership. Overall, according to SAMIEEE, more than 50 new members were added to the IEEE COMSOC/SP Chapter in Austin in 2014.

Collaboration with other societies/MGA/IEEE:

•The Austin COMSOC chapter collaborated with several other chapters/councils. These included DSP, CS, I&M, TMC, CTCN, and CEDA. Among the examples, COMSOC joined other chapters for a full-day educational program on career development, including a workshop with more than 65 attendees. The theme was "Skills

(Continued on Newsletter page 5)

IEEE ComSoc Ecuador Chapter

Winner of the 2015 Chapter Achievement Award, Latin America Region

By Hernán A. Samaniego Armijos, Ecuador Chapter Chair

The Ecuador Chapter of the IEEE Communications Society has 147 active members (as of Sept. 2015) between professionals and students in the area of telecommunications. The objective of the ComSoc Ecuador Chapter is to strengthen the professional growth of its members, creating spaces for sharing knowledge and technological advances in the area of communications. ComSoc Ecuador has four technical student chapters at the Technical University of Loja – UTPL, the Polytechnic School of the Coast – ESPOL, the National Polytechnic School – EPN, and the Technical University of the North – UTN.

The executive committee is composed of:

Chair: Hernán Samaniego Armijos, Master in telematics and telecommunication networks, Manager of the National Telecommunications Corporation CNT-EP in Loja, and Visiting Professor at the Technical University of Loja – UTPL.

Vice-Chair: Vanessa Cuesta, Engineer in Electronics and Telecommunications, works for the Bank of Loja as a telecommunications technician.

Treasurer–Secretary: Eduardo Suarez, Engineer in Electronics and Telecommunications, works as an independent professional.

Organized Events: ComSoc Ecuador participated as the moderator in a workshop about Basic Infrastructure in the Plan of Territorialization and Innovation, integrating university, government, and industry, in the city of Loja in the south of Ecuador, proposing several projects taking advantage of Loja is an IEEE Smart City Affiliate. In addition, ComSoc Ecuador participated with a lecture on Smart Cities by Hernán Samaniego Armijos.

ComSoc Ecuador participated in the International Smart Grid and Smart Cities International Seminar in the city of Cuenca, Ecuador, where experts from prestigious appointments, professionals, and renowned companies in the field of Smart Grid and Smart Cities, are gathered to show expertise, practical experience and the progress that has been achieved so far in these fields. Hernán Samaniego Armijos gave a conference about wireless sensor networks for smart cities.

In Ibarra City in the north of Ecuador ComSoc participated in the High Tech–Smart Cities: New Opportunities for Telecommunications Professionals in Ecuador, an event that had the participation of CNT EP, the Ecuador telecom company, and ARCOTEL, the telecommunications regulatory entity. Hernán Samaniego participated with a lecture on Sensor Networks for Smart Cities.

In addition, we took this opportunity to encourage the creation of the ComSoc technical chapter in the Northern Technical University – UTN Student Branch.

Summer Course on IP Telephony Systems: In order to prepare the professionals in communication multimedia services, ComSoc Ecuador made two webinars. "Introduction to the IP Telephony Systems" reviewed aspects related to the systems of traditional telephony, voice over IP and operating platforms for IP telephony systems, dictated by Daniel Guevara, who has several years of experience in this type of communications systems. "Protocols of VoIP" reviewed the protocols used in VoIP systems such as SIP/SDP/RTP/RTCP/IAX, and was dictated by Hernán Samaniego Armijos, Master in telematics and telecommunication networks.

Hernán Samaniego at the conference on Wireless Sensor Networks for Smart Cities.

Hernán Samaniego moderating workshop.

Participants in the International Smart Grid and Smart Cities International Seminar.

ComSoc Ecuador recognition by the UTN for the conference.

Chaskieee Project: Chasqui is the term used at the time of the Incas to identify the person responsible for the comunication between different tribes, and likewise a project called Chaskieee has been initiated. Professionals in ComSoc travel through different cities of Ecuador, giving lectures at the universities where ComSoc technical chapters have assets. This project is managed by Eduardo Suarez, Treasurer–Secretary of ComSoc Ecuador.

Internet Day: With the purpose of commemorating the International Day of Telecommunications and Information Society, ComSoc Ecuador organized an event named Internet Day, with joint activities in Ibarra and Loja in collaboration with the UTPL and UTN ComSoc student technical chapters. Vanessa Cuesta, ComSoc Ecuador Vice-Chair, participated with a lecture about corporate social networks.

DLT Program: We are expecting the visit of ComSoc Distinguished Lecturer Sonia Aissa in October 2015 in the cities of Guayaquil and Ibarra, discussing the topics: "Cognitive Radio" and "Energy Harvesting Communication Networks."

IEEE ComSoc Sweden Vehicular Technology/Communications/Information Theory Joint Chapter

Winner of the 2015 Chapter Achievement Award, Europe, Middle East, Africa Region

By Tommy Svensson, Sweden Chapter Chair, Chalmers University of Technology, Sweden

Chapters are important activity centers within the IEEE. It is in the chapters where the technical and scientific daily life takes place for our members. To this end, on the IEEE ComSoc Sweden VT/COM/IT joint chapter board, we have focused on creating value for our members via networking opportunities in Sweden in conjunction with high quality technical/scientific seminars and workshops, and by catalyzing important information flows from IEEE to our members. There are only about 9.6 million people in Sweden, so we are proud of our 599 unique VT/COM/IT members (797 memberships: VTS-117, Comsoc-583, ITS-97).

In 2014 we organized 13 IEEE Technical seminars by distinguished speakers: Dr. Nicola Marchetti (Trinity College Dublin); Prof. Linda Doyle (Trinity College Dublin); Prof. John Thompson (The University of Edinburgh); Prof. Ali H. Sayed (University of California); Prof. Maïté Brandt-Pearce (University of Virginia); Dr. Matthieu Bloch (Georgia Institute of Technology); Prof . João Hespanha (University of California–Santa Barbara); Assoc. Prof. Serdar Yuksel (Queens University); Prof. Virgil Gligor (Carnegie Mellon University); Prof. Gene Tsudik (University of California–Irvine); Prof. Ness Shroff (The Ohio State University); Prof. Merouane Debbah (Supelec); and Dr. Laurent Schmalen (Alcatel-Lucent Stuttgart). The seminars were hosted by the KTH Royal Institute of Technology, Chalmers University of Technology, and Linköping University.

We have initiated two annual Swedish workshops. The Swedish Communication Technologies Workshop (Swe-CTW) was first organized in 2011; the Workshop on Wireless Vehicular Communications was first organized in 2010. The Swe-CTW brings together researchers and research students in the general area of communication technologies and related areas. It is a three day workshop that provides an opportunity for researchers and research students in Sweden to gather in a largely informal setting to share ideas, make contacts, and foster new collaborative links for their future careers. The 2014 edition of Swe-CTW was organized June 2-5 2014 in Västerås in collaboration with Mälardalen University. In addition to the poster presentations by young researchers, there were two tutorials: "Sparse Systems" presented by Dr. Saikat Chatterjee (KTH Royal Institute of Technology), and "An Introduction to Wireless Sensor Networks," presented by Dr. Carlo Fischione (KTH Royal Institute of Technology). Swe-CTW'2014 attracted 62 attendees. The 2014 edition of the Workshop on Wireless Vehicular Communications was organized, as traditionally, in Halmstad, this year on Nov 11 2014. There were 40 participants listening to seven presentations, including the invited speech on "Dependable Vehicular Communications in Non-Stationary Propagation Conditions," given by Thomas Zemen, FTW Telecommunications Research Center, Vienna.

Tommy Svensson, Sweden Chapter Chair.

Swe-CTW'2014, Västerås, June 2-5 2014.

Workshop on Wireless Vehicular Communications 2014, Halmstad, Nov 11, 2014.

We are also acting as technical sponsor for IEEE conferences and workshops organized by others in Sweden, where we help to advertise the event in our online media (our IEEE web site, LinkedIn and Facebook, see links below). In 2014 we sponsored The 18th International Conference on Optical Networking Design and Modeling (ONDM'2014), and the IEEE International Workshop on Signal Processing Advances in Wireless Communications, (SPAWC) 2015.

To promote the visibility of young talented researchers, we have initiated two annual best paper awards for young researchers in Sweden: the IEEE Sweden VT-COM-IT Joint Chapter Best Student Journal Paper Award, and the IEEE Sweden VT-COM-IT Joint Chapter Best Student Conference Paper Award. These awards are given to the winners during the annual Swe-CTW. We are also catalyzing nominations to awards programs within our societies, such as the ComSoc EMEA Regional Young Researcher of the Year Award.

It seems that we do a couple of things right. On the board we are proud to have received six awards: the IEEE Communications Society Chapter Achievement Award 2011, 2012, 2013, and 2015, and the Chapter of the Year Award 2013 from the IEEE Information Theory Society. In addition, we received the IEEE Communications Society Special recognition award "For long-time achievements in chapter activities and ongoing excellence serving its members" in December 2013. We think that the composition of our board is an important enabler, comprising very committed persons who span competencies across our technical/scientific areas within the scope of the Vehicular Technology, Communications and Information Theory Societies; academia and industry; senior and younger professionals; as well as a geographical span in Sweden. We have been quite efficient in creating activities for our members. We think the key there is to distribute responsibilities among the board members, taking opportunities when they appear, and working with minimal administrative overhead. For instance, we promote individual initiatives, and we meet using regular telephone conferences (around once per month) where we try to sort out most things, to avoid lengthy email discussions. We also store all our documents on a common server, so that all board members can always be up-to-date even if they cannot attend a telephone meeting.

(Continued on Newsletter page 5)

IEEE ComSoc Kolkata Chapter

Winner of the 2015 Chapter Achievement Award, Asia Pacific Region

By Iti Saha Misra, Kolkata Chapter Chair, India

It is our great pleasure to share the ComSoc Kolkata Chapter activities for the year 2014, which led to our Chapter Achievement Award this year. The Chapter activities start with executive meetings in the first week of February, when members decide the next year's activities and make their commitment to volunteer for various programs. We the members of Chapter executive committee thought of conducting a DLT/DLP event, technical seminars, student activity programs, technical programs involving industry, along with membership development and awareness generation programs. Our chapter hosted an "Algomaniac Student Contest Event," one of the major events under Convolution-2014, organized by the students of Electrical Engineering, Jadavpur University, Kolkata, on 14–15 Feb. 2014. Interdepartmental Engineering students participated the contest to test their programming capability along with algorithmic design.

The next big event was the Distinguished Lecturer Tour: DLT 2014. The Chapter organized two DLT tours in the months of March and December involving engineering students from Jadavpur University and the neighboring Institutes, Ph.D. scholars, and young faculty members.

Mobile health care, sensor networks, and automation have become very popular areas of research in the recent years. The Kolkata ComSoC Chapter conducted two very successful DLTs on these areas. One of the DLT programs was on the topic "Towards an Intelligent and Ubiquitous Healthcare Infrastructure: Challenges and Trends." The DLT speaker was Prof. Pradeep Kumar Ray, Director of the WHO Collaborating Centre on eHealth, Asia Pacific Ubiquitous Healthcare Research Centre (APuHC), University of New South Wales, Australia. It was a very nice experience to gather 104 participants from Jadavpur University itself and the other Institutions, as well as industry, for this DLT, which included extensive interaction with the speaker. The Chapter Chairperson, Iti Saha Misra, highlighted the various activities of the Kolkata Chapter and invited young professionals and students to join IEEE as part of their career growth path. The second DLT program was on the topic "Robotics, Automation and Machine Learning," delivered by Prof. (Dr.) G. Sengupta, Masse University, Newzeland, on 22 Dec. 2014. It was a mesmerizing lecture event with 70 enthusiastic participants.

DLT interactive session: Pradeep K. Ray with Iti Saha Misra, ComSoc Kolkata Chapter Chair.

DL speaker Prof. Gourav Sengupta with the audience.

Dr. Arijit Mukherjee of TCS, delivering his lecture on "Grid, Cloud, Fog and Beyond."

3-day workshop on Android Program, students of ETCE have joined the program.

The year 2014 was the year of industry and R&D interaction in the Kolkata Chapter. Several technical lectures and interactive session were organized for the academic benefit of the student community and researchers. The Chapter invited ISRO Senior Scientist, Tosicul Wara (Scientist-SF) to deliver the video of the ISRO Satellite for Navigation Application in Indian Context on 13 May 2014. A series of lectures was organized with industry Interaction in a program involving TCS Innovation Lab of Kolkata Centre, where the current industry research activities were discussed on the topics of Internet of Things (IOT), Security Aspects and Signal Processing aspects on 29 Auguat 2014. The one-day seminar program was very special to the young professionals and students of ECE, inspiring them in their research activities. This also helped the young students to develop their interest in recent research beyond their curricula. A technical training program from Industry personnel for PG Communication Specialization students held on April 3, 2014, ETCE Department, Jadavpur University. These types of technical programs help students link their subject of study with the application domains.

There was a wonderful program for the three-day workshop on the topic "Introduction to Android Operating Systems for Touch Screen Smart Phones" for the undergraduate students of Electronics and Telecommunication Engineering in the midst of summer vacation, under the auspices of TEQIP Phase–II of Jadavpur University on 11-13 June 2014. The Dean of the Engineering Faculty and also theTEQIP Coordinator inaugurated the workshop. CASS Chapter Chair, Prof. S.K. Sanyal, delivered a won-

(Continued on Newsletter page 5)

AUSTIN CHAPTER/*Continued from page 1*

for a Multi-Faceted Career." Among the speakers were Gary Black, IEEE-USA 2014 President, Bob Krause, IEEE-USA AICN, and others.

•Among other activities, the chapter helped Central Texas chapter chairs take advantage of the IEEE PACE Program. As a section PACE chair, I organized several PACE projects, which increased from three projects in 2013 to seven projects in 2014. I also led a PACE automation program that is expected benefit Regions 1-6.

•Among the new IEEE initiatives is Collabratec, for which I participated as a moderator representing the Austin Community in the pilot study, and engaged in creating messages and discussions to bring about building a community of interest for Austin.

Activities with Young Professional/Students:

•COMSOC co-sponsored a young professional mixer in Austin during SxSW in 2014. The event attacted more than 100 attendees. The chapter organized an information table. During that event, presentations were made by Glenn Zorpette, Executive Editor of *IEEE Spectrum*, and Dr. Robin Murphy, an expert on search and rescue drones.

•The chapter supported GLOBECOM 2014 (Dec. 8-12) and secured more than 100 students from the University of Texas, San Antonio, State University of San Marcos, and others who volunteered as room monitors and at the registration desk. As the chair of marketing and publicity for GLOBECOM, I sent an appreciation note to the students for their dedication and exemplary performance at the event.

Visibility at the Community Level: As part of our efforts to promote GLOBECOM 2014, we reached out to:

•Austin Chamber of Commerce. We engaged the Austin Chamber of Commerce with promotional activities for GLOBECOM 2014 for several months prior to the event targeting the technical community and startups. The Chamber of Commerce provided $2500 for the Conference.

•City of Austin. We engaged with the Mayor's office in Austin and succeeded in obtaining a Proclamation declaring December 8-12 to be IEEE week in Austin. Later, the Mayor's office issued a quote that was used in a press release issued by COMSOC promoting GLOBECOM 2014

Collaboration IEEE/Hilton/AT&T for GLOBECOM 2014 IT Infrastructure:

•The chapter successfully led an infrastructure IT project in delivering an enhanced WiFi/Cellular signal and coverage that supported more than 3300 client devices and 2500 persons attendees. With the upgrade of the Wifi network by Hilton and the cellular network by AT&T, the peak traffic reached only 30 percent of total bandwidth capacity.

•This was the first time ever in COMSOC that a feasibility assessment followed by implementation of upgrading equipment was undertaken for an international event.

Financial Health: The chapter continued enjoying good financial health due to careful budgeting and spending on meetings and events. Collaborating and joining with other chapters helped share expenses, and at the same time enabled us to add more meetings for the year. We do not charge for meetings. Regarding food and beverages (F&B), officers helped in ordering catered food and carrying them to the meeting facilities. Promotion of events, printing of feedback forms and transportation were done at no cost to the society. The appreciation plaques were typically purchased at a 50 percent discount at the beginning of the year, and personalization and execution were typically done by the chapter chair.

SWEDEN CHAPTER/*Continued from page 3*

We are proud of our achievements so far, but there are important grand challenges for the future, such as improving the quality of life for all people, sustainability, and maintaining the interest for engineering among young people. IEEE is a fantastic organization of talented and knowledgeable people! Together we can tackle these challenges. In the IEEE Sweden section we are currently discussing suitable initiatives that could address these important future challenges.

http://sites.ieee.org/sweden-vtcomit/
http://www.linkedin.com/groups/
IEEE-Sweden-VT-COM-IT-5070706
https://www.facebook.com/IeeeSwedenVtcomit
http://www.ieee.se/chapterindex.php?code=vt_com_it
http://www.es.mdh.se/swe-ctw2014/Welcome.html
http://www.hh.se/english/schoolofinformationtechnology/
eventsandseminars/
workshoponwirelessvehicularcommunications2014.65442012.html

KOLKATA CHAPTER/*Continued from page 4*

derful lecture to the students on how the technology is changing and encouraged the students to adapt in this ever changing environment of technology. The workshop was a great success with the fantastic delivery of the speaker, Mr. Dipanjan Biswas from a local company, SkyBits Kolkata.

The Chapter is also aware of its membership development program. During each program, the Chairperson, Prof. Iti Saha Misra, addresses the audience, inviting them to be a part of the IEEE programs through their membership, and she discusses the benefits of IEEE. She encourages young researchers to volunteer for the organization of the programs as a way to improve their organizing capabilities, as well as for the added opportunitiues to network with others in the field. They will be the future torch bearers of IEEE societies. In return, volunteers are given a certificate from the Chapter Chair.

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

A publication of the IEEE Communications Society

www.comsoc.org/gcn
ISSN 2374-1082

MILITARY COMMUNICATIONS

Torleiv Maseng

Randall Landry

Modern military communications systems continue to face a number of unique challenges not present in commercial enterprises. These challenges include among other things, operating in contested environments, low bandwidth communications links, unpredictable topological changes, and the need for special priority and security level considerations. In the face of these challenges, there is a need to leverage and build upon the explosive information technology advances being made in the commercial markets. This Feature Topic explores an interesting and important cross-section of topics along these lines, spanning communications layers from physical, through networking and higher layer IT services.

The first article by Mailloux *et al.*, tackles the important problem of secure communications in military systems through the use of Quantum Key Distribution (QKD). The authors describe QKD systems as part of a secure communications solution, and examine the performance of decoy state QKD systems in support of metropolitan and long-haul quantum communication networks. They propose a practical variation of the decoy state protocol which does not require a priori knowledge of the quantum channel loss. The authors argue that this is an advancement towards practical QKD systems to detect an entire class of attacks which introduce photon number specific interference.

Disruption Tolerant Networking (DTN) is a technology that has been successfully applied to domains such as satellite communications, where link errors can have a particularly negative impact on system throughput due to typically large bandwidth-delay products. DTN can also be useful in environments where topological changes are not easily handled by traditional IP routing protocols, thereby resulting in poor throughput performance. The article by Amin *et al.*, examines the use of DTN Bundle Protocol and applies it to an aerial high capacity backbone network that carries traffic from ships/subs to shore command post. The authors explore applying DTN proxies, tunnels, and interfaces to both the plaintext and ciphertext side of military networks to understand architecture and design considerations and limitations.

The value of information-based approaches to networking has become a topic of great interest with the explosion of networked devices and the anticipated adoption of the Internet of Things (IoT). Information-centric solutions also have the potential to address many of the challenges posed by tactical military environments in which low bandwidth links and changing topologies are a fact of life. The article by Suri *et al.*, explores the notion of determining the value of information in order to prioritize and filter information that is disseminated over these tactical networks, focusing on the dissemination of information to and from dismounted soldiers in a battlefield environment. This is a promising approach to mitigate the constraints of tactical networks and to reduce information overload on soldiers.

Messaging services represent some of the most important, and heavily utilized, IT services used in military systems. Delivery of messages depends of the recipient's role in the organization, trustworthiness and the type of terminal equipment used. Messages also have different priorities and classifications. In the article by Cailleux *et al.*, an innovative correspondence model for a future military messaging services is described, which overcomes some of the shortcomings of current email systems for military use. The main interest of this model is to provide new prospects for military messaging systems with the capability of providing different levels of services by defining and enforcing policies on a per-usage basis. The authors argue that such a dynamic correspondence model provides a much richer set of email services that can be tailored to the unique conditions and requirements of military environments.

The article by Brannsten *et al.*, explores the difficult problem of extending a rich set of IT services, largely built for enterprise networks, to the tactical domain. The authors focus on core services defined by the NATO Network Enabled Capabilities (NNEC), which include among others, messaging, collaboration, discovery and security. NNEC focuses on interoperability and promotes the use of Service Oriented Architecture (SOA) and Web services as enabling technologies. The authors evaluate the core services defined in NNEC for deployment at the tactical level, discuss the challenges related to extending support for these services into the tactical domain, and identify potential solutions.

We trust that you will find these articles to be interesting and relevant to the challenges facing military communications today. Please look for our call for papers soliciting article submissions for the 2016 Feature Topic on Military Communications.

BIOGRAPHIES

TORLEIV MASENG (torleiv.maseng@ffi.no) is director of research at the Norwegian Defense Research Establishment, where he is responsible for communications and information systems. He worked as a scientist at SINTEF in Trondheim for 10 years, involved in design and standardization of GSM. For seven years he was a scientist at the NC3A NATO research center in The

Hague. During 1992–1994 he was involved in the startup of the new private mobile operator NetCom GSM in Norway, where he had technical responsibility. Since 1994 he has held a chair in radio communications at the University of Lund, Sweden. In 1996 he took up his employment at the Norwegian Defense Research Establishment located at Kjeller, 20 km outside Oslo. Since 2005 he is also Professor II at the University of Oslo. He is the author of more than 150 papers, holds patents, and is a Technical Editor of *IEEE Communications Magazine*. He has received an award for outstanding research and has arranged large international conferences.

RANDALL LANDRY (rlandry@mitre.org) is a technical director at the MITRE Corporation, where he runs a division specializing in advanced computing and networking technologies. He has served in a number of roles in support of cyber defense and large-scale network design and deployment for the U.S. Air Force and Department of Defense. Prior to joining MITRE, he worked in Corporate R&D at Texas Instruments, and holds several patents in network algorithms and highly integrated switching architectures for gigabit networking. He has also served as a director of optical and wireless networking in the telecommunications industry. He has led research programs in satellite communications, tactical wireless networking, cross-layer design methodologies, and network science. His current research interests include named data networking for highly mobile and disadvantaged tactical networks, management and security for the Internet of Things, and software-defined networking. He has published numerous technical articles, and served as a Technical Program Committee member and session organizer for the major IEEE conferences in communications . He received his M.S. and Ph.D. in electrical engineering from the University of Vermont in 1992 and 1994.

Quantum Key Distribution: Examination of the Decoy State Protocol

Logan O. Mailloux, Michael R. Grimaila, John M. Colombi, Douglas D. Hodson, Ryan D. Engle, Colin V. McLaughlin, and Gerald Baumgartner

ABSTRACT

Quantum key distribution (QKD) is an innovative technology that exploits the laws of quantum mechanics to generate and distribute a shared cryptographic key for secure communications. The unique nature of QKD ensures that eavesdropping on quantum communications necessarily introduces detectable errors which is desirable for high-security environments. QKD systems have been demonstrated in both freespace and optical fiber configurations, gaining global interest from national laboratories, commercial entities, and the U.S. Department of Defense. However, QKD is a nascent technology where realized systems are constructed from non-ideal components, which can significantly impact system performance and security. In this article, we describe QKD technology as part of a secure communications solution and identify vulnerabilities associated with practical network architectures. In particular, we examine the performance of decoy state enabled QKD systems against a modeled photon number splitting attack and suggest an improvement to the decoy state protocol security condition that does not assume a priori knowledge of the QKD channel efficiency.

INTRODUCTION

Quantum key distribution (QKD) is the most mature application of quantum cryptography and is heralded as a revolutionary technology offering the means for physically separated parties to generate unconditionally secure shared cryptographic keying material. Unlike conventional key distribution techniques, the security of QKD systems rests in the laws of quantum mechanics and not computational complexity. This is due to the fundamental nature of quantum communications, where observation of a quantum state collapses the state, which enables detection of eavesdropping on the channel. In theory, these attributes make QKD well suited for high-security applications such as military, banking, and government environments. For example, the U.S. Navy is exploring the feasibility of ship-to-ship freespace QKD, the U.S. Army is conducting basic research toward distributed quantum systems, and the U.S. Air Force is assessing the utility of quantum technologies in multiple areas, to include satellite-based QKD. Moreover, there are commercial QKD systems available for purchase from several vendors, including ID Quantique, SeQureNet, Quintessence Labs, MagiQ Technologies, and QuantumCTek [1].

While QKD advocates are building increasingly practical systems with demonstrated networked key sharing configurations, QKD is still a nascent technology where design and implementation trade-offs are not well understood. Moreover, critical questions remain unanswered regarding the validity of its "unconditional security" claim. These concerns are well justified as real-world QKD systems are constructed from non-ideal components, which can adversely impact system performance and security. Thus, there is a clear need to further understand and study these concerns in realized QKD systems.

In this article, we introduce QKD in an assessable way for communication engineers to effectively engage the established QKD community. First, we explain fundamental QKD principles and describe the technology's integration into a secure communication solution. Additionally, practical engineering limitations and common implementation vulnerabilities in "unconditionally secure" QKD communication architectures are highlighted.[1] Second, we examine decoy state enabled QKD systems, and particularly, their ability to detect photon number splitting (PNS) attacks over metropolitan networks.

QKD FUNDAMENTALS

The genesis of QKD can be traced back to Wiesner, who proposed the idea of encoding information on polarized photons using two conjugate bases in the late 1960s [2]. In 1984, Bennett and Brassard built on this work to introduce the first QKD protocol, known as BB84, to generate shared secret keying material between two parties [3]. While BB84 remains a popular implementation choice, several alternative QKD protocols have also been developed such as con-

Logan O. Mailloux, Michael R. Grimaila, John M. Colombi, Douglas D. Hodson, and Ryan D. Engle are with the U.S. Air Force Institute of Technology.

Colin V. McLaughlin is with the Naval Research Laboratory.

Gerald Baumgartner is with the Laboratory for Telecommunication Sciences.

[1] QKD systems are subject to many of the same weaknesses as conventional networked devices (i.e., vulnerabilities in protocols, applications, and operating systems, as well as side channel attacks and others); however, it is our intent to study problems specific to QKD implementations; thus, we have chosen to study the QKD system's ability to detect attacks specific to the technology.

0163-6804/15/$25.00 © 2015 IEEE

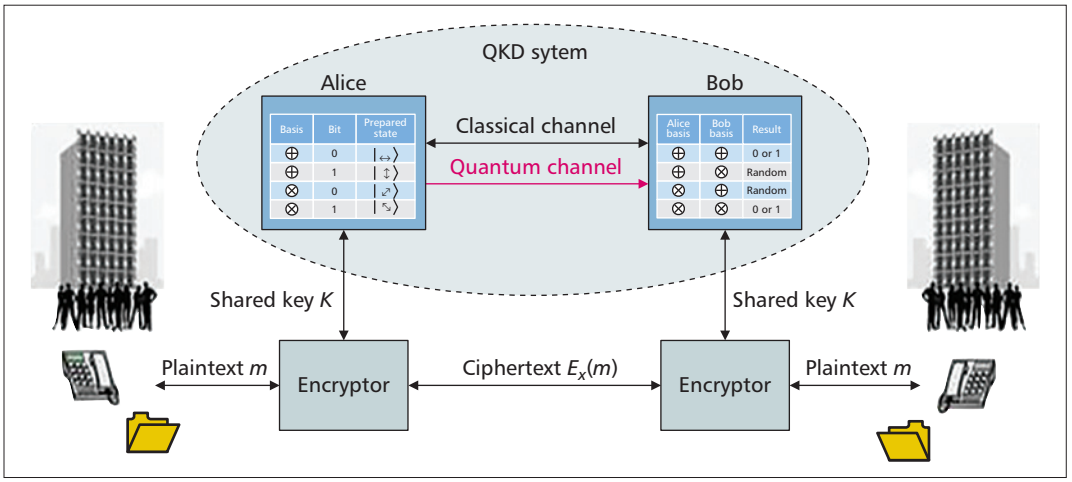


Figure 1. A BB84 polarization-based prepare and measure QKD system. The sender Alice and receiver Bob are configured to generate a shared secret key for use in bulk encryptors, where the quantum channel (i.e., a freespace or optical fiber link) is used to securely transmit qubits, and the classical channel is used to facilitate the BB84 protocol.

The security of QKD is based on quantum uncertainty, where directly measuring a qubit changes its quantum state. In classical communications, bits exist in a deterministic state of either 0 or 1 where they can be measured and re-measured as necessary. In contrast, qubits are encoded in a simultaneous combination or "superposition" of states.

tinuous variable, coherent-one-way, and entanglement-based QKD [1]. An engineering oriented review of quantum communications is available in [4], while an accessible introduction to QKD is available in [5].

QKD Protocols

Figure 1 depicts a BB84 polarization-based prepare and measure QKD system, where the sender, Alice, prepares quantum bits (qubits) in one of four polarization states: $|\leftrightarrow\rangle$, $|\updownarrow\rangle$, $|\nearrow\rangle$, or $|\nwarrow\rangle$. These qubits are encoded according to a randomly selected basis ("rectilinear" $\oplus$ or "diagonal" $\otimes$) and bit value (0 or 1), and sent over an optical fiber "quantum channel" to Bob, where they are measured using a second randomly selected basis. If Bob measures the qubits using a matching basis, the encoded message is obtained with a high degree of accuracy. Conversely, if he measures the qubits using an incorrect basis, a random result occurs, and the encoded information is destroyed. This unique phenomenon is inherent to quantum communications, where measuring an encoded qubit disturbs its quantum state. The process of preparing, sending, and measuring qubits is known as "quantum exchange," resulting in a "raw key" at Alice and Bob.

After quantum exchange, the QKD system employs classical information theory techniques to generate an error-free secure shared key [6]. First, Alice's and Bob's raw keys are sifted to eliminate incorrect (non-matching) basis measurements, which results in a shared sifted key (in both Alice and Bob) approximately half the length of the raw key due to Bob's random (50/50) basis selection. This is accomplished over the classical channel where Bob publically announces the basis he used to measure each pulse, and Alice acknowledges correct basis measurements. Note that the basis information is exchanged but not individual bit values, and thus the sifting process is secure against eavesdroppers as each bit has equal likelihood of being a 0 or 1 regardless of the basis measurement in a mutually unbiased measurement system.

Next, error reconciliation is performed to correct errors in the sifted keys where specialized error correction algorithms (e.g., Winnow, Cascade, or low-density parity check) are used to minimize the amount of information "leaked" over the classical channel regarding the secret keys. Lastly, the error reconciled key is subject to privacy amplification — an information theory technique used to ensure an eavesdropper has negligible information regarding the final shared secret key. The result is a smaller, more secure final shared secret key, K, which can be used to encrypt voice, video, or data communications.

The shared secret key K can then be used to enhance the security of conventional symmetric encryption algorithms such as 3DES or AES through frequent rekeying. Alternatively, QKD-generated key is often discussed in conjunction with the One-Time Pad (OTP) encryption algorithm to create an unbreakable cryptosystem. Despite its great appeal, OTP is not commonly implemented because of its strict keying requirements — the key must be as long as the message to be encrypted, truly random, and never used.

QKD's appeal is generally found in its ability to meet these requirements by generating unlimited amounts of unconditionally secure random keys, thus making previously unrealistic OTP configurations possible and gaining worldwide interest.

Quantum Security

The security of QKD is based on quantum uncertainty, where directly measuring a qubit changes its quantum state. In classical communications, bits exist in a deterministic state of either 0 or 1 where they can be measured and remeasured as necessary. In contrast, qubits are encoded in a simultaneous combination or "superposition" of states. This means a qubit probabilistically exists in a state between 0 and 1, typically described in a two-dimensional vector space $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$. Probabilistically, a 0 occurs with likelihood $|\alpha|^2$ and a 1 occurs with likelihood $|\beta|^2$, where $|\alpha|^2 + |\beta|^2 = 1$. This quantum phenomenon forms the foundation of QKD's unconditional

As networked devices, QKD systems are subject to a myriad of vulnerabilities in software, hardware, and system interfaces, requiring stringent security practices such as operating from a known secure state and continuous monitoring.

security claim, where attempting to copy or "clone" qubits randomly prepared in BB84's two conjugate bases (⊕ and ⊗) increases the system's quantum bit error Rate (QBER) [7]. For example, during an intercept-resend (man in the middle) attack, where an eavesdropper, Eve, attempts to measure and replace photons during key generation, she must pick a measurement basis (⊕ or ⊗) and will inevitably be wrong 50 percent of the time. When she selects the incorrect basis, she will randomly select the right bit value (0 or 1) 50 percent of the time. Overall, she will determine the correct bit value 75 percent of the time; thus, she introduces a readily detectable 25 percent QBER in key she resends to Bob. The unconditionally secure, theoretical nature of QKD rests on this unavoidable quantum phenomenon, which necessarily increases the system's QBER as captured in formal security proofs [6].

In practice, the source of all quantum bit errors is attributed to Eve in QKD security proofs. Unlike classical systems, quantum communication error rates are relatively high with operational QBERs of 3–5 percent due to device imperfections and transmission errors with theoretical QBER thresholds set at 11 percent with practical limits as low as 8 percent [6]. If the QBER threshold is ever exceeded, it is assumed an adversary is interfering with the key exchange, and the key generation process is aborted. Therefore, the QBER must be closely monitored and well characterized as every QKD architecture is a unique implementation. For further explanation see [5, 6].[2]

Practical Engineering Limitations and Security Vulnerabilities

While information theorists have proven the "unconditional security" of QKD, several practical security concerns remain unanswered as these systems do not undergo rigorous security hardening or formal certification [8]. As networked devices, QKD systems are subject to a myriad of vulnerabilities in software, hardware, and system interfaces, requiring stringent security practices such as operating from a known secure state and continuous monitoring. Moreover, as cryptographic devices, they should be physically protected against reverse engineering and side channel attacks through anti-tamper, code obfuscation, RF shielding, and other fundamental security techniques.

More specifically, QKD systems have practical engineering limitations and device non-idealities inherent to their design and implementation, which introduce vulnerabilities into the shared secret key generation process. For example, QKD systems are vulnerable to attacks over the quantum channel, including man in the middle (authentication failures), photon splitting (stealing photons), and blinding optical receivers (overpowering photon detectors to control their detections). In this article, we focus on multiphoton vulnerabilities associated with decoy state enabled QKD systems as they are commonly employed and provide a rich example of trade-offs between theoretical security, desired key generation rates, and real-world system implementation limitations. We describe this complex security issue in a systematic way with respect to the sender (Alice), quantum communication channel, and receiver (Bob).

Sender: Alice

Ideally, Alice would generate perfectly encoded qubits using a perfect on-demand single photon source; however, such devices are not currently available, nor are they expected in the near term. Instead, she uses a laser to produce classical optical pulses (i.e., pulses with millions of photons), which are attenuated to "weak coherent" levels (i.e., pulses with a mean photon number, MPN, of less than 1 photon per pulse). These sub-quantum energy levels are represented using a Poisson distribution. This means, for example, when the MPN = 0.1, 90.48 percent of the pulses have no photons, 9.05 percent have one photon, and 0.47 percent have two or more photons. While higher MPNs are desired to increase quantum throughput, low MPNs are chosen in an attempt to retain QKD security requirements, as each multiphoton optical pulse exposes information about the secret key to eavesdroppers.

Quantum Communications Channel

Because QKD optical pulses operate with low MPNs (e.g., 0, 1, 2, or 3 photons per pulse), propagation loss on the quantum channel severely constrains system performance. For example, since most QKD systems utilize existing single-mode fiber infrastructure with losses of 0.2 dB per km at 1550 nm, a propagation distance of 20 km has $\gtrsim$ 4 dB loss, and a 50 km link has $\gtrsim$ 10 dB loss. While greater operational distances are possible over fiber, desired key rates generally limit practical network distances to < 100 km.

When considering freespace propagation, losses are largely dependent on changing atmospheric conditions, line of sight, and pointing and tracking mechanisms. While not generally conducive to sustained QKD operation, recent experiments toward satellite-to-ground QKD have been conducted with total losses reported at $\gtrsim$ 30 dB near 800 nm wavelengths.

Additionally, timing synchronization and accurate basis alignment are limiting factors over the quantum channel (i.e., precise timing synchronization is required as pulse durations are 10^{-10} s, while device alignment imperfections generally contribute to ~1 percent QBER). There are multiple approaches to address these problems; however, they are generally limited by the performance of the compensation devices.

These non-idealities typically result in lower key rates and introduce additional quantum bit errors.

Receiver: Bob

In QKD systems, specialized single photon detectors (SPDs) generate "clicks" in response to minute energy levels corresponding to single photons (~1.28×10^{-19} J/photon at 1550 nm). The most prevalent type of SPD is the avalanche photodiode (APD), which has detection efficien-

[2] In addition to the security provided by quantum phenomenon, transactional authentication is often assumed in QKD.

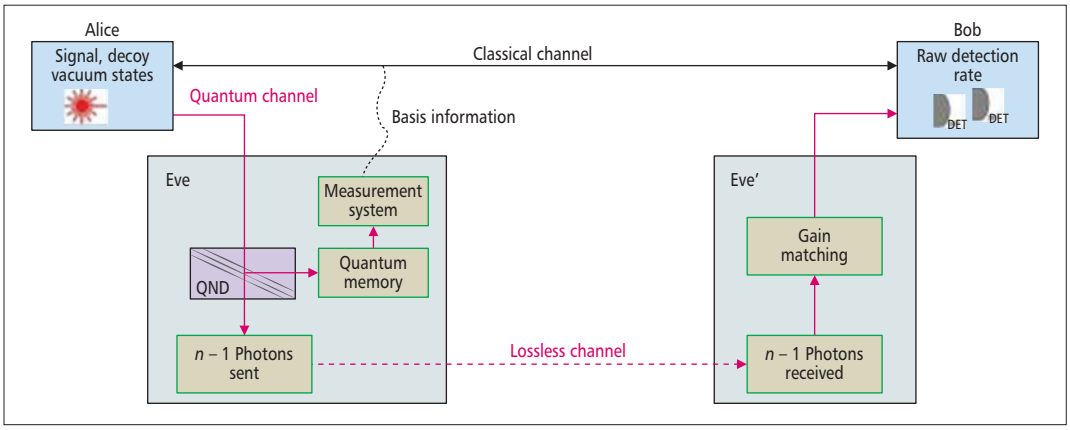


Figure 2. A generalized depiction of Eve's photon number splitting attack conducted against Alice's and Bob's decoy state enabled QKD system.

cies of 10–20 percent at 1550 nm [9]. These devices are coupled with advanced control circuitry to increase performance, where precise detection "gates" are used to reduce system noise, and detection "avalanches" are actively quenched to reduce relatively long recovery times (10^{-5} s). These recovery times limit the sender's pulse rate to ~10 MHz and thus overall quantum throughput. Despite seemingly poor performance, APDs are commonly implemented because they operate with low-cost thermo-electric coolers, unlike high-performance SPDs (i.e., >90 percent detection efficiency) such as superconducting nanowire SPDs (SNSPDs) or transition edge sensor (TES) arrays, which require elaborate multi-stage cooling devices to achieve near absolute zero operating temperatures [9].

Vulnerabilities in QKD Implementations

Each of the above engineering limitations and non-idealities contributes to practical security vulnerabilities, where the use of highly attenuated classical laser sources exposes information to eavesdroppers. Furthermore, losses and errors associated with poor propagation provide opportunities for eavesdropping without detection. In response to these vulnerabilities, a powerful theoretical attack, the photon number splitting (PNS) attack, was developed to gain maximum information on the secret key [6].

Photon Number Splitting Attacks

Figure 2 depicts the eavesdropper, Eve, executing a PNS attack against the subject QKD system where both Eve and Eve′ are necessary to account for the geographic distance between Alice and Bob. As is consistent with QKD security proofs, we consider an all-powerful Eve, limited only by the laws of quantum mechanics [10]. Eve is allowed to interfere with the quantum channel (introducing losses and errors) and listen to the classical channel.

In this configuration, Eve replaces the quantum channel with a lossless channel (achieved through quantum entanglement) and uses a quantum non-demolition (QND) measurement to determine the number of photons n in each pulse generated by Alice [10]. If $n \leq 1$, Eve blocks the pulse and sends nothing to Bob. If $n \geq 2$, Eve stores one photon in a quantum memory and sends the remaining $n - 1$ photons to Bob via the lossless channel. Because of this lossless channel, Eve′ must apply attenuation matching in order to not exceed Bob's expected detection rate and thus avoid becoming too obvious. This attack allows Eve to store a copy of Alice's and Bob's raw key buffer, and once Alice and Bob announce their measurement basis information over the classical channel, Eve is able to correctly measure her stored photons, generating a copy of the QKD-generated "unconditionally secure" shared secret key.

Decoy State Protocol

The decoy state protocol was introduced in 2003 as a means of detecting PNS attacks [11] and quickly improved [12]. It is now widely implemented in a number of QKD systems and technology demonstrations, as it simultaneously increases both performance (delivered secret key rate) and security (ability to detect PNS attacks). As described in Table 1, decoy state enabled QKD systems typically utilize a conventional "signal" state plus two dedicated security states: "decoy" and "vacuum" [13].

The decoy state protocol is relatively easy to implement as it extends the BB84 protocol with little additional hardware/software. During quantum exchange, Alice randomly generates optical pulses in one of the three states (signal, decoy, or vacuum) according to their occurrence percentage and unique MPN. For example, according to Table 1, the signal state is transmitted 70 percent of the time with an MPN of 0.6, the decoy state is transmitted 20 percent of the time with an MPN of 0.2, and the vacuum state is transmitted 10 percent of the time with an MPN near zero. In order to maintain integrity of the protocol, each pulse must have identical characteristics (e.g., wavelength, duration, and shape) other than the MPN, such that Eve cannot distinguish a decoy state from a signal or vacuum state.

During sifting, Alice and Bob announce the bases used to prepare and measure each pulse along with their state (signal, decoy, or vacuum). Communication errors in each state must then

The decoy state protocol was introduced in 2003 as a means for detecting PNS attacks [11] and quickly improved. It is now widely implemented in a number of high performance QKD systems and technology demonstrations, as it simultaneously increases both performance (delivered secret key rate) and security (ability to detect PNS attacks).

State	State description	Mean photon number (MPN)	Occurrence percentage
Signal "μ"	The signal state is used to transmit weak coherent optical pulses for generating shared secret key and facilitates increased key distribution rates through the use of higher MPNs.	0.6	70%
Decoy "ν"	The decoy state is used to detect PNS attacks on the quantum channel through statistical differentiation with the signal state.	0.1	20%
Vacuum "Y_0"	The vacuum state is used to determine the system's noise level, known as "dark count" when no photons are present at the detectors.	~0	10%

Table 1. Example decoy state protocol configuration.

be identified, counted, and corrected. For the signal state this is accomplished through complex error reconciliation algorithms, while the decoy and vacuum states can be accomplished through simple comparisons over the classical channel because they do not contribute to the final secure key. Next, a number of decoy state protocol calculations are made from system measurements (see the next section for details). Finally, the estimated signal and decoy photon number dependent yields Y_n^{signal} and Y_n^{decoy} are compared, where Y_n is the conditional probability of detecting a pulse at Bob given Alice sent an n-photon pulse.

When operating normally, the security condition $Y_n^{\text{signal}} = Y_n^{\text{decoy}}$ should always be true for a given QKD architecture. This is because the estimated yields depend primarily on the quantum channel efficiency and the number of photons per pulse ($n = 1, 2, 3, \ldots, n$), not the state type (signal or decoy). Thus, any deviation from the security check is an indication of adversarial eavesdropping [12].

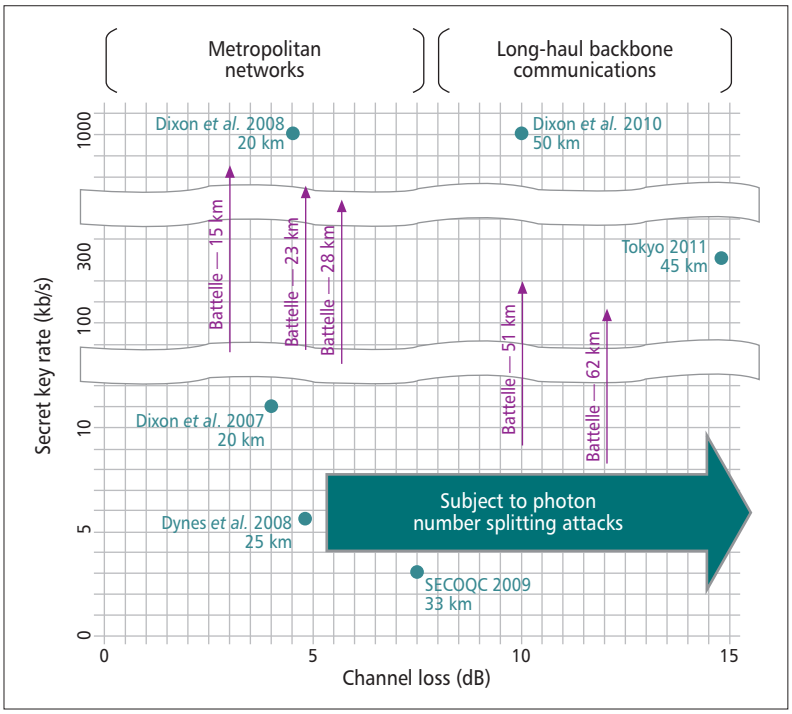


Figure 3. Results from recent experimental network demonstrations categorized into metropolitan and backbone distances, with longer propagation distances (and their corresponding higher channel losses) being subject to PNS attacks.

Efficiency Based Security Condition

In this section, we propose a more precise variation of the decoy state security condition that allows PNS attacks to be detected without a priori knowledge of the quantum channel efficiency. This approach is particularly advantageous as it is nearly impossible to guarantee a known secure state of the quantum channel over many kilometers. Similarly, it is difficult to ensure secure system performance [14].

During operation of decoy state enabled systems, the signal and decoy state yields are estimated from the measured gains Q^{signal}, Q^{decoy}, and the dark count rate Y_0

$$Q^{\text{signal}} = \frac{\text{Number of detections during signal pulses}}{\text{Total number of signal pulses sent}} \quad (1)$$

$$Q^{\text{decoy}} = \frac{\text{Number of detections during decoy pulses}}{\text{Total number of decoy pulses sent}} \quad (2)$$

$$Y_0 = \frac{\text{Number of detections during vacuum pulses}}{\text{Total number of vacuum pulses sent}} \quad (3)$$

The gain Q is the product of the probability of Alice sending out an n-photon pulse (a Poisson distribution) and the conditional probability of Alice's n-photon pulse leading to a detection event at Bob (Y_n) formally represented as

$$Q = \sum Poisson(n|\mu) * Y_n = \sum \frac{\mu^n e^{-\mu}}{n!} * (Y_0 + 1 - (1-\eta)^n) \quad (4)$$

where μ is the signal MPN (or ν is the decoy MPN), n is the number of photons in each pulse leaving Alice, Y_0 is the dark count rate, which is typically characterized during calibration activities, and η is the system's measured quantum efficiency, including the quantum channel, Bob's optical components, and Bob's detector efficiency. Summing Eq. 4 for all n and solving for η, we obtain

QKD system	Reported Operational Configuration					Simulation Results		
	Distance (km)	Signal MPN	Decoy MPN	Occurrence percentage (signal/decoy/ vacuum)	Reference quantum η (signal/decoy)	Normal quantum η (signal/decoy)	Efficiency security tolerance	PNS quantum η (signal/ decoy)
Dynes 2007 [15]	20	0.55	0.098	92.2/6.2/1.6	2.299E-2/ 2.254E-2[1]	2.288E-2/ 2.290E-2	2.834E-3	Below PNS Threshold
Dixon 2008 [16]	20	0.55	0.10	80/16/4	1.503E-2/ 1.524E-2[1]	1.543E-2/ 1.542E-2	1.294E-3	Below PNS Threshold
Dynes 2008 [17]	25	0.425	0.204	75/25/0	1.986E-2/ 1.918E-2[1]	2.180E-2/ 2.179E-2	7.924E-4	Below PNS Threshold
SECOQC 2009 [18]	33	0.48	0.16	92/6/2	6.020E-3[2]	6.008E-3/ 5.997E-3	5.687E-4	6.230-3/ 2.346E-3
Tokyo 2011 [19]	45	0.5	0.1	98.83/0.78/0.39	2.979E-3[2]	3.217E-3/ 3.476E-3	1.167E-3	3.222E-3/ 6.888E-4
Dixon 2010 [20]	50	0.5	0.1	98.83/0.78/0.39	7.293E-3/ 7.303E-3[1]	7.396E-3/ 7.382E-3	1.292E-3	7.449E-3/ –5.084E-5[3]

Notes: [1] Numbers are calculated from reported signal gains. [2] Numbers are estimated from reported losses and efficiencies. [3] The efficiency check becomes negative when the dark count rate exceeds the measured state gain (noise is greater than the signal strength).

Table 2. Quantum key distribution system experimental results from recent Toshiba Research Europe Ltd. Research efforts.

$$\eta = \frac{-\ln\left|1 + Y_0 - Q\right|}{\mu} \tag{5}$$

where signal and decoy efficiencies can be directly calculated and compared from the measured gains Q^{signal}, Q^{decoy}. We also realize that due to non-ideal devices and probabilistic single photon sources, we expect variations Δ in the calculated efficiency and the security condition becomes

$$\eta^{\text{signal}} = \eta^{\text{decoy}} \pm \Delta \tag{6}$$

The proposed secure condition should always be true unless a PNS attack is occurring. In contrast, $\eta^{\text{signal}} \neq \eta^{\text{decoy}} \pm \Delta$ implies eavesdropping on the quantum channel.

Examining Decoy State Enabled QKD Networks

While the decoy state protocol was introduced to detect PNS attacks, to date it has been primarily used to increase key generation rates and demonstrate the viability of QKD as an unconditionally secure key distribution solution.

High-Performance Decoy State QKD Networks

QKD is a method for growing a secret key between two authorized users, and therefore a point-to-point connection. However, over the past decade several experimental multi-node QKD networks have been created around the world: the DARPA Quantum Network in 2002, the European SECOQC network in 2008, the SwissQuantum network in 2009, and the Tokyo QKD network in 2010. Additionally, two significant network demonstrations are planned for the near future with the 5th International Conference on Quantum Cryptography to be held in Tokyo, Japan in late 2015 and the U.S.-based research group, Battelle, building a 1000 km trusted node network in 2016. In general, each of these networks employs several QKD systems over metropolitan distances of ~20 km and typically one long-haul communication link of 50–60 km. For this article, we chose to examine six fielded high-performance decoy state enabled QKD systems by Toshiba Research Europe, Ltd.

Figure 4 shows experimental results from Toshiba's efforts according to reported key generation rates and quantum channel losses [15–20]. Planned communications links supporting Battelle's QKD network around Columbus, Ohio are also shown [1]. Of the six QKD architectures considered, three are subject to PNS attacks while meeting Bob's expected detection rate. This is because Eve's PNS attack blocks a significant number of single photon pulses inducing ~6 dB loss for the described operating regimes. Therefore, sufficient loss must be available on the quantum channel for Eve to compensate for using her lossless channel.

Examination of the Efficiency-Based Security Condition

Table 2 provides configuration details and simulation results for the six identified Toshiba QKD systems from Fig. 3. Each system is reported with relevant operational parameters (distance, MPN, and occurrence percentages) and its calculated quantum efficiency for both the signal

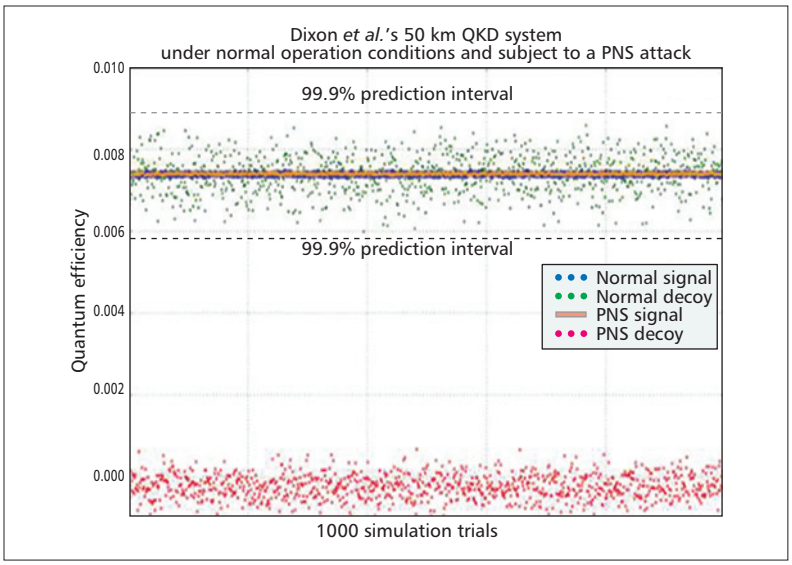


Figure 4. Simulation results from normal operating conditions and when subject to a photon number splitting attack.

and decoy state. Results of the simulation study are shown for normal operating conditions and when subject to the PNS attack, including the decoy state tolerances used to detect Eve.

For this experiment, security tolerances were established from calibration scenarios of 100 trials each with 100,000 detections. Based on the modeled system's performance, we chose security tolerances of 99.9 percent from the expected decoy state mean to detect the PNS attack with high confidence and minimize false positives. Operational tolerances are established from the decoy state to account for expected variations in system performance because of its lower MPN and occurrence percentage. Additionally, increasing statistical confidence can be achieved through repeated rounds of quantum exchange as desired.

Each system was modeled using a QKD experimentation framework [21] with average efficiencies reported from 1000 simulations with more than 2 billion pulses sent for each scenario of interest. Comparing the reference and normal simulated quantum efficiencies, they are in close agreement, validating the model. From a practical security perspective, these efficiencies can also be compared against physical measurements of the quantum channel, providing additional evidence to justify the system's security posture.

When Eve conducts PNS attacks, the signal state efficiency remains relatively unchanged, while the decoy state efficiency drops considerably. This is primarily due to the difference in MPNs, where Eve blocks a majority of the lower MPN decoy state pulses and must send slightly more signal state pulses in order to meet Bob's expected detection rate. A detailed example of this behavior is presented in Fig. 4 based on Dixon *et al.*'s 50 km QKD link [20].

Figure 4 depicts the signal and decoy efficiencies under normal operating conditions and the resulting efficiencies when Eve is conducting a PNS attack. The overlapping signal and decoy efficiencies (i.e., within the 99.9 percent security tolerances) imply $\eta^{signal} = \eta^{decoy}$ and indicate secure operation. Conversely, Eve's PNS induced decoy state efficiency is significantly reduced, clearly outside the 99.9 percent tolerances.

As indicated by the relatively large difference between the PNS induced decoy state efficiency and the normal efficiencies, Eve's PNS attack is readily detectable when properly configured. In all 3000 simulations conducted, the PNS attack was successfully detected. This simulation study illustrates how the efficiency-based decoy state security condition can be used to easily and accurately detect PNS attacks without assuming a known secure quantum channel.

Conclusions

In this article, we describe QKD systems as part of a secure communications solution, and specifically examine the performance of decoy state enabled QKD systems in support of metropolitan and long-haul quantum communication networks. A more accurate variation of the decoy state protocol is proposed that does not require a priori knowledge of the quantum channel. We believe this to be an advancement toward practical QKD architectures, as the system can quickly verify the security posture of the quantum channel against PNS attacks and, more generally, photon number specific interference. Future work includes performance-security characterization of the decoy state protocol to aid in the understanding, evaluation, and development of secure QKD network architectures for use in high-security environments.

Acknowledgments

This work was supported by the Laboratory for Telecommunication Sciences [grant number 5743400-304-6448]. This work was supported in part by a grant of computer time from the DoD High Performance Computing Modernization Program at the Air Force Research Laboratory, Wright-Patterson AFB, Ohio.

Disclaimer

The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, the Department of Defense, or the U.S. Government.

References

[1] L. Oesterling, D. Hayford, and G. Friend, "Comparison of Commercial and Next Generation Quantum Key Distribution: Technologies for Secure Communication of Information," *2012 IEEE Conf. Technologies for Homeland Security*, 2012.
[2] S. Wiesner, "Conjugate Coding," *ACM Sigact News*, vol. 15, no. 1, 1983, pp. 78–88.
[3] C. H. Bennett and G. Brassard, "Quantum Cryptography: Public Key Distribution and Coin Tossing," *Proc. IEEE Int'l. Conf. Computers, Systems and Signal Processing*, 1984.
[4] S. Imre, "Quantum Communications: Explained for Communication Engineers," *IEEE Commun. Mag.*, vol. 51, no. 8, 2013, pp. 28–35.
[5] L. O. Mailloux *et al.*, "Performance Evaluations of Quantum Key Distribution System Architectures," *IEEE Security and Privacy*, vol. 13, no. 1, pp. 30–40, 2015.
[6] V. Scarani *et al.*, "The Security of Practical Quantum Key Distribution," *Rev. Modern Physics*, vol. 81, no. 3, 2009, pp. 1301–50.
[7] W. K. Wootters and W. H. Zurek, "A Single Quantum Cannot Be Cloned," *Nature*, vol. 299, no. 5886, 1982, pp. 802–03.

[8] V. Scarani and C. Kurtsiefer, "The Black Paper of Quantum Cryptography: Real Implementation Problems," arXiv:0906.4547v2, 2009.
[9] R. H. Hadfield, "Single-Photon Detectors for Optical Quantum Information Applications," *Nature Photonics*, vol. 3, no. 12, 2009, pp. 696–705.
[10] G. Brassard *et al.*, "Limitations on Practical Quantum Cryptography," *Phys. Rev. Lett,*, vol. 85, no. 6, 1330, 2000.
[11] W.-Y. Hwang, "Quantum Key Distribution with High Loss: Toward Global Secure Communication," *Phys. Rev. Lett.*, vol. 91, no. 5, 057901, 2003.
[12] H.-K. Lo, X. Ma, and K. Chen, "Decoy State Quantum Key Distribution," *Phys. Rev. Lett.*, vol. 94, no. 3, 230504, 2005.
[13] X. Ma *et al.*, "Practical Decoy State for Quantum Key Distribution," *Phys. Rev.*, vol. 72, no. 1, 012326, 2005.
[14] Z. L. Yuan, A. W. Sharpe, and A. J. Shields, "Unconditionally Secure One-Way Quantum Key Distribution Using Decoy Pulses," *Appl. Phys. Lett.*, vol. 90, no. 1, 011118, 2007.
[15] J. F. Dynes *et al.*, "Practical Quantum Key Distribution over 60 Hours at an Optical Fiber Distance of 20km Using Weak and Vacuum Decoy Pulses for Enhanced Security," *Optics Express*, vol. 15, no. 13, 2007, pp. 8465–71.
[16] A. R. Dixon *et al.*, "Gigahertz Decoy Quantum Key Distribution with 1 Mbit/s Secure Key Rate," *Optics Express*, vol. 16, no. 23, 2008, pp. 18,790–18,979.
[17] J. F. Dynes *et al.*, "Decoy Pulse Quantum Key Distribution for Practical Purposes," *Optoelectronics*, IET, vol. 2, no. 5, 2008, pp. 195–200.
[18] M. Peev *et al.*, "The SECOQC Quantum Key Distribution Network in Vienna," *New J. Physics*, vol. 11, no. 7, 075001, 2009.
[19] M. Sasaki *et al.*, "Field Test of Quantum Key Distribution in the Tokyo QKD Network," *Optics Express*, vol. 19, no. 11, 2011, pp. 10,387–10,409.
[20] A. R. Dixon *et al.*, "Continuous Operation of High Bit Rate Quantum Key Distribution," *Appl. Phys. Lett.*, vol. 96, no. 16, 2010, p. 161,102.
[21] L. O. Mailloux *et al.*, "A Modeling Framework for Studying Quantum Key Dstribution System Implementation Non-Idealities," *IEEE Access*, 2015, pp. 110–30.

Biographies

Logan O. Mailloux [M] (Logan.Mailloux@afit.edu, Logan-Mailloux@yahoo.com), CISSP, CSEP (B.S. 2002, M.S. 2008) is a commissioned officer in the United States Air Force and a Ph.D. candidate at the Air Force Institute of Technology (AFIT), Wright-Patterson AFB, Ohio. His research interests include system security engineering, complex information communication and technology implementations, and quantum key distribution systems. He is a member of Tau Beta Pi, Eta Kappa Nu, INCOSE, and ACM.

Michael R. Grimaila [SM] (Michael.Grimaila@afit.edu, MichaelGrimaila@yahoo.com), Ph.D., CISM, CISSP (B.S. 1993, M.S. 1995, Ph.D. 1999) is a professor, head of the Systems Engineering Department, and member of the Center for Cyberspace Research (CCR) at AFIT, Wright-Patterson AFB. He is a member of Tau Beta Pi, Eta Kappa Nu, and ACM, and a Fellow of the ISSA. His research interests include computer engineering, mission assurance, quantum communications and cryptography, data analytics, network management and security, and systems engineering.

John M. Colombi [M] (B.S. 1986, M.S. 1992, Ph.D. 1996) is an associate professor of systems engineering at AFIT, Wright-Patterson AFB. His research interests include systems and enterprise architecture, complex adaptive systems, acquisition process modeling, and human systems integration. He is a member of INCOSE.

Douglas D. Hodson, Ph.D. (B.S. 1985, M.S. 1987, Ph.D. 2009) is an assistant professor of computer engineering at AFIT, Wright-Patterson AFB. His research interests include computer engineering, software engineering, real-time distributed simulation, and quantum communications. He is also a DAGSI scholar and a member of Tau Beta Pi.

Ryan D. Engle [M] (B.S. 2007, M.S. 2015) is a commissioned officer in the U.S. Air Force and a Ph.D. student at AFIT, Wright-Patterson AFB. His research interests include software engineering, computer engineering, model-based systems engineering, modeling and simulation, and quantum key distribution systems. He is a member of ACM.

Colin V. McLaughlin (B.A. 2003, Ph.D. 2010) is a research physicist at the U.S. Naval Research Laboratory, Washington, DC. He specializes in photonic communication devices and systems.

Gerald Baumgartner, Ph.D. (B.S. 1971, M.S. 1973, Ph.D. 1980, Illinois Institute of Technology) is a research physicist at the Laboratory for Telecommunications Sciences, College Park, Maryland. He is a member of the American Physical Society, the Optical Society of America and the Society for Industrial and Applied Mathematics. His research interests include quantum optics, quantum communications, quantum information, communications security, communications system modeling and simulation, and statistical signal processing.

A more accurate variation of the decoy state protocol is proposed that does not require a priori knowledge of the quantum channel. We believe this to be an advancement toward practical QKD architectures, as the system can quickly verify the security posture of the quantum channel against PNS attacks and, more generally, photon number specific interference.

Design Considerations in Applying Disruption Tolerant Networking to Tactical Edge Networks

Rahul Amin, David Ripplinger, Devanshu Mehta, and Bow-Nan Cheng

ABSTRACT

In recent years there has been a strong desire in the U.S. Department of Defense to augment traditional ground and satellite communications with a high capacity aerial tier. High capacity airborne links are often directional in nature and highly affected by aircraft body blockage, exhibiting unique outage characteristics compared to ground or satellite networks. To mitigate the effects of periodic link outages that last seconds to minutes, disruption tolerant networking (DTN) technology has been proposed. In this article we examine applying the DTN Bundle Protocol (RFC 5050) to ship-to-shore networks for traffic flowing over the aerial nodes. Specifically, we examine applying DTN proxies, tunnels, and interfaces to both the plaintext and ciphertext side of military networks to understand architecture and design considerations and limitations.

INTRODUCTION

In deployed military networks, disparate surface nodes are often unable to communicate effectively with each other through line-of-sight radios due to long distances, terrain blockages, or adversarial attack. In addition, satellite communications may also be degraded due to a number of factors. An aerial high capacity backbone (HCB) network can seamlessly bridge between terrestrial and aerial communication segments to connect users from the tactical edge to command centers [1]. The aerial HCB consists of a set of manned and unmanned aircrafts that provide range extension to surface nodes while also providing reach back to the global information grid (GIG). These capabilities are achieved through a downward facing antenna on the aircrafts that acts as a surface-to-surface communication relay, as well as directional, high data rate cross link antennas that connect the aircrafts to each other and to a GIG entry-point, forming a backbone. In order to maintain stable end-to-end network connectivity, the system needs to provide high availability on each of the component links. The performance of these links (the uplink, downlink, and cross-links) is heavily dependent on the system dynamics that include the movement of the aircraft, the orbit the aircraft flies in, the placement of antennas, and the blockage characteristics of the platform.

In previous work that evaluated a multi-hop airborne IP backbone with heterogeneous radio technologies [2], it was shown that the average air-to-ground link availability was 75 percent and the average air-to-air link availability was 71.8 percent, with outages on the order of minutes. All links in the multi-hop airborne IP backbone suffer from various factors such as aircraft body blockage, aircraft location, aircraft orbits, and weather. The combined effect of these outages is that end-to-end connection availability between surface nodes is often very poor, suffering from long periods of lost connectivity when at least one of the component links is unavailable or when routes need to be re-established. The outages are often long enough to cause link layer retransmission timeouts, route failures, and TCP timeouts. Furthermore, multiple poor links along a path result in a compounded loss rate as data is dropped after traversing part-way. Delay (or disruption) tolerant networking (DTN), which spans the network and application layers of the OSI stack, can be used to ensure delivery across these long outages by storing the data at intermediary nodes during outages.

There are several key challenges when applying DTN techniques to aerial high capacity backbone networks, as shown in Fig. 1. First, there is currently a set list of traditional TCP/IP based applications deployed. Because of the challenges of upgrading ship/shore systems and the training required, it is difficult to completely swap out all current applications with DTN-enabled applications. Second, ciphertext/plaintext boundaries affect DTN capabilities, depending on where the DTN technologies are placed. Although there has been some work in applying DTN technologies to tactical edge networks, most previous papers either assume completely new applications [3, 4], or in the case of leveraging application layer proxies or gateways that adapt current applications to use DTN, do not consider how encryption would affect architecture design [5–7].

In this article we examine tradeoffs in applying the DTN concept to representative airborne

The authors are with MIT Lincoln Laboratory

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

0163-6804/15/$25.00 © 2015 IEEE

high capacity network scenarios to improve the end-to-end packet delivery rate of several user applications. We present several solutions that offer varying tradeoffs in terms of steps required to conform to military security architectures, ease of deployment, and modifications required to support existing applications. We show that although DTN has been studied extensively as a stand-alone solution, applying DTN to tactical edge networks comes with significant additional challenges due to security boundaries and pre-existing applications. While we use a ship-to-shore scenario as an example tactical edge network, techniques presented in this article apply to many similar military tactical edge networks with security boundaries. While the Bundle Security Protocol (BSP) and related efforts in the Internet Engineering Task Force are promising in this context, we are limiting our discussion in this article to approaches that are compatible with the current military IP security architecture.

The article is organized as follows. We provide background on the suite of technologies encompassed by the term DTN. We discuss design considerations for applying DTN solutions in tactical edge networks. We describe several proposed DTN architectural options and discuss pros and cons of each solution. We conclude the article and discuss unsolved challenges for DTN solutions in a military environment as a part of future work.

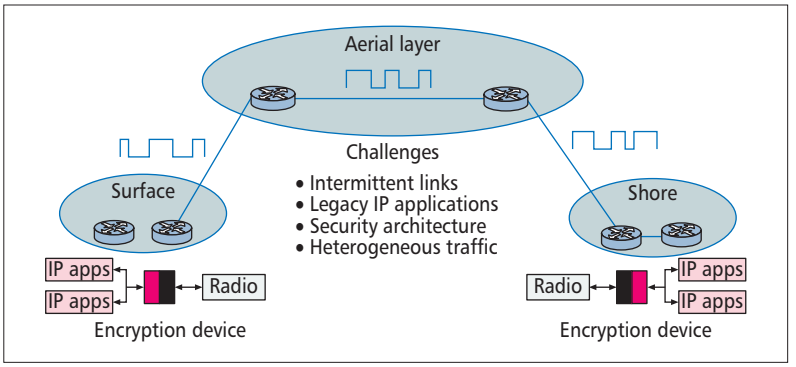


Figure 1. Main challenges in tactical edge networks.

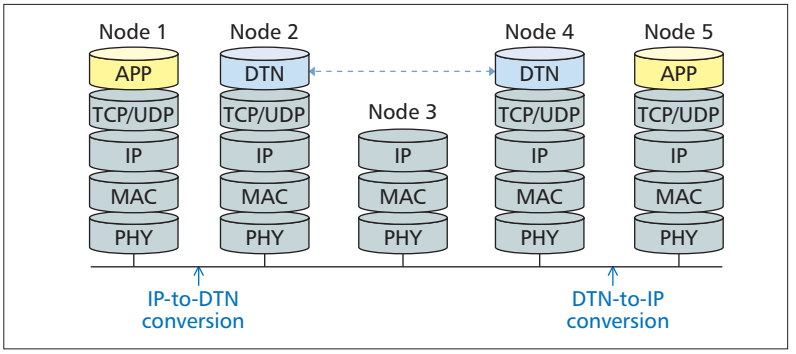


Figure 2. DTN example.

Disruption Tolerant Networking Overview

In general terms, DTN is a concept: providing data delivery in spite of long delays or disruptions. Delivery is assured by storing the data at intermediary DTN-capable nodes during disruptions and by using robust transport and link-layer protocols that can tolerate long disruptions and delays for transmission between DTN-capable nodes. In the Internet Research Task Force (IRTF) research community, DTN has come to be referred to as a specific suite of protocols built around the Bundle Protocol (RFC 5050 available: http://tools.ietf.org/html/rfc5050) that was originally developed for high delay deep space applications such as the InterPlanetary Network (IPN). At its core, the Bundle Protocol defines the intermediary format for storage and transmission of data at DTN-capable ("store-and-forward") nodes, in addition to protocol features that allow for reliable delivery across a high delay or disrupted network.

Traditional Internet protocols require the end-to-end path between source and destination nodes to be available simultaneously. DTN, on the other hand, is often compared to the postal service where packets are held at intermediary nodes ("post offices") that take responsibility for delivering the data ("packages") to the next DTN-capable node or to the destination. The act of storing the data and taking responsibility for it enables DTN to deliver data through severe disruptions, delays, and outages. An example of DTN usage is shown in Fig. 2, and its main features are discussed next.

There are several options as to how an application can make use of DTN's store-and-forward capability via the Bundle Protocol. A bundle is the common unit of transmission and storage of data in DTN. It is intended to be a complete application data unit, such that a receiving application has sufficient independent data and metadata to take action based on a received bundle without relying on subsequent bundles. An application can either be written so that it has native Bundle Protocol support where all data packets are converted to bundles before they are transmitted, or a proxy can be implemented that converts normal IP data packets to bundles so that DTN-capable nodes can store those bundles during network outages. Figure 2 shows an example of the latter approach where node 1 and node 5 implement a normal IP application. Node 1 is the source node and node 5 is the destination node for the implemented application. After node 1 sends out normal IP application packets, a proxy encapsulates the IP packets to DTN bundles before the bundles are sent to DTN-capable node 2. In addition, the proxy determines the DTN path the bundle needs to take based on the destination address. In Fig. 2, the DTN path exists from node 2 to node 4. Once node 2 receives the bundle, its job is to store and forward the bundle to node 4 as necessary. Figure 2 also illustrates that not all nodes in a DTN network have to be capable of storing and forwarding bundles. Node 3 is not a DTN-capable node. Node 2 uses a DTN routing protocol to determine the next appropriate DTN-hop, which is node 4 in our example. However, once node 2 has determined that node 4 is the next DTN hop, it utilizes the underlying IP routing protocol (e.g. OSPF) to determine that the bundles need to go through node 3 to get to

Figure 3. HCB scenario example.

node 4 even though node 3 does not have any store-and-forward capability. Thus, if any part of the path consisting of the DTN hop is not available, node 2 will keep storing bundles until all bundles can be successfully delivered to node 4. Once node 4 receives a bundle, the bundle needs to be converted to the original IP packet sent by the application on node 1. A second proxy is required that implements this bundle-to-IP conversion functionality. Once the bundle is converted to an IP packet, normal IP protocols take care of forwarding the packet to the actual destination (node 5 in our Fig. 2 example).

Design Considerations

We consider an example ship-to-shore scenario, which is representative of a common military tactical edge scenario, that connects dispersed ships as well as provides reach-back capability to the command center via an aerial HCB communication network, as shown in Fig. 3. Trajectories of the aircraft and number of simultaneous links that need to be established for the aerial HCB have a significant impact on overall network availability. A typical military application such as email and chat uses a client-server architecture that requires successful data transfer from ship-to-multiple aircrafts-to-shore and in the reverse direction. Each individual link in the end-to-end path has to be available for the application to work successfully in a normal IP network. As an example, trying to maintain a total of three simultaneous links (two uplink/downlink and one cross-link for the leftmost aircraft shown in Fig. 3) for a topology with each aircraft having four antennas and flying in a specific racetrack orbit with 7 degree bank angle, Satellite Tool Kit (available: http://www.agi.com/products/stk/stk-10) modeling produces average uptime of around 93 percent with average outage duration of approximately 13.2 seconds and largest outage duration of approximately 50 seconds. Moreover, when degrees of freedom decrease, the end-to-end path availability decreases even further and the average outage duration increases. Previous work [2] that studied link dynamics in an airborne HCB network using an electronic switch beam radio system showed that end-to-end availability measured using ping packets was between 64 percent and 77 percent with average link downtime of 85.2 seconds over measurements collected for a period of four days. Clearly, these end-to-end path outages in the order of minutes are going to cause a significant performance degradation for normal IP applications. Depending on the application and amount of traffic in the network, DTN can provide significant benefits over traditional IP networks.

In implementing a DTN solution, all additional equipment/software that support DTN functionality has to conform to the different military security levels and the associated red/black boundaries. The red (plaintext) side of the network handles unencrypted traffic, while the black (ciphertext) side of the network handles encrypted traffic. All applications reside on the red side network. Before the application data is transmitted over-the-air, all data packets are encrypted, and typically the radios that transmit these messages lie on the black side. The red and black boundary is separated by a security device that we represent as one endpoint of an IPsec tunnel. The security device that initiates and terminates an IPsec tunnel is only available on certain nodes, such as the aircraft carrier, in the topology. So the DTN solution, which processes either unencrypted or encrypted data depending on if the solution resides on the red side or black side network, respectively, is constrained in terms of where additional interfaces/ equipment that support DTN functionality are inserted in the topology.

The implemented DTN solution also has to consider which applications are to be supported via DTN's store-and-forward capability. The first choice that has to be made in supporting applications is to decide whether to add native support in the applications themselves to convert IP data packets to the required bundles, or to implement a proxy that performs the required IP-to-bundle conversion. If support is added natively, the new application has to pass military conformance tests and has to go through an approval process. Also, each individual application that will use DTN will have to be rewritten. If a proxy is implemented, all existing applications do not change. Only the placement of the proxy (either on the red side or the black side) has to obtain military approval. Next, not all applications in an operational tactical edge network can tolerate large latencies (e.g. voice and video). Traffic of applications that are delay tolerant (e.g. file transfer and email) has to be separated from non-delay tolerant application traffic (e.g. via DSCP values).

DTN Architectural Options

We investigate four DTN architectural options that present tradeoffs in terms of conformance to military security architecture, ease of deployment, and user applications that are supported under each solution. A high level overview of each of the four solutions is presented in Fig. 4. We discuss the implementation details of each solution in the following subsections.

DTN Applications Description

The first option involves creating *DTN applications*. Applications such as file transfer and email can be written so that they have built-in support for the Bundle Protocol's store-and-forward capability. Data packets created by the applications are encapsulated into bundles before they

are transmitted over the network. Since this option is applied directly at the application level, it is implemented on the red side network on ship/subs where applications originate. The bundles created by DTN applications are encrypted and sent over the black side network over the aerial links. The next DTN hop that is able to process the bundle on the red side are the aircraft carriers/shore. Thus, DTN code capable of processing the bundles and storing them (if needed) must be present on aircraft carriers/shore on the red side network.

DTN Interface Description

The *DTN interface* option is implemented as a transparent proxy (i.e. bump in the wire) and applications have no knowledge of its presence. The transparent proxy is implemented on the red side network of ships/subs, which gives it the capability to examine data packet contents and implement the required IP packet-to-bundle conversion. For UDP traffic, the conversion is pretty straightforward and does not really require examination of packet contents. But for TCP traffic, the transparent proxy has to break up the end-to-end TCP connection and convert it to hop-by-hop TCP connections. In doing so, an ACK packet for each TCP data packet arriving at the proxy is required, and the TCP sequence number from each packet has to be obtained before the appropriate ACK packet can be created. The goal is to ensure TCP does not throttle packet transmission by sending a fake ACK packet back to the sender. Again, the next DTN hop that is able to process the bundle on the red side are the aircraft carriers/shore. Thus, DTN code capable of processing the bundles and storing them (if needed) has to be present on aircraft carriers/shore on the red side network.

DTN Proxy Description

The *DTN proxy* option implements an explicit DTN proxy that converts traffic of current applications to support the DTN Bundle protocol. Delay tolerant applications on ships/subs configure proxy settings so that their traffic reaches the DTN proxy, where regular IP traffic is converted to bundles. The proxy is located on the red side network of ships/subs and performs the IP packet-to-bundle conversion after examining the contents of the packet. If the IP packet is using TCP, an ACK is sent back to the source and the end-to-end TCP connection is split into hop-by-hop TCP connections. The next DTN hops that process the bundle on the red side are the aircraft carriers/shore. Thus, DTN code capable of processing the bundles and storing them (if needed) has to be present on aircraft carriers/shore on the red side network.

The difference between the DTN interface and the DTN proxy approach is that the DTN interface is transparent to the application while the DTN proxy approach explicitly requires applications to point to the proxy server. The DTN interface has not been widely studied yet as all traffic (including non-DTN traffic) passes through a single possible point of failure and incurs packet processing delay. The proxy architecture is well understood and widely deployed in most information technology (IT) infrastructures. An HTTP-based DTN proxy described in [5] and jabber based DTN proxy described in [6] are examples of actual DTN proxy implementations. In our example topology shown in Fig. 3, if the links between ships/subs to the aircraft carriers/shore are fairly stable, it is possible to only implement the DTN proxy on the aircraft carrier/shore nodes and have the ship/sub nodes connect to the proxy on the aircraft carrier/shore. This limits the installation requirements at the cost of potential unavailability of the proxy.

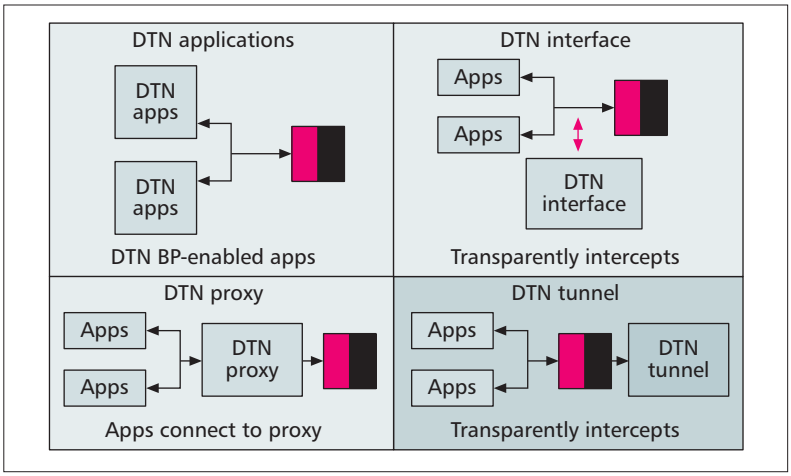


Figure 4. DTN solutions overview.

DTN Tunnel Description

The *DTN tunnel* option is implemented on the black side network of ships/subs as a transparent proxy applying DTN to encrypted data. Only limited information about the traffic passing through the proxy can be determined as the traffic is encrypted. Delay tolerant vs. non-delay tolerant traffic can be separated at the proxy via the use of tagging that passes through the encryption process untouched (e.g. DSCP values). The proxy converts the encrypted IP packets to bundles. The next DTN hop that processes the bundle on the black side are the aircraft carriers/shore. Thus, DTN code capable of processing the bundles and storing them (if needed) has to be present on aircraft carriers/shore on the black side network.

Implications of DTN Solutions in Tactical Edge Networks

A summary of pros/cons of each DTN solution option is presented in Table 1. We discuss the major implications of DTN solutions in tactical edge networks in this section in terms of availability, overhead, ease of deployment, applications supported, performance impacts on normal IP applications, and robustness.

Availability

Availability is defined as the packet completion percentage between an end-to-end path. The DTN application option results in the greatest availability gain compared to a non-DTN solution, as the store-and-forward capability is supported on the complete end-to-end data path. Any application that requires high availability can be rewritten

Option	Pros	Cons
DTN applications (red side)	• Greatest availability gain • All traffic types can be supported • Lower overhead	• Getting approval to install DTN code on the red side is challenging • Significantly rewrite current applications
DTN interface (red side)	• Good availability gain • All traffic types can be supported • All current applications can be supported without any modification	• Getting approval to install DTN code/proxy on the red side is challenging • Traffic has to be tagged to separate DTN/non-DTN traffic • Processing tagged traffic incurs unnecessary delay for non-DTN traffic • Possible single point of failure for all traffic • Higher overhead
DTN proxy (red side)	• Good availability gain • All traffic types can be supported	• Getting approval to install DTN code/proxy on the red side is challenging • Current applications have to be slightly modified to add proxy configuration support • Higher overhead
DTN tunnel (black side)	• Good availability gain • Ease of deployment/approval due to black side DTN code/proxy installation	• Traffic has to be tagged to separate DTN/non-DTN traffic • Processing tagged traffic incurs unnecessary delay for non-DTN traffic • Possible single point of possible failure for all traffic • Only UDP traffic can be supported • Higher overhead

Table 1. Pros/cons of DTN architectural options.

with built-in Bundle Protocol functionality support. All the other solutions (*DTN interface*, *DTN proxy*, and *DTN tunnel*) result in good availability gain compared to a non-DTN solution, as the store-and-forward capability is supported between all DTN proxies used by these solutions. The only data path segment that might reduce availability is the path between the application source-to-proxy and proxy-to-application destination. However, these paths are generally stable wired connections and do not impact availability.

Overhead

Each IP packet that gets converted to a bundle incurs bundle protocol header overhead. The overall overhead depends on the mapping of data packets to bundles. For *DTN application*, due to native application support, the application can optimize what goes in a single bundle (e.g. create small packets so multiple data packets form a single bundle) and can thus be the most efficient approach in terms of incurred overhead, whereas *DTN interface*, *DTN proxy* and *DTN tunnel* solutions do not have control over the packet size used by applications. Thus, additional bundle header overhead can be expected for these solutions if one data packet is mapped to one bundle.

Ease of Deployment

Ease of deploying a DTN solution depends on two factors:

- Changes that need to be made to the current application set.
- Getting approval for placing a DTN solution that can process packets at a specific security level.

The *DTN applications* solution requires replacement of existing applications with newly written Bundle Protocol enabled applications, which involves going through the military approval procedure that accepts the replacement of current applications. The drawback of this solution is that it requires change on all existing platforms. The *DTN proxy* solution uses an explicit DTN proxy in the network. Applications such as email and HTTP already come with proxy configuration support so that traffic from these applications can be configured to reach the DTN proxy. However, applications such as file transfer need to be rewritten so that a proxy setting can be configured. The *DTN interface* and *DTN tunnel* solutions process all traffic through an in-line proxy. Thus, for these solutions, applications do not have to be modified in any way to include DTN support. As far as the placement of a DTN solution is concerned, getting deployment approval for DTN solutions implemented on the red side of the network (*DTN application*, *DTN proxy*, and *DTN interface*) which processes unencrypted data is much more challenging than the one implemented on the black side of the network (*DTN tunnel*) which processes encrypted data.

Applications Supported

All delay tolerant applications benefit from DTN solutions implemented on the red side (*DTN applications*, *DTN proxy*, and *DTN interface*), but that is not the case with DTN solutions implemented on the black side (*DTN tunnel*). We specifically focus on UDP and TCP applications in our analysis. An example enhancement seen by UDP and TCP applications for both red side and black side solutions, respectively, is shown in Fig. 5. The network shown in Fig. 3 is used to generate the results, where traffic is sent over two DTN hops: carrier 1 (DTN)–aircraft 1–aircraft 2–carrier 2 (DTN)–aircraft 2–aircraft 3–carrier 3 (DTN). The traffic generator mgen is used to send constant bit rate traffic of one 1500 byte packet per second for a duration of 10 minutes

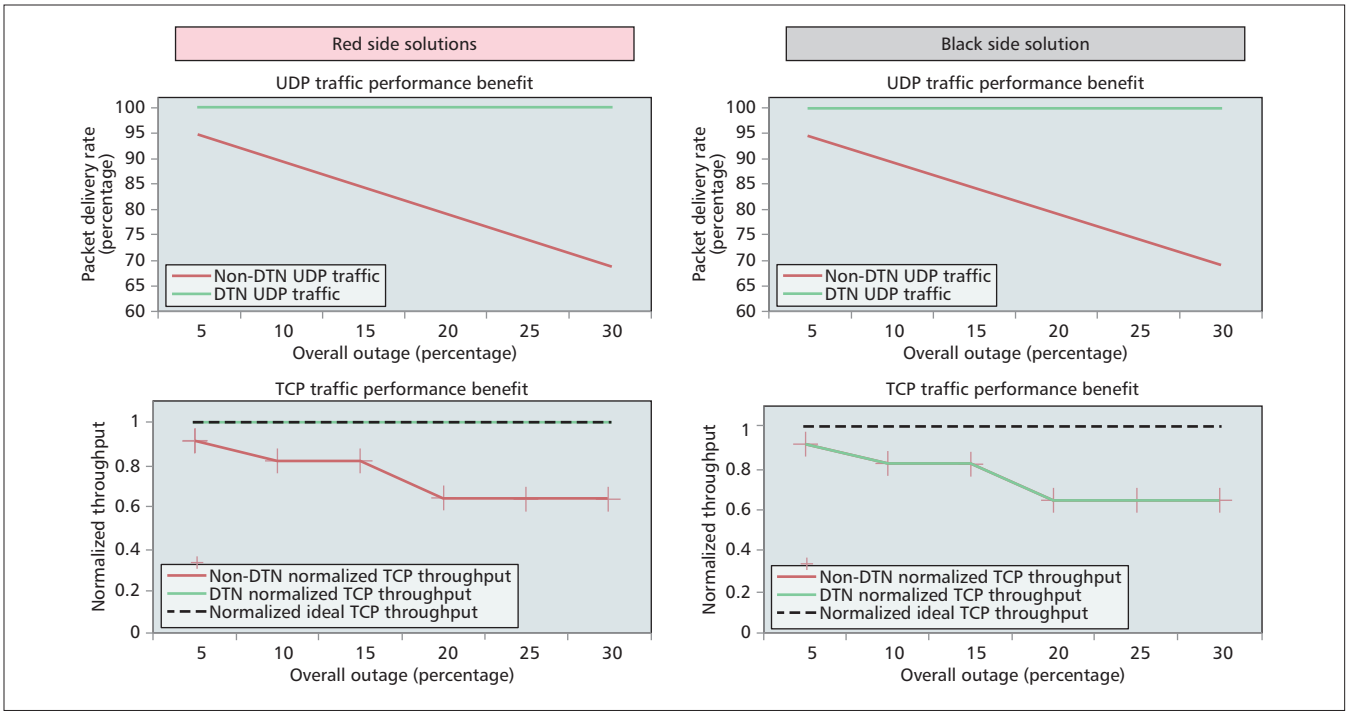


Figure 5. DTN performance benefit.

either using UDP or TCP. The aircraft 2–aircraft 3 link is taken down to generate the different percentages of path outages. As seen in Fig. 5, the performance of UDP applications is the same for both red side and black side solutions. Packets get stored by carrier 2 during link outages. When the aircraft 2–aircraft 3 link comes back up, all stored packets are then forwarded to carrier 3. For the non-DTN case, all packets are dropped during link outages, which results in packet loss equal to the percentage of link outages. The same phenomenon would be experienced in case of network congestion (instead of link outages), where more data was sent than could be supported over the data links.

Due to TCP's built-in congestion control and reliability mechanisms, TCP relies on ACK packets for every data packet that is sent to determine whether to continue transmitting data. If no ACK packets are received, TCP assumes the network is congested and throttles traffic. In the case of red side DTN solutions, the intermediary DTN hops can inspect TCP packet sequence numbers (carrier 2 in our example) and generate an ACK packet that gets sent to the original sender (carrier 1). This ensures that the sender continues to transmit TCP packets even during outages. With the black side DTN solution, packet content is unknown due to encryption, making it impossible to generate ACK packets early. Thus, the TCP session must wait until the destination sends ACKs, which could be delayed or lost. For the black side solution, this results in TCP timing out and throttling traffic, reducing the throughput. The reason for the gradual (and stepwise) decrease in non-DTN and black side DTN solution throughput is because of the TCP timeout that is experienced during the link outage and the doubling of retransmission timer due to TCP's exponential back-off mechanism, which makes packets wait approximately the same time for certain outage durations (e.g. 10 percent and 15 percent outages) before being retransmitted. It is important to note that because of TCP's reliability mechanisms, delivery success with and without DTN is roughly equal.

Performance Impacts on Normal IP Applications

An operational network consists of both delay tolerant traffic as well as non-delay tolerant traffic. Delay tolerant traffic is processed by DTN nodes and is stored during network outages. When the network comes back up, the stored traffic is forwarded by the DTN nodes. Depending on the convergence mechanism used by DTN, the delay tolerant traffic will impact the performance of non-delay tolerant traffic. For example, once the network comes back up, if the DTN solution forwards all stored packets at once, much of the non-delay tolerant traffic will suffer from increased latency and possible congestion. To avoid this situation, non-delay tolerant traffic should be prioritized and delay tolerant traffic should only be sent if additional network bandwidth is available. Alternatively, a fair queuing mechanism could be implemented to process delay tolerant and non-delay tolerant traffic based on expiration time of delay tolerant traffic.

Robustness

The *DTN application* and *DTN proxy* send only delay tolerant traffic to the special DTN boxes. All the non-delay tolerant traffic does not go through any DTN proxies. As a result, these solutions are robust and the failure of any DTN boxes does not affect any non-delay tolerant traffic. Only the delay tolerant traffic would be affected by a DTN proxy failure. *DTN interface* and *DTN*

While DTN Applications and DTN Tunnel remain unchanged, DTN Interface and DTN Proxy solutions require the placement of an intermediary box for each security level on the red side. The only safe way to segregate different security levels is to have a separate proxy for each level.

tunnel, on the other hand, use a transparent proxy. The transparent proxy only processes traffic that is delay tolerant. The other traffic is just allowed to pass through. Since even the non-delay tolerant traffic's tags are checked by the transparent proxy, a slight processing delay might be introduced via these solutions, even for applications that do not use DTN. Yet another concern for these solutions is the possible failure of the transparent proxy. If the transparent proxy device fails, it not only takes down delay tolerant traffic but also non-delay tolerant traffic, since all traffic passes through the transparent proxy device. A failover mechanism for non-delay tolerant traffic is therefore necessary to avoid complete communication failure over the affected portion of the military network for these solutions.

Conclusions

Disruptions to high capacity airborne links are frequently caused by factors such as aircraft body blockage, which exhibit unique characteristics compared to ground or satellite networks. These link outages might last seconds to minutes. To mitigate the effects of periodic link outages, DTN technology has been proposed. In this article we examined the use of DTN Bundle Protocol and applied it to an aerial HCB network that carries traffic from ships/subs to shore command posts. Specifically, we examined applying DTN proxies, tunnels, and interfaces to both the plaintext and ciphertext side of military networks to understand architecture and design considerations and limitations.

Future Work

There are a number of unsolved DTN challenges that could impact benefits achieved by proposed DTN solutions. First, the new Streamlined Bundle Security Protocol (SBSP) could improve options for black side storage of bundles within the network, but integrating these protocols with the existing military security architecture needs to be examined. Second, converting IP packets to bundles for DTN support incurs an additional processing penalty on all packets which affects performance for certain applications. Even though DTN traffic by definition is delay tolerant, each packet still has an expiration time associated with it based on how much delay the corresponding applications can tolerate. In the presence of network outages, the DTN proxies help improve application performance via the store-and-forward mechanism. However, when there are no outages, traffic should be sent as efficiently and quickly as possible (i.e. performance in terms of delay should strive to reach that of a non-DTN application). Investigating storage mechanisms (file system vs. memory) and selecting an appropriate convergence layer between DTN hops would help achieve this goal. Third, the use of DTN solutions for larger operational scenarios needs to be further investigated, e.g. the impact that multi-level security would have on the presented DTN deployment solution options. While DTN applications and DTN tunnel remain unchanged, DTN interface and DTN proxy solutions require the placement of an intermediary box for each security level on the red side. The only safe way to segregate different security levels is to have a separate proxy for each level. Fourth, in this article we assume a fixed DTN box for each DTN proxy/router. Using concepts such as software defined networking to deploy DTN proxies and the impact this has on overall robustness of the solution remains a topic of future work.

References

[1] B. Cheng *et al.*, "Characterizing Routing with Radio-to-Router Information in a Heterogeneous Airborne Network," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, Aug. 2013.

[2] B. Cheng *et al.*, "Evaluation of a Multi-hop Airborne IP Backbone with Heterogeneous Radio Technologies," *IEEE Trans. Mobile Computing*, vol. 13, no. 2, Feb. 2014.

[3] K. Scott *et al.*, "Robust Communications for Disconnected, Intermittent, Low-Bandwidth (DIL) Environment," *IEEE MILCOM*, Nov. 2011.

[4] J. Green and J. Schultz, "Collaborative Applications at the Tactical Edge Through Resilient Group Dissemination in DTN," *IEEE MILCOM*, Oct. 2012.

[5] K. Scott, "Disruption Tolerant Networking Proxies for On-the-Move Tactical Networks," *IEEE MILCOM*, Oct. 2005.

[6] R. Metzger and M. C. Chuah, " Opportunistic Information Distribution in Challenged Networks," *ACM Wksp. Challenged Networks*, Sept. 2008.

[7] S. Parikh and R. Durst, " Disruption Tolerant Networking for Marine Corps CONDOR," *IEEE MILCOM*, Oct. 2005.

Biographies

Rahul Amin (rahul.amin@ll.mit.edu) is a member of technical staff in the Tactical Networks Group at MIT Lincoln Laboratory. His research interests include design, analysis, and prototyping of end-to-end network solutions for ad hoc and challenged wireless networks. His recent research has focused on novel network architectures, scheduling and routing algorithms for both ad hoc and infrastructure based wireless networks. He received the B.S. (summa cum laude), M.S., and Ph.D. degrees in computer engineering from Clemson University. As a graduate student he worked on the topics of vehicular broadband networks with the BMW IT Research Center, and next generation heterogeneous cellular wireless networks and smart grid communications at the Networking Lab at the School of Computing at Clemson University.

David Ripplinger (david.ripplinger@ll.mit.edu) is an associate staff member in the Tactical Networks Group at MIT Lincoln Laboratory. His research interests include theoretical optimization, design, simulation, and testing of new protocols for wireless networks, with an emphasis on medium access control (MAC) design. His work has included characterization of the behaviors of random access and scheduling MACs in a frequency hopping environment; the design of low-overhead, distributed random access protocols; the stability of systems with various levels of zero-backoff (high-priority) traffic; and the design of fast, high-fidelity interference models in a frequency hopping environment. He received an M.S. in computer science and a B.S. in physics and Spanish translation from Brigham Young University.

Devanshu Mehta (mehta@ll.mit.edu) is an associate staff member in the Tactical Networks Group at MIT Lincoln Laboratory. His research involves developing network architecture and prototypes for mobile and disadvantaged networks. He has an MS from Worcester Polytechnic Institute, where his research focused on the graceful degradation of applications over unreliable wireless networks.

Bow-Nan Cheng (bcheng@ll.mit.edu) is a member of technical staff in the Tactical Networks Group at MIT Lincoln Laboratory. His research interests include the design, development, prototyping, and test and evaluation of next generation routing and information disseminations solutions for airborne backbone and tactical networks. He received M.S. and Ph.D. degrees in computer systems engineering from Rensselaer Polytechnic Institute, and he holds a B.S. degree in electrical engineering from the University of Illinois at Urbana-Champaign.

Exploring Value-of-Information-Based Approaches to Support Effective Communications in Tactical Networks

Niranjan Suri, Giacomo Benincasa, Rita Lenzi, Mauro Tortonesi, Cesare Stefanelli, and Laurel Sadler

ABSTRACT

Tactical networking environments present many challenges in terms of bandwidth, latency, reliability, stability, and connectivity. Sensors can today generate very large data sets that exceed the ability of tactical networks to transfer and disseminate them in a timely manner. Furthermore, the desire to cover larger areas with persistent sensing capabilities, have resulted in the widescale deployment of inexpensive sensors, further widening the gap between the volume of information that is generated and the subset that can successfully be delivered to consumers. This article explores the notion of determining the value of information in order to prioritize and filter information that is disseminated over these tactical networks, focusing on the dissemination of information to and from dismounted soldiers in a battlefield environment. This is a promising approach to mitigate the constraints of tactical networks and to reduce information overload on soldiers.

INTRODUCTION

Tactical networking environments present many challenges from the communications perspective in terms of bandwidth, latency, reliability, stability, and connectivity [1]. While sensing, computation, and storage capabilities have advanced rapidly, communication capabilities in tactical edge networks have not been able to achieve a similar growth rate. Sensors increasingly generate very large data sets that exceed the ability of tactical networks to transfer and disseminate them in a timely manner. Furthermore, a rapid decrease in the cost of sensors, combined with the desire to cover larger areas with persistent sensing capabilities, have resulted in widescale deployment of sensors, further widening the gap between the volume of information that is generated and the subset of that information which can be successfully delivered to consumers. Soldiers as sensors, equipped with smartphones or other portable computing devices, are placing a further load on the already congested networks.

These trends have motivated researchers to increasingly focus on the challenging problem of filtering information, and prioritizing and transmitting only those subsets that would be useful to consumers. In fact, recent research in multiple disciplines has raised the question of determining the value of information (VoI) as an enabler for effective decision making [2], thus enabling the filtering and prioritization of information according to the corresponding value perceived by the consumer on an individual basis [3].

Solutions that can analyze information and infer its value represent a natural complement to tactical communications middleware. In fact, the latter were designed to withstand node mobility and communication path disruptions, and to exploit the scarce communication resources in the most efficient way, typically by implementing smart and reliable message prioritization mechanisms [1] and data fusion [4]. VoI-based solutions help by further reducing the bandwidth requirements and improving the communication latency, essentially trading off the delivery of non-critical information to ensure that important and high-priority information can reach consumers that need it in a timely manner.

An equally important motivation for filtering information based on value to the consumer is to reduce information overload. Delivering and presenting unnecessary information to soldiers actively performing a task at the very least results in an unnecessary increase in their cognitive workload. In the worst case, it could become a distraction and cause them to make mistakes.

This article provides a working definition of VoI and a short survey of how other researchers and systems have applied this concept. Then the article explores VoI-based concepts for the purpose of timely dissemination of essential information to and from dismounted soldiers in a battlefield environment. The approach we describe is generalizable to the dissemination of information to other platforms and vehicles that are also interconnected via tactical networks, and is an extension of an earlier realization of information selection and prioritization based on relevance, which we described in [5].

Niranjan Suri is with the Florida Institute for Human and Machine Cognition (IHMC) and the U.S. Army Research Laboratory.

Giacomo Benincasa and Rita Lenzi are with IHMC.

Mauro Tortonesi and Cesare Stefanelli are with the University of Ferrara.

Laurel Sadler is with the U.S. Army Research Laboratory.

 0163-6804/15/$25.00

Any reduction in bandwidth utilization alleviates the constraints of TENs and is beneficial. Furthermore, the sorting and delivery of IOs in priority order reduces the latency of delivery of important information, which is an added benefit.

Tactical Edge Networks

The reader is likely to be familiar with the characteristics of tactical edge networks (TENs), but we summarize them in this section for completeness. Many types of nodes typically operate in the tactical environment. Some are mobile, such as manned and unmanned ground and air vehicles and portable devices carried by dismounted soldiers. Other nodes are stationary, such as tactical operation centers and unattended sensors (often grouped into wireless sensor networks to perform coordinated information gathering and object tracking tasks). Most of the nodes communicate through wireless links of various types (satellite, cellular, and ad hoc), usually in a hostile RF environment. As a result, in the tactical environment, severely constrained bandwidth, highly varying communication latencies, disconnected nodes, and network partitions are more the norm than the exception.

From the information-centric perspective, the objective of communications middleware operating in TENs is to manage discretized units of information content (henceforth referred to as information objects or IOs) and deliver them to consumers in the most effective way. IOs may be as simple as an assertion of some facts, location data of a friendly or enemy unit, a graphic such as a map or a picture, a document such as an intelligence report, or a full motion video clip. In a military context, information may be deemed to be of value if it increases the situational awareness (SA) of consumers and/or causes them to alter their course of action for a better outcome. The overall objective is to convey useful information to support decision making while reducing bandwidth consumption, delivery latency, and cognitive workload. Any reduction in bandwidth utilization alleviates the constraints of TENs and is beneficial. Furthermore, the sorting and delivery of IOs in priority order reduces the latency of delivery of important information, which is an added benefit.

Defining Value of Information

Research in sensor networks, where strict constraints on computation, energy, and channel access make communications particularly expensive, has recently identified two interesting metrics for ranking IOs with filtering and prioritization purposes: quality of information (QoI) and VoI [3, 6]. In this article, we generalize these metrics to other IOs, such as tracks and documents, in addition to sensor reports.

More specifically, QoI represents an "*internal*" *and objective metric* that considers the *intrinsic characteristics* of an IO. For example, an IO containing a photograph or an image might have a QoI value defined by aspects including level of detail (or resolution), clarity, contrast, exposure, and so on. In some cases, it is possible to devise automated systems that determine the QoI of an IO by analyzing its contents. For example, IOs generated by infrared sensors will typically have higher QoIs than those generated by visual sensors at night, and vice versa by day. In more complex cases, such as IOs containing documents/reports, the determination of the QoI may have to rely on a human operator.

On the other hand, VoI represents an "*external*" *and subjective metric* that classifies an IO according to the *utility it provides to its consumer* — that is, to its ability to support the consumer in more effective decision making. In fact, the VoI concept originated in economic and decision making research communities, with the purpose of investigating the advantages that additional information provided to decision makers [7, 8]. VoIs are dynamic values that change according to many factors, such as a consumer's needs and information availability. Also note that the same IO may have different VoI values for different consumers and that an IO may have a very high QoI (e.g., it may be a very accurate recent location report for an enemy unit) but a very low VoI for a particular consumer, for whom it represents irrelevant information. In other words, an IO's VoI is a function of its QoI, of its suitability or applicability to the consumer given the consumer's current context, as well as of the previous history of IOs sent to the consumer.

While QoI and VoI are novel concepts, researchers have already started investigating their adoption in mostly static/steady-state operational conditions, through the application of either multiple-criteria decision making solutions such as the analytic hierarchy process [3] or Von Neumann-Morgenstern utility functions [9]. These earlier works focus on static VoI values and essentially adopt the congestion control objective. More sophisticated solutions have applied the VoI concept to optimize the scheduling of message transmissions [10] or the traveling path of unmanned data harvesters [11] in underwater wireless sensor networks, considering time-varying but system-wide (i.e., non-consumer-specific) VoI measures.

However, the highly varying characteristics of TENs, as well as the ever changing contexts and mission objectives of all the personnel and devices operating in those environments, call for significantly more intelligent middleware solutions. In particular, they must be capable of considering multiple and dynamic VoI values for each IO, according to the corresponding consumers and their current contexts, and of integrating with tactical communication solutions [1] to implement dynamic IO filtering and prioritization policies. Since the contextual information about consumers and their relative interests might be incomplete or out of date, the reasoning component of the middleware should be capable of dealing with missing data and uncertainty, and of predicting personnel's future context and mission objectives — and consequently of forecasting the future VoI of a given IO for all the potentially interested consumers. Finally, the middleware should allow VoI determination to dynamically adapt based on feedback from the consumers in the field.

Calculating the actual VoI of an IO for a particular consumer is challenging, as it requires the system to model each consumer in terms of their existing knowledge, their objectives, their information needs, and their decision making strategy. A general solution that comprehensively addresses this problem does not exist to the best of our knowledge. This article describes a specific implementation, DSPro, which realizes VoI-

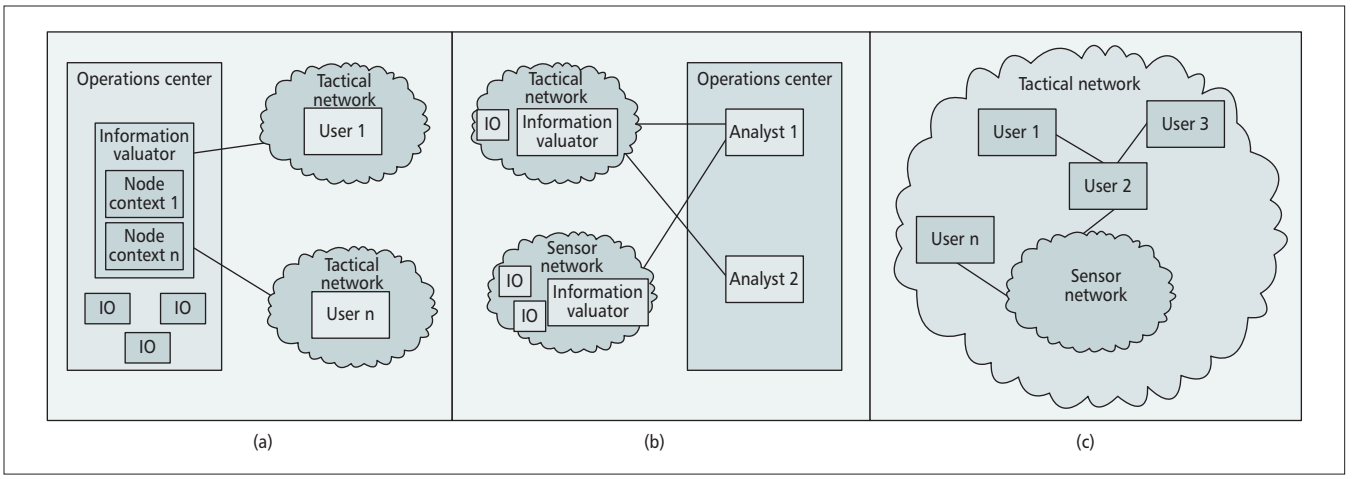


Figure 1. Deployment patterns for information value driven dissemination: a) from OC to edge network; b) from edge network to OC; c) at the edge network.

based filtering for a small set of critically important IO types in tactical environments, such as tracks, sensor reports, and other documents with metadata that supports such evaluation.

Value-Based Information Dissemination Patterns

As part of our ongoing research efforts in the application of VoI concepts within tactical environments, we have identified three different deployment patterns for exploiting information value-based dissemination, which are shown in Fig. 1. The first deployment pattern applies to information from a command/operations center flowing to dismounted soldiers using a tactical network. The intuition here is that the operations center is essentially an enterprise network node, with very few constraints on computation, storage, and network resources. However, not all of the information available at the operations center can be transmitted to the tactical edge given the capacity limitations. Therefore, each of the dismounted soldiers pushes their current context (denoted as the user context) to the operations center where an information valuator component examines the available IOs (and new incoming IOs, e.g., from deployed sensor networks) to determine their value, filter IOs that do not satisfy a relevance threshold, prioritize the IOs that are selected, and transmit those to dismounted soldiers. In this deployment pattern, data from sensor networks is first exfiltrated back to the operations center from one or more sensor network gateways (via a number of possible network links) before being evaluated and transmitted to dismounted users.

The second deployment pattern applies to information from the tactical network, including dismounted soldiers and sensor networks, being transmitted back to an operations center, where the consumer may be an analyst, or other tactical network users who are also connected to the operations center. In this case, the IOs are generated by sensors as well as soldiers, and include tracks, detections, pictures, reports, and other potentially large objects. As discussed before, the tactical network does not have the capacity to transfer all of this data back to the operations center. Therefore, in this second deployment pattern, IOs increasingly stay where they are gathered, and are pushed out of the tactical network to the operations center based on demand. For example, an analyst may express his or her interest in different types of IOs, which would represent his/her user context. These user contexts would be pushed out to sensor network gateway nodes, where valuator components would match locally generated IOs to the consumers. Again, IOs that satisfy a relevance threshold would be selected, prioritized based on their relevance, and transmitted to the analysts.

The third deployment pattern applies to information sharing directly at the tactical edge, for example, from one soldier to another or from a sensor network to a soldier. Unlike the first two cases, these peer-to-peer exchanges occur on an ad hoc basis based on potentially opportunistic contacts between the edge users. Hence, the user contexts are not pre-shared but exchanged upon contact, at which point a node with relevant IOs would push those to the other node. While not shown in the diagram, user nodes and sensor network gateway nodes contain information valuators as well as the node contexts for connected peers/users (we use the word peer and user interchangeably: a user is represented by a node that is a peer on the network).

DSPro Middleware

The DSPro middleware is a concrete realization of a VoI-based dissemination system that currently supports deployment patterns 1a and 1c (it is being extended to handle pattern 1b as well). DSPro implements a peer-to-peer architecture, where the same node can act as a consumer and provider of information to other nodes, and builds on top of the experience we developed with information dissemination in tactical environments in the context of the DisService research project [12].

Figure 2 shows the high-level architecture of DSPro, and sketches the key components in the

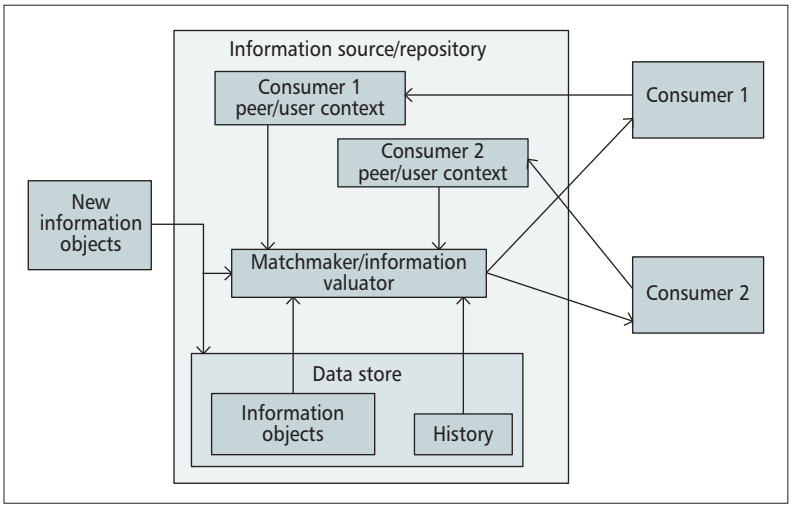


Figure 2. High-level component architecture of the DSPro middleware.

system and the interactions with information producers and consumers. In this middleware architecture, consumer nodes generate and push their context to other nodes that are potential sources or repositories of information. The source nodes have a matchmaker component that evaluates the value of each IO against a consumer node's context. IOs that have a value higher than a chosen or predetermined relevance threshold are then transmitted to the appropriate consumer. IOs that have been sent to a consumer are tracked in a history store so that they will not be considered again in the future. The matchmaker is triggered by one of two events, either the arrival of new IOs or changes to a consumer node's context. When a new IO arrives, it is stored in the local data store and evaluated by the matchmaker for all known consumers. On the other hand, if a peer node context changes, all IOs in the data store that were not already transmitted are examined to see if any of them would be of value to the consumer given the updated node context. If so, those are transmitted to the consumer and tracked in the history store.

DSPro adopts a pragmatic, computationally efficient approach to determining the VoI of an IO for each user/consumer. The evaluation function depends on the data type of the IO. The user node context representation in DSPro consists of a user's current location, a projected route (or routes), along with any temporal information (usually the result of a route planning tool), type of mission, role of the consumer, current activity, as well as policies about the value to be assigned to an IO based on the metadata. Given that a vast majority of tactical information is geographically driven, extra flexibility is provided in controlling the valuation based on geographical proximity. In particular, the user context can specify the notion of a "useful distance" for different IO types, which is taken into account when evaluating the VoI based on geographical proximity. Note that the evaluation is more than a simple geographical filter. In particular, the evaluation considers future planned positions (based on planned routes) in addition to the current position. Therefore, DSPro will match IOs that may not currently be relevant to a consumer based on the consumer's current position, but might be relevant where the consumer is expected to be in the future, with the value being inversely proportional to the expected time at that position.

Another enhancement to DSPro's evaluation of the value of information is to consider the previous history of information that has been matched and delivered to a consumer. For example, with tracks (position updates about entities), DSPro will compute the degree of change of the track from the last update sent to a consumer as part of determining the value of that track. This also takes into account the distance of the track from the consumer. Thus, a track that moves 100 m but is 1 km away is assigned a higher value than a track that moves 100 m but is 10 km away.

Proximity-based valuation also considers the range of influence of the entity represented by the IO. For example, an airborne platform, given its speed, has a much wider range of influence than a dismounted soldier walking on the ground. The range of influence policies in DSPro are specified based on the MIL-STD-2525 symbol code for the entity, which is part of the metadata of the IO. Using the MIL-STD-2525 symbol code makes changing the valuation policies in DSPro very flexible. For example, it is simple to express a policy in DSPro that assigns a high value to airborne elements vs. ground elements. Future implementations will also take into account the lines of bearing/direction of motion (moving toward a consumer vs. moving away) when computing the value of that particular track for a given consumer.

The user context also specifies the weights for a ranking function, which can adjust the relative importance across the different parameters of geographical proximity, temporal proximity, mission relevance, role relevance, and activity relevance. DSPro also supports custom policies for evaluating VoI on other metadata attributes of IOs. For example, Fig. 3 shows a simple policy that computes value based on the Affiliation attribute in the metadata. This particular example is a static policy (i.e., does not change based on other information in the user context) with a weight of 4.0 (out of a maximum of 10.0). If the attribute matches the word Coalition, the value is determined to be 5.0. On the other hand, if the attribute matches the word Hostile, the value is determined to be 9.0.

Determining the VoI of an IO for a user then consists of evaluating the metadata of the IO against that user node context. In addition to the above parameters, other attributes of the IO, such as the source (e.g., the commander of the mission), pedigree, and designated importance level are taken into consideration. As mentioned earlier, once the VoI is calculated for an IO, if the VoI falls below a configured "minimum value/worth" threshold (specified as part of the consumer node context), the IO is not transmitted to the consumer. If the VoI is higher than the relevance threshold, the IOs are sorted in priority order, based on their VoI, and transmitted to the consumer accordingly. Additional details about the matchmaking process are described in [5].

The context that is pushed to other nodes can change dynamically over time to reflect changes in the nature of information that is desired (or would be of value) to the consumer. Changes can include updates to policies (e.g., the one shown in Fig. 3), adjustments to the weights assigned to the different factors of geographical and temporal proximity, range of relevance based on the MIL-STD-2525 symbol code, changes to planned routes, current position, and so on. This allows a consumer to specify and control the nature of information that is desired. DSPro has been integrated with tactical applications such as Android Tactical Assault Kit (ATAK)[1], which provide the user interface to the consumer. Due to space considerations, we do not address the user interface in this article.

Since the matchmaker may run on resource-constrained nodes, the computational cost should be considered. In DSPro, the computational complexity is $O(n)$, where n is the number of IOs in the data store. If a peer node context changes, all n IOs in the data store have to be evaluated against the updated peer node context. On the other hand, when a new IO arrives, it has to be evaluated against each peer (so if there are m peers, there are m evaluations). The computational cost is kept low by not considering interactions between IOs. For example, DSPro does not consider that sending a report from sensor 1 may make it unnecessary to send a different report from sensor 2 since both reports cover the same target.

IOs are also immutable, so an update to an existing IO is handled as a new IO through the system.

Furthermore, DSPro linearizes the IOs so that they can be handled in some sequential order. The simplest approach to linearization is to use a metric such as the creation time or arrival time of the IO at a node in the network (e.g., a node that has to process the IOs).

Because military operations are inherently group efforts, often multiple consumers have similar node contexts, and therefore have interest in receiving similar subsets of IOs. For instance, all the members of a team may be interested in receiving situational awareness data from the area of deployment. A naïve implementation of value-based information dissemination that transmits the selected IOs objects via unicast may even underperform a simpler IP multicast-based implementation that blindly transmits all the IOs to every consumer if the overlap among the matched subsets is large. In DSPro, while the IO selection is performed by matching a single consumer node context against a single IO metadata, the actual transmission of the IOs is performed by taking the network topology into account. The matchmaker component aggregates IOs that need to be sent to multiple consumers in the same subnetwork and then transmits a single copy of the message via multicast.

DSPro relies on a reliable multicast capability provided by DisService [12] (which, in turn, is based on hop-by-hop UDP multicast). Because IOs are multicast, nodes may receive irrelevant IOs, which are not processed but are locally cached for possible later use. If a previously cached IO becomes relevant in the future, the node is simplify notified that the IO is now relevant to the consumer without having to transmit the IO again, thereby saving on bandwidth. DisService goes even further with its implementation of opportunistic listening [13], where each network packet is self-contained and self-describing, which allows intermediate nodes that listen in on this traffic to be able to cache the packets, and potentially make them available to other peers at a later point in time. This approach increases the availability of the information in the face of disconnections and network partitioning.

It is important to note that false positives (where the information is not relevant but is not filtered) can be tolerated because while the system may not save as much bandwidth as possible, it would still be an improvement over the baseline. However, false negatives (where information is relevant but still filtered) are detrimental as they might cause the consumer to lose SA and/or make an incorrect decision.

```
<?xml
version="1.0"?>
<RankerPolicy>
    <Type>Static</Type>
    <Attribute>Affiliation</Attribute>
    <Weight>4.0</Weight>
    <Alternative>
        <Match>Coalition</Match>
        <Value>5.0</Value>
    </Alternative>
    <Alternative>
        <Match>US</Match>
        <Value>6.0</Value>
    </Alternative>
    <Alternative>
        <Match>Hostile</Match>
        <Value>9.0</Value>
    </Alternative>
</RankerPolicy>
```

Figure 3. Custom ranking policy in DSPro for metadata.

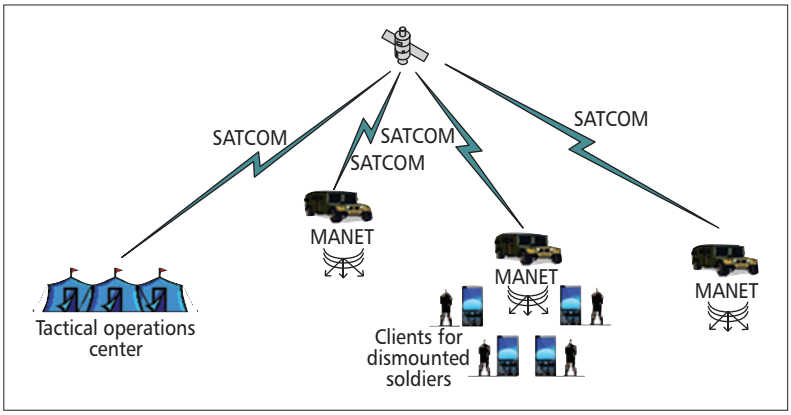


Figure 4. DSPro experiment scenario.

Experimental Results

We performed an experimental evaluation of DSPro using a reference scenario based on Agile Bloodhound — an annual U.S. Department of Defense (DoD) Office of Naval Research (ONR) sponsored technology demonstration, where

[1] https://atakmap.com/

	Selectivity	Tracks				Sensor reports				Bandwidth (kb/s)		Performance improvement	
		Total number	Delivered per client	Filtered out	Average output	Total number	Delivered per client	Filtered out	Average overlap	SAT-COM	MANET	SAT-COM	MANET
Baseline	N/A	3445	3445.0	0	100.0%	510	510.0	0	100.0%	642.1	936.2		
Naïve	High	3445	997.0	2448	100.0%	510	105.5	315	75.9%	361.9	466.8	43.64%	50.14%
DSPro		3445	997.0	2448	100.0%	510	107.8	315	76.9%	350.7	328.9	45.39%	64.86%
Naïve	Medium	3445	997.0	2448	100.0%	510	124.2	303	82.1%	369.4	515.7	42.47%	44.92%
DSPro		3445	997.0	2448	100.0%	510	125.0	303	82.6%	361.5	338.9	43.71%	63.81%
Naïve	Low	3445	997.0	2448	100.0%	510	173.5	223	80.8%	436.7	644.5	32.00%	31.16%
DSPro		3445	997.0	2448	100.0%	510	176.5	223	83.6%	469.7	481.2	26.85%	48.60%

Table 1. DSPro performance results.

DSPro has been deployed over the last four consecutive years. This scenario realizes the deployment pattern from Fig. 1a. As shown in Fig. 4, the scenario consists of multiple military hub vehicles connected via SATCOM to a command center. Each hub vehicle supports multiple dismounted soldiers, either on foot or in vehicles of their own, using a tablet/smartphone and communicating through a mobile ad hoc network (MANET). The scenario emulates a typical operation that involves a variety of information flows, including friendly (blue) force tracks, enemy (red) force tracks, sensor reports, documents such as intelligence reports and logistics reports, and messaging. For the purpose of this experiment, we consider only two data types: tracks and sensor reports. The metadata and data format for tracks and sensor reports were XML messages as defined by the Marine Air-Ground Task Force (MAGTF) Command and Control (C2) Tactical Service-Oriented Architecture (TSOA). However, it should be relatively easy to adapt the system to use other standardized metadata and data formats, particularly if they are XML-based.

Tracks were fed into DSPro at the operations center from the Joint Tactical Common Operating Picture Workstation (JTCW), but could come from any number of track management systems such as the Distributed Common Ground System (DCGS). Tracks are also generated by vehicles on the move and by dismounted soldiers at the edge. Note that there are a wide variety of tracks, representing platforms ranging from ships to airborne vehicles to ground units, that are part of multiple, ongoing, sometimes unrelated missions. Therefore, not all tracks are relevant to every consumer.

Sensor reports are generated from a variety of sensors deployed throughout the area of operations. A sensor report typically contains metadata and a payload, typically an image. Each sensor report affects or covers a geographic subregion. Again, not every sensor report is relevant to every consumer.

We compared three different information dissemination strategies: a baseline strategy, in which all tracks and sensor reports are delivered via DisService's reliable multicast communication function to each of the four clients; and the naïve and DSPro strategies, which are VoI-based approaches that independently evaluate and select the IOs to deliver to each consumer. The naïve approach to implementing such a system is to independently transmit the relevant IOs to each consumer. But this approach could be highly inefficient in a scenario where there is a large degree of overlap between the IOs that are selected for a set of consumers in physical proximity, since the overlapping IOs would be duplicated on the MANET network as they are transmitted independently to each consumer. As mentioned earlier, the DSPro implementation aggregates data and avoids multiple transmissions of the same IOs on the same network links, thereby saving bandwidth.

It is important to note that the results could vary widely based on the policy selected for the VoI-based filtering and prioritization, and on the actual scenario itself (in terms of the number of tracks, their positions, the number of sensor reports, their sizes, and their coverage). In fact, DSPro provides many customizable options for determining the VoI as well as setting the filtering relevance thresholds. As a result, the number of IOs transmitted by the three strategies to consumers are significantly different.

Table 1 shows the results we obtained for the subset of soldiers supported by one of the hub vehicles. In total, there were 3445 tracks and 510 sensor reports generated during the mission execution phase. Results for both the naïve and DSPro optimized implementations are shown for the three different selectivity thresholds (high, medium, and low) defined by setting the corresponding useful distance to 0.5 km, 1 km, and 2 km for sensor reports. In all cases, the geographic proximity filter for tracks was set to be 1000 km, a large value to ensure that all nodes have the same common operating picture regarding tracks in their area of operations. This setting resulted in tracks for unrelated missions being filtered out, but each of the four clients that are part of the same mission receive the same set of tracks (i.e., an overlap of 100 percent).

With regard to sensor data, the number of reports filtered out for each client varies according to selected dissemination strategy and, for the naïve and DSPro strategies, to the configured rel-

evance threshold. As can be seen in the results, applying a VoI-based filtering mechanism in this scenario is very effective, reducing SATCOM bandwidth utilization by between 26.85 and 45.39 percent, and MANET bandwidth utilization by between 48.60 and 64.86 percent. Note that there is a small difference between the naÏve and DSPro implementation with the same selectivity, with delivery count being higher with DSPro. This is caused by DSPro being more efficient, and being able to send more reports to clients before their position (and consequently their node context) changes. Once the node context changes, any unsent reports that are no longer relevant are simply dropped.

Conclusions

Value-of-information-based approaches to information management and dissemination are a particularly promising direction for tactical network communications. They are an effective mechanism to counter the increasing disparity between the volume of data gathered/generated and the bandwidth available in the network to move all of that data. VoI-based approaches also have the potential to reduce operator overload by filtering out unnecessary information that can be distracting. Determining the VoI in a generic open-ended system is a difficult unsolved problem. However, this article has described the DSPro middleware, which exploits VoI in a tactical information management context. DSPro has been applied to the problem of disseminating information from an operations center to dismounted soldiers, and also between soldiers and sensor networks at the edge. Initial results are promising in terms of bandwidth reduction. Future efforts will focus on more comprehensive representations of users' contextual information that is used to evaluate VoI, and on more flexible VoI evaluation mechanisms that can accommodate multiple deployment scenarios.

Acknowledgments

The authors would like to thank John Moniz at the Office of Naval Research for initial sponsorship under Grant N00014-09-1-0012, and the U.S. Army Research Laboratory for continued sponsorship under cooperative agreement W911NF-11-2-0095.

References

[1] N. Suri *et al.*, "Communications Middleware for Tactical Environments: Observations, Experiences, and Lessons Learned," *IEEE Commun. Mag.*, vol. 47, no. 10, Oct. 2009, pp. 56–63.

[2] R. Laxminarayan and M.K. Macauley, Eds., *The Value of Information: Methodological Frontiers and New Applications in Environment and Health*, Springer, 2012.

[3] C. Bisdikian, L. Kaplan, and M. Srivastava, "On the Quality and Value of Information in Sensor Networks," *ACM Trans. Sensor Networks*, vol. 9, no. 4, article 48, July 2013, pp. 48:1–48:26.

[4] H. Mitchell, *Data Fusion: Concepts and Ideas*, Springer, 2012.

[5] S. Rota *et al.*, "Supporting Information on Demand with the DisServicePro Proactive Peer-to-Peer Information Dissemination System," *Proc. MILCOM 2010*, San Jose, CA, 31 Oct.–3 Nov. 2010, pp. 984–91.

[6] I. Todoran *et al.*, "Assessing Information Quality in Fusion Systems," *NATO SAS-106 Symp. Analysis Support to Decision Making in Cyber Defence & Security*, Tallinn, Estonia, 9–10 June 2014, 2014.

[7] S. Galanis, "The Value of Information under Unawareness," *J. Economic Theory*, vol. 157, May 2015, pp. 384–96.

[8] J. Quiggin, "The Value of Information and the Value of Awareness," *Theory and Decision*, 2015.

[9] D. Cansever, "Value of Information," *Proc. MILCOM 2013*, San Diego, CA, 18–20 Nov. 2013, pp. 1105–08.

[10] L. Bölöni *et al.*, "Scheduling Data Transmissions of Underwater Sensor Nodes for Maximizing Value of Information," *Proc. IEEE GLOBECOM 2013*, 9–13 Dec. 2013, pp. 460–65.

[11] S. Basagni *et al.*, "Maximizing the Value of Sensed Information in Underwater Wireless Sensor Networks via an Autonomous Underwater Vehicle," *Proc. IEEE INFOCOM 2014*, Toronto, Canada, 27 Apr.–2 May, 2014.

[12] N. Suri *et al.*, "Peer-to-Peer Communications for Tactical Environments: Observations, Requirements, and Experiences," *IEEE Commun. Mag.*, vol. 48, no. 10, Oct. 2010, pp. 60–69.

[13] N. Suri and G. Benincasa, "Opportunistic Listening System and Method," U.S. Patent No. 8,493,902 issued July 23, 2013.

Biographies

Niranjan Suri (nsuri@ihmc.us, niranjan.suri.ctr@mail.mil) is a senior research scientist at the Florida Institute for Human & Machine Cognition (IHMC) and also a visiting scientist at the U.S. Army Research Laboratory, Adelphi, Maryland. He received his Ph.D. in computer science from Lancaster University, England, and his M.Sc. and B.Sc. in computer science from the University of West Florida, Pensacola. His current research activity is focused on the notion of agile computing, which supports the opportunistic discovery and exploitation of resources in highly dynamic networked environments. His other research interests include information management, coordination algorithms, distributed systems, networking, communication protocols, virtual machines, and software agents.

Giacomo Benincasa (gbenincasa@ihmc.us) is a research associate at IHMC. He holds an M.Sc. in computer engineering from the University of Modena and Reggio Emilia, Italy, an M.Sc. in computer engineering from the University of Florida, Gainesville, and a B.Sc. from the University of Modena and Reggio Emilia, Italy. His current research activity is focused on agile computing approaches to information dissemination that opportunistically discover and exploit excess communications, storage, and processing capacity in a distributed network. His other research interests include distributed systems, networking, communications protocols, opportunistic sensing, and software agents.

Rita Lenzi (rlenzi@ihmc.us) is a research associate at IHMC. She received both her B.Sc. and M.Sc. in computer science from the University of Modena and Reggio Emilia. Her research interests lie in the areas of information management, federation, and network visualization.

Mauro Tortonesi (mauro.tortonesi@unife.it) received his Laurea degree in electronic engineering and his Ph.D. in computer science engineering from the University of Ferrara, Italy. Currently, he is an assistant professor in the Engineering Department of the University of Ferrara, Italy. His research interests include distributed and mobile computing, Internet of Things, QoS management, network and service management, business-driven IT management, and e-maintenance.

Cesare Stefanelli (cesare.stefanelli@unife.it) is a professor of computer science engineering in the Engineering Department of the University of Ferrara. His research interests focus on distributed and mobile computing in wireless and ad hoc networks. He received his Laurea degree in electronic engineering and his Ph.D. in computer science Engineering from the University of Bologna, Italy.

Laurel Sadler (laurel.c.sadler.civ@mail.mil) received her B.Sc. in electrical engineering from the University of Maryland and her M.Sc. in electrical engineering from the Johns Hopkins University focusing on signal processing. She has over 21 years of experience as an electronics engineer at the U.S. Army Research Laboratory in Adelphi, Maryland. Her research interests are in robotics communications and control software, video surveillance systems, databases, and information dissemination.

Initial results are promising in terms of the bandwidth reduction. Future efforts will focus on more comprehensive representations of users' contextual information that is used to evaluate VoI, and on more flexible VoI evaluation mechanisms that can accommodate multiple deployment scenarios.

A Correspondence Model for a Future Military Messaging Handling System

Laurent Cailleux and Ahmed Bouabdallah

ABSTRACT

Messaging service still remains fundamental for military operations. This service is an important tool to fulfill missions, either in time of peace or in time of crisis. Military operational needs must comply with specific requirements. Military standards describe the services that meet these needs. For different reasons, current military messaging standards have become obsolete, thus demanding the definition of new military standards. The main requirements concerning such systems have been put together. The main document, "Future Military Messaging," issued in 2005, identifies the different usages of a future military messaging system with the associated requirements in each case. Even if this document has become the reference model in the military messaging area, new theoretical developments are required. Indeed, it quickly appears that the development of a messaging system compliant only with this model could lead to rigid implementation, preventing any evolution of the system. This article reformulates the messaging system through a new concept called correspondence applied to military messaging. The main interest of this model is to decorrelate the deployment of the policies from the implementation of the messaging handling system. Therefore, it offers new prospects for military messaging systems with the ability of providing different levels of services dedicated to various military environments.

INTRODUCTION

For many years, military systems were based on Allied Communications Publication 127 (ACP127),[1] which is the legacy protocol used for military formal messaging. ACP127 is based on telegraph format, which only permits the exchange of messages with a restricted set of characters. Besides, it does not provide attachment capabilities or enriched text.

With the evolution of messaging technologies, and in order to replace ACP127 systems, in 1987 the North Atlantic Treaty Organization (NATO) decided to create a Military Messaging Handling System Working Group (MMHS WG). The objective of this working group was to define a new standard for military messaging, and it provided a definition of MMHS, the purpose of which is to convey military messages between (NATO/military) organizations or individuals.

In 1999, a first version of the STANAG 4406 Edition 1 standard[2] was published. This standard was based on the Consultative Committee for International Telegraphy and Telephony (now the International Telecommunication Union, ITU) X.400 series of Recommendations.[3] The military context may involve the exchange of different types of information. In 2005, a new version of this standard, STANAG 4406 Edition 2,[4] introduced different levels of services (high grade, medium grade, and basic grade), each level associated with dedicated policies.

For their part, the Combined Communications Electronics Board (CCEB) have also defined ACP123 as an MMHS standard. To ensure interoperability with NATO members, this ACP was developed in close coordination with NATO's Core Enterprise Services Working Group (CESWG,WG/5). This ACP does not specifically cover messaging between individuals.

The changing military context and the evolution of technology and Internet standards imply the definition of a new generation of MMHS. Moreover, the obsolescence of standards (e.g., X.400 Recommendations) and the cost of systems based on these standards have made it necessary to define such a new generation. A study called Future Military Messaging (FMM)[5] has been issued with the objective to define all the requirements of a future military messaging system and to clarify the design of such a system. FMM document introduces three categories of requirements concerning information-oriented services (IOS), system-oriented services (SOF), and mixed functions (MF). IOS concerns end-to-end exchange of information. SOF focuses on the properties of the underlying transfer system. MF groups all the remaining requirements. The FMM document involves 48 mandatory (M) or optional (O) detailed requirements distributed between the three previous categories. Another important requirement concerning interoperability has recently been introduced for more general purposes in NATO Federated Mission Networking (FMN).[6]

The development of a military messaging system using traditional design approaches could lead to rigid and inefficient implementations, which will complicate the evolution of such a system. Indeed, traditional email systems provide a

Laurent Cailleux is with the French Ministry of Defence.

Ahmed Bouabdallah is with Institut Mines-Telecom/Telecom Bretagne.

[1] CCEB, "Communication Instructions — Tape Relay Procedures," ACP 127 (G), Nov. 1988.

[2] STANAG 4406, "Military Message Handling System, Edition 1," NATO, 1999.

[3] CCITT Eighth Plenary Assembly, "Recommendations X.400–X.420: Data Communication Networks Message Handling Systems," ITU, Nov. 1988, CCITT Blue Book Volume VIII.7.

[4] STANAG 4406, "Military Message Handling System, Edition 2," NATO, 2005.

[5] NATO Military Message Handling Working Group, "Requirements and Rationale for Future Military Messaging," 5 edition, Oct. 2003.

[6] http://www.act.nato.int/fmn

0163-6804/15/$25.00 © 2015 IEEE

set of features for all their users. In some cases, it is possible to set policies for specific users or groups of users, but the description of a dedicated policy is not standard and could be difficult to implement. For this reason, most email systems provide a single policy for all their users. This solution could be efficient but limited. For example, the security services are the same for all users, regardless of the type of exchange. Of course, the user can choose whether to apply a cryptographic signature or not, but it is the choice of the user. There is no relation between the type of exchange and applied policy. Email systems allow multiple usages but with a single policy. This is a major limitation of current email systems. Moreover, as seen with STANAG 4406, the capability of exchanging different kinds of information may be required. In traditional design approaches, each type of information may induce different kinds of services, which may necessitate the implementation of a dedicated system with associated policies; for example, the exchange of official information does not impose the same constraints as the transmission of critical information on the system.

In this article, we propose to take into account these aspects and add such capabilities to a traditional email system. We reformulate the messaging system through a new concept called correspondence applied to military messaging. A type of correspondence can be seen as a correspondence model on which a policy expressing various requirements is grafted. From this model, it is possible to describe different types of correspondence, thereby allowing the application of an appropriate policy to each grade of services. For instance, a user could have a basic grade of service, while another might have a high grade of service to exchange operational information.

This article is organized as follows. The following section introduces the main messaging systems and exposes our problems. In the next section, we present our main contributions in the form of a military correspondence model, policies, and types of correspondence and we conclude in the final section.

Messaging Systems

Email Systems

An email system or message handling system (MHS) is composed of at least the set of elementary services: Internet Message Format (IMF) [1], Simple Mail Transfer Protocol (SMTP) [2], and Internet Message Access Protocol (IMAP) [3] enabling the creation, sending, and receiving of emails. The body section of an email object is limited to a string of US-ASCII characters. Structuring the body with Multipurpose Internet Mail Extensions (MIME) [4] will enable us to solve this issue.

Implemented at the beginning of the 1980s, its ease of use combined with its efficiency gave this service an irreversibly growing success which significantly contributed to the development of the Internet. Even though significant developments have taken place, the main protocols, IMF and SMTP, have evolved in slight proportions. IMF defines the syntax for text messages sent between

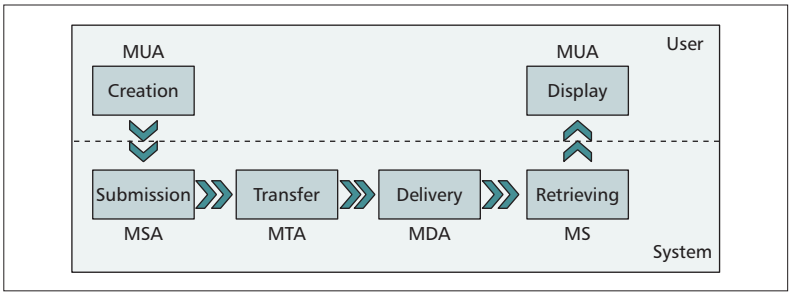


Figure 1. Components and events in an email system.

computers within the framework of electronic mail messages. SMTP has to ensure reliable and efficient transfer of email between two components of an email system and, more generally, between two end users. A global and comprehensive description of Internet Mail Architecture (IMA) is given in [5]. In terms of architecture, IMA describes six basic types of components:

- Message
- Message user agent (MUA), the component that works on behalf of user actors and provides access to the MHS
- Message submission agent (MSA), the component that accepts each email from the MUA and submits it to the MHS
- Message transfer agent (MTA), the component in charge of transfer of emails
- Message delivery agent (MDA), the component that delivers each email from the MHS to a recipient's mailbox
- Message store (MS), the component that provides long-term email storage

These components and the associated events are described in Fig. 1.

Military Messaging Handling System

As mentioned in the introduction, the military context requires the use of appropriate messaging services. These services must provide capabilities depending on the type of exchange. STANAG 4406 defines three grades of service (high, medium, and basic) for the use of messaging services, and each grade provides different capabilities.

High grade service: The high grade service (HGS) corresponds to the most constrained level of requirements. The use of an HGS-compliant system guarantees the exchange of critical information and official correspondence between users in a military operational environment. Critical information is defined as information where people's lives or military missions are jeopardized if a message is not delivered to the target recipient within the time defined by the priority of the message. Official correspondence is formal information exchanged between military organizations. HGS is also referred to as "fire and forget" service.

Medium grade service: The medium grade service (MGS) enables the exchange of important information between individuals. The use of an MGS-compliant system guarantees delivery and security. STANAG 4406 defines important information as that exchanged between individuals with a formal function in the military structure, which contributes to the preparation of orders.

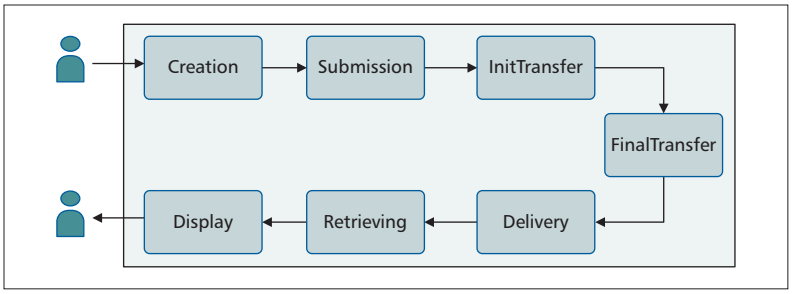


Figure 2. Email life cycle.

There are several differences between HGS and MGS. MGS is not intended for critical mission operations but provides a secure messaging service to allow exchanges linked with administrative activities. In addition, the user must ensure that message delivery is achieved. MGS is also referred to as "fire and watch" service.

Basic grade service: The basic grade service (BGS) is a simple military messaging service that only offers general email service. In BGS, there is no advanced service; for example, the implementation of security services provided by a public key infrastructure (PKI) is not required. BGS allows the exchange of routine information between persons throughout Defense and their partners. BGS is also referred to as "fire and hope" service.

MMHS over SMTP

Due to the obsolescence of standards such as the X.400 Recommendations and the cost of systems based on these standards, there is a need to use new standards. Nowadays, the obvious solution is to specify a new military messaging system based on Internet Engineering Task Force (IETF) standards, particularly based on three main protocols: IMF, SMTP, and IMAP.

A first study[7] suggests the applicability of Internet email standards for military high-grade messaging. The objective of the study is to compare the MMHS requirements and the functionalities of Internet email standards. Some requirements (priorities, deferred delivery, alternate recipient, message header fields, etc.) are not met by the transport and format protocols.

Another study[8] describing some drafts[9,10] and new standards [8–10] has completed this work.

Problems

To take into account particular threats and security objectives, trusted email systems have been defined. These systems meet specific requirements of an organization, but they have several important limitations.

These systems require the implementation of many features, including security services encapsulated in a single monolithic policy, which is the same for all its users. These solutions could be efficient but limited because there is no relation between the nature of the exchange and the applied policy. A user may indeed have various and different usages of email systems: private, professional, administrative, official, military, and so on. But for a given usage, a user should ideally be able to select the policy that encapsulates the exact set of associated features. The main solution to solve this issue would be to define a system that encapsulates security services required by the most restrictive usage. But although this approach seems to be a good solution, some usages may have contradictory requirements. For example, a system that implements an archiving mechanism of all the outgoing messages might not be in compliance with the requirements of privacy. Thus, the encapsulation of security services is not the solution.

Furthermore, email systems can evolve. For example, the appearance of a new threat can lead to the definition of new security services and specification of a new policy. Thus, email systems must be able to evolve in a fast way and according to the new requirements.

Correspondence Model

To address the different limitations presented in the previous paragraph, we introduce a general concept of correspondence based on interpersonal communication. Basically, an interpersonal communication is an exchange between two or more people. During the exchange, there is message sending and message receiving. We propose to apply the concepts of interpersonal communication in an email system. In this system, there are one sender, one or more recipients, a message, and an MHS to transfer it.

The Email Life Cycle

In order to reach its recipient, a message goes through an email system and is handled by different components of the system. To describe this concept, we use the term email life cycle (ELC). ELC is a sequence of ordered events. It begins with the event of Creation and terminates with the event of Display. An ELC is complete when the Display event is processed successfully (the message is displayed by the recipient). Otherwise, the ELC is incomplete. A full intradomain email life cycle is illustrated in Fig. 2.

The ELC is composed of the following events:

- Creation
- Transmission
  - –Submission
  - –Relay
    - -InitTransfer
    - -Transfer
    - -FinalTransfer
  - –Delivery
  - –Retrieving
- Display

An ELC could be composed of several events of *Transfer*. For example, the message could be sent to another administrative management domain (ADMD). In this document we assume that the email is exchanged in the same ADMD and with only one MTA.

Below, we propose an example of the concept of complete and incomplete ELC. It is illustrated in Fig. 3. Alice (A) sends a message to Bob (B), Charlie (C), and Dave (D). User B retrieves the message and displays it. User C does not retrieve the message, and user D is unknown. This exchange is composed of three sequences of ELC. Events are shown as follows:

- *c*: creation
- *s*: submission

[7] M. Schmeing and N. Haak, "Applicability of Internet-Email for Military High-Grade Messaging," report FGAN/FKIE, Feb. 2005

[8] http://www.isode.com/whitepapers/mmhs-over-smtp.html.

[9] A. Melnikov, G. Lunt and A. Ross, "Military Message Handling System (MMHS) over SMTP," Internet draft, expired on 2 Jan. 2015; July 2014.

[10] A. Melnikov, "Draft and Release Using Internet Email ," Internet draft, expired on 25 Apr. 2015; Oct. 2014.

- *i*: initial transfer
- *f*: final transfer
- *d*: delivery
- *r*: retrieving
- *p*: display

Sequence 1, denoted ELC(c1, p1), is complete because user B displays (with p1 event) the message. Sequence 2 ELC(c1, d1′) is incomplete because user C does not display the message. Sequence 3 ELC(c1, d1″) is incomplete because user D does not exist.

Determination of Intention of Communication

When a user sends a message to a recipient, it performs an act of interpersonal communication that can take different forms (private, professional, etc.). This act is preceded by an intention of communication. J. Searle [6] points out that an act of communication is not only the recognition of words in a given language but "grasping the gist of what is being conveyed," and defines intention of communication as intention "to make the listener recognize what I want to say, that is, to understand me." The communication is successful when the recipient understands the intention of the sender [7].

The application of these concepts in the domain of email systems will allow applying specific policies according to the user's intention of communication. The challenge is to describe the email system that is able to take into account the sender's intention of communication. Therefore, we define the correspondence model concept.

Concepts of the Correspondence Model

We define a correspondence as an abstract model that formalizes one or more sequences of the ELC and together the applied policies representing the user intention of communication (UIC).

A correspondence is composed of one sequence of ELC when there is only one recipient. The number of sequences depends on the number of recipients. We define an elementary correspondence (EC) as a one-sequence of ELC. An EC is complete when the ELC is complete; otherwise, the EC is incomplete. A correspondence is complete when all its associated ECs are complete; if not, the correspondence is incomplete.

The first message, referred to as root message, could initiate a thread of message. In this case, all the correspondences are gathered within a global correspondence. The global correspondence is composed of all the correspondences originating in the same root message. We can find notifications like delivery status notification (DSN)[11] or message disposition notification (MDN).[12]

We define a correspondence as a model allowing us to define and enforce policies specific to each usage of email sending and taking into account the current context of the sender.

Figure 4 depicts one way to graft the correspondence model on the email architecture. For each step of email routing, each component should be solicited to enforce the local part of the correspondence policy according to the usage and context of the email.

We propose to define a successful EC when it is complete and when all applied policies according to the usage and the current context are satisfied.

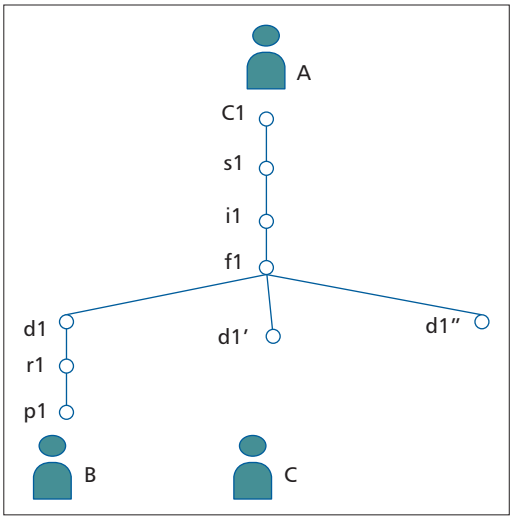


Figure 3. Example of correspondence.

Concerning the message, it is a transmitted object from a sender to a recipient. IMF does not contain explicit information related to the used correspondence. In order to explicitly provide this kind of information, we have extended IMF with the definition of eXtended Internet Message Format (XIMF) technology.[13] A prototype implementing the SMTP transmission of the correspondence information has been developed and is currently being tested.

Correspondence Policies

With our model, it is possible to define an advanced email system that applies appropriate policies according to usage. The correspondence policy concept is composed of distinct parts: policy model, policy discovery, and policy enforcement.

The policy model aims to provide a framework allowing specifying rules and constraints according to the usage and context of current correspondence. When the message is transmitted through the system, each agent of the system must apply rules present in the correspondence policy. To that end, the agent ought to know the part of the policy that has to be applied. Two options exist. Either the policy is installed in the agent, or the agent has to retrieve it. For scalability reasons, we suggest the second option.

Each policy is identified with a correspondence policy reference and is stored in a policies server. A policy reference is composed of a name, a version, and a name domain. This reference must be unique. The name of the policies server can be obtained from the DNS. The agent sends an SRV record request to the DNS and obtains in response the name of the policies server. Then the agent can retrieve the correspondence policy from the policies server. In our system, each ADMD can have a policy server. This allows defining policies and deploying them easily on the basis of HTTP and DNS services. But it is also possible to define a policy used by several ADMDs and to centralize it in a central policies server. In this case, all organizations agree to implement the policy.

With our model, it is possible to define an advanced email system that applies appropriate policies according to the usages. The concept of correspondence policy is composed of several distinct parts: policy model, policy discovery and policy enforcement.

[11] K. Moore and G. Vaudreuil, "An Extensible Message Format for Delivery Status Notifications," RFC 3464, IETF, January 2003.

[12] T. Hansen and G. Vaudreuil, "Message Disposition Notification," RFC 3798, IETF, May 2004.

[13] http:// www.trusted-bird.org

Our model allows, for each end user's usage, the precise definition of the associated correspondence. We can define an advanced military messaging system that allows implementation of multiple policies adapted to multiple usages, such as high, medium, and basic grade services.

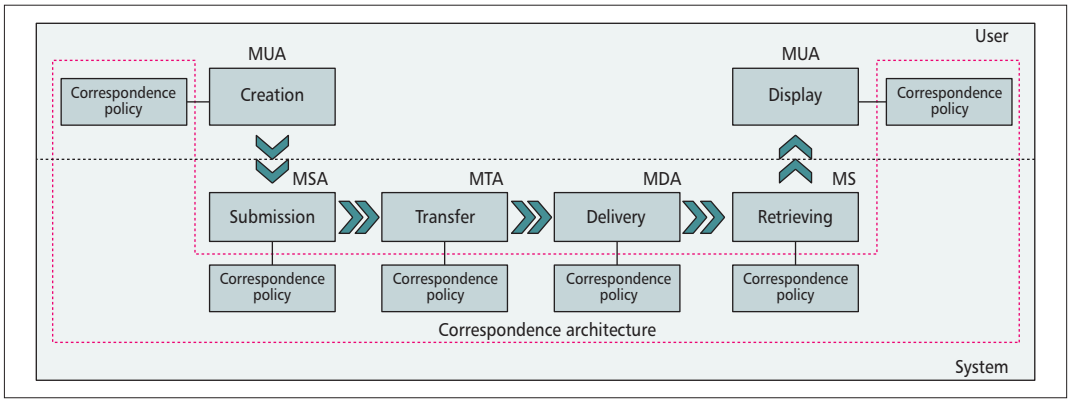


Figure 4. Correspondence architecture.

In order to apply correspondence policies, each agent has to integrate a dedicated component. In our work, we have defined a correspondence enforcement agent (CEA) component. This component is in charge of enforcing the correspondence policies. In an ideal system, each agent should integrate one CEA. However, in some cases, email systems could implement a CEA only within the MUA. CEA applies the correspondence policy according to the event. For example, the CEA within the MSA applies the policy of a submission event, and the CEA within the MDA applies the policy of a delivery event. Hence, with this policy model, it is possible to apply an appropriate policy according to the user intention of communication.

Use Cases

We can define a type of correspondence as a correspondence model instantiation. A type of correspondence could be associated with only one ADMD or with several ADMDs. We illustrate our approach through a comprehensive example covering both the intra- and inter-ADMD cases.

Let us take three users, Alice, Bob, and Dave: Alice and Bob are in ADMD1 and Dave in ADMD2. Alice is responsible for the J6 division (communication and information systems) for nation X. Dave is responsible for the J4 division (logistics) for nation Y. To organize exchanges between the two ADMDs, we assume that prior negotiation between them led to the definition of a common correspondence policy. The latter describes the header fields to be integrated into a military message in accordance with RFC 6477 [10] together with the security profile. This security profile requires that a military message sent between such organizations should be signed by the sender and countersigned by the associated organization.

Alice, Bob, and Dave use MUA A, B, and D, respectively, and there are two email servers, S1 and S2, one for each ADMD. The three scenarios below express the capacity of our approach to homogeneously handle three very different cases:

- Scenario 1: Alice wants to send an important and sensitive professional message M1 with her email client to Bob.
- Scenario 2: Alice wants to send a private message M′1 to Bob.
- Scenario 3: Alice, responsible for J6, sends a military message M″1 to Dave, responsible for J4.

Each one of the previous exchanges is precisely managed by a correspondence type deployed and implemented by a set of three policies, {P1, P2, P3},{P′1, P′2, P′3} and {P″1, P″2, P″3}, respectively.

Figure 5 depicts this email exchange. We detail the main steps of the three scenarios below:

1. A retrieves the right correspondence policy (P1, P′1 or P″1 depending on the scenario).
2. Alice opens her creation form, fills out the displayed message form, selects information about the current correspondence, and selects send action in the form.
3. A generates an M1 (resp. M′1, M″1) message based on P1 (resp. P′1 or P″1) that includes:
   3.1 *The addition of a correspondence header field that specifies the current correspondence*
   3.2. *The addition of secured extra header fields only for scenarios 1 (security label header field) and 3 (security label and military header fields)*
   3.3. *A signature with a particular cryptographic algorithm only for scenarios 1 and 3*
4. A opens a connection with S1. A and S1 negotiate the current correspondence during the SMTP transaction.
5. S1 retrieves current correspondence policy P2 (resp. P′2 or P″2).
6. S1 checks if Alice is authorized to send a private message (P′2) or a professional message (P2). S1 also checks whether Alice has been granted the role and whether this role is authorized to send a formal military message (P″2). If the authorization process succeeds, S1 continues the transfer of M1 (resp. M′1, M″1); otherwise, S1 rejects the negotiation.
7. S1 enforces P2 (resp. P′2 or P″2) and generates the M2 message (resp. M′2, M″2).
   7.1. *Adding extra header fields only for scenario 3*
   7.2. *Adding a countersignature with a particular cryptographic algorithm only for scenario 3*
8. Two cases:
   Cases 1 and 2: S1 delivers M2 (resp. M'2) in Bob's mailbox.
   Case 3: S1 sends M″2 to S2 and S2 delivers M″2 in Dave's mailbox.

9. B (resp. D) for cases 1 and 2 (resp. 3) downloads the M2 or the M′2 message (resp. the M″2 message).
10. B (resp. D) for cases 1 and 2 (resp. 3) retrieves P3 or P′3 (resp. P″3).
11. Cases 1 and 2: B displays the message and checks that M2 (resp. M′2) is in accordance with P3 (presence of a cryptographic signature with a particular cryptographic algorithm and secured extra header fields of step 3) or P′3.
Case 3: B displays the message and checks that M″2 is in accordance with P″3 (presence of a cryptographic signature with a particular cryptographic algorithm, presence of secured extra header fields of step 3, presence of a cryptographic countersignature and extra header fields of step 7).
Cases 1, 2, and 3: If the verification process succeeds, Bob has the guarantee that the correspondence policy has been fully applied to the message. If the verification process fails, a warning is displayed.

Conclusion

In this article, we have presented an innovative correspondence model applied to military messaging. Basic email systems allow multiple usages but with a single policy. This is a major limitation of current email systems. Our model allows, for each end user's usage, the precise definition of the associated correspondence. We can define an advanced military messaging system that allows implementation of multiple policies adapted to multiple usages, such as high, medium, and basic grade services.

Development of the correspondence model has been carried out on an email client: Trustedbird.[14] This email client (a project managed by the French Ministry of Defence) based on Mozilla Thunderbird provides extended security services. Applying several types of policies according to the correspondence type thus becomes possible. For example, it allows the application of policies to military message format by adding metadata in the form of header fields. It also allows the application of signature policies on military messages, in particular including the description of security labels [11] and metadata that need to be secured [12, 13]. A prototype global architecture of correspondence has also been developed. This prototype has been used to verify how to integrate the concepts of correspondence in an email system [14, 15].

The main interest of this model is to provide new prospects for military messaging systems with the ability to provide different levels of services dedicated to a military environment.

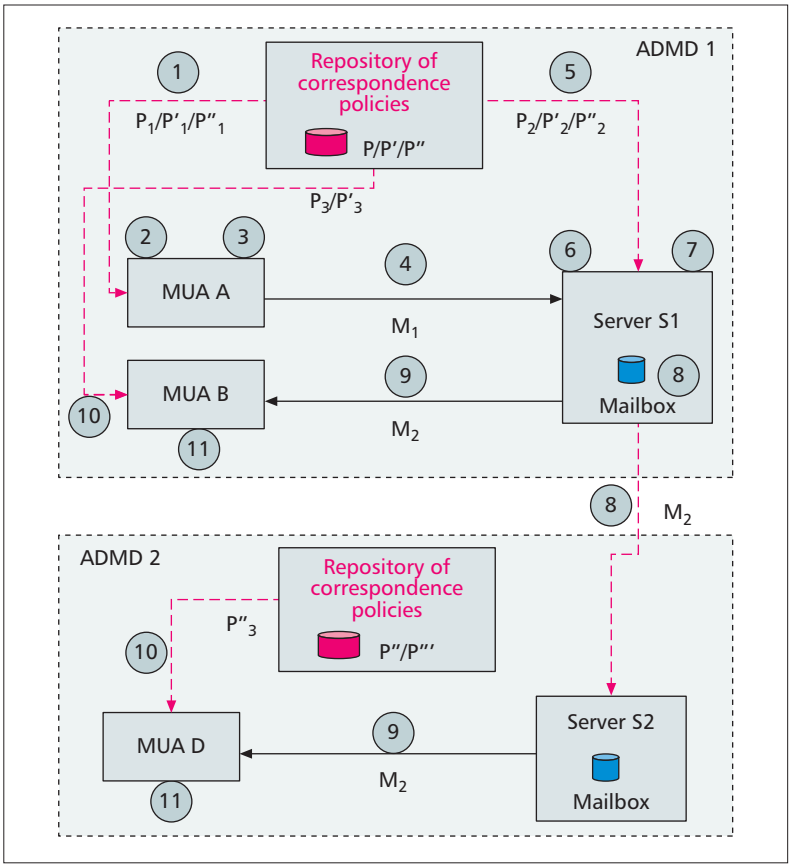


Figure 5. Email exchange based on correspondence policies.

References

[1] P. Resnick, "Internet Message Format," IETF RFC 5322, Oct. 2008
[2] J. Klensin, "Simple Mail Transfer Protocol," IETF RFC 5321, Oct. 2008
[3] M. Crispin, "Internet Message Access Protocol — Version 4rev1," IETF RFC 3501, Mar. 2003.
[4] N. Freed and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies," IETF RFC 2045, Nov. 1996.
[5] D. Crocker, "Internet Mail Architecture," IETF RFC 5598, July 2009.
[6] J. Searle, *Mind, Language and Society: Philosophy in the Real World*, Basic Books, 1998, p. 144.
[7] M. Haugh, "Intention in Pragmatics," *Intercultural Pragmatics J.*, vol. 5, no 2, 2008, pp. 99–110.
[8] A. Melnikov and K. Carlberg, "Simple Mail Transfer Protocol Extension for Message Transfer Priorities," IETF RFC 6710, Aug. 2012.
[9] A. Melnikov and K. Carlberg, "Tunneling of SMTP Message Transfer Priorities," IETF RFC 6758, Oct. 2012.
[10] A. Melnikov and G. Lunt, "Registration of Military Message Handling System (MMHS) Header Fields for Use in Internet Email," IETF RFC 6477, Jan. 2012.
[11] P. Hoffman, Ed., "Enhanced Security Services for S/MIME," IETF RFC 2634, June 1999.
[12] L. Cailleux, "A New Security Service for Future Military Messaging," *IEEE MCC 2013 Military Commun. and Info. Sys. Conf.*, 7–9 Oct. 2013, Saint-Malo, France.
[13] L. Cailleux and C. Bonatti, "Securing Header Fields with S/MIME," IETF RFC 7508, Apr. 2015.
[14] L. Cailleux, A. Bouabdallah, and J.-M. Bonnin, "Building a Confident Advanced Email System Using a New Correspondence Model," *28th Int'l. Conf. Advanced Info. Networking and App. Wksps.*, 13–16 May 2014, Victoria, Canada, pp. 85–90.
[15] L. Cailleux, *Un Nouveau Modèle de Correspondance Pour un Service de Messagerie Electronique Avancée*, Ph.D., 2014, telb0335, Telecom Bretagne — Network, Multimedia and Security Department, defended 12 Jan. 2015.

Biographies

Laurent Cailleux (laurent.cailleux@intradef.gouv.fr) received his Ph.D. degree in computer science from Telecom Bretagne, France. He is an IT expert in the Direction Générale de l'Armement, the procurement agency of the French Ministry of Defence.

Ahmed Bouabdallah (ahmed.bouabdallah@telecom-bretagne.eu) received his Ph.D. in computer science from the University of Franche-Comté, France. He joined Telecom Bretagne as an assistant professor in the Network, Security, and Multimedia Department. His research interests concern telecommunication services, security, and formal methods.

[14] http:// www.trusted-bird.org

Toward Federated Mission Networking in the Tactical Domain

Marianne R. Brannsten, Frank T. Johnsen, Trude H. Bloebaum, and Ketil Lund

ABSTRACT

NATO is currently working on the Federated Mission Networking (FMN) concept, which will become the foundation for establishing mission networks in the future. The realization of the FMN concept is described in the NATO FMN Implementation Plan (NFIP). The information infrastructure outlined in NFIP today builds on the concept of service-oriented architecture in order to achieve interoperability, and bases itself on many of the same standards and specifications as the ones identified through NATO Network Enabled Capabilities (NNEC). The NNEC SOA Baseline [1] identifies a number of core enterprise services that represent the common functionality needed to build an interoperable service-oriented infrastructure in a federation. It further identifies which standards should be used to realize these core services while ensuring interoperability between the federation members. A subset of these capabilities includes messaging services, collaboration services, service discovery, and security services. This article looks into each of these foundational core services, presents the challenges related to extending support for these services into the tactical domain, and identifies potential solutions.

INTRODUCTION

The North Atlantic Treaty Organization (NATO) is working on the Federated Mission Networking (FMN) concept to enable efficient establishment of mission networks in the future. FMN consists of three parts:

- The FMN framework, which serves as a template for how to build mission networks
- A number of mission network instances
- A governance structure that oversees both the FMN framework and the specific mission network instances

FMN as a capability will continue to develop over time, and the approved concept[1] uses a spiral approach to the development of FMN.

To realize the FMN concept, NATO is working on the NATO FMN Implementation Plan (NFIP), which is divided into three volumes. Volume I [2] covers the overall concept and governance; Volume II covers the FMN framework; and Volume III describes the common NATO capabilities. At the time of this writing, the current version of the NFIP is 3.0, which outlines a spiral approach for FMN implementation that aims to have an initial capability with limited functionality defined in Spiral 1. The Spiral 1 ambition level is to establish a basic capability, which supports a limited set of mission threads, and enables information exchange down to the deployed headquarters (HQ) level. Extending the capability to other mission threads and enabling interoperability in the tactical domain is left for future spirals.

The NFIP today consists of many standards identified in the NATO Network Enabled Capabilities (NNEC). Both FMN Spiral 1 and the NNEC Service-Oriented Architecture (SOA) Baseline focus on interoperability between federation members on the strategic and operational levels. Neither of these addresses the additional requirements that arise when includbeing interactions at the tactical level, but later spirals of FMN include this in their ambition levels. When extending support for the core services into the tactical domain, the service implementations need to be adapted to the specific limitations encountered in tactical communications networks. A disconnected, intermittent, and limited (DIL) environment is a common description of an environment characterized by the possibility of periodic communication disruptions, bad connectivity, and limitation problems when it comes to both network (e.g., low data rate) and node capabilities (e.g., battery life, storage capacity, CPU power).

The Core Enterprise Services (CES) identified through the NNEC SOA Baseline cover a broad set of capabilities. In the tactical domain, the set of functional services required by the users is smaller than what one would expect to see at higher operational levels [3]. As a consequence of this, the set of CES required at the tactical level is likely to be smaller than what is required in a mission network as a whole.

The NATO RTO/IST-118 Working Group "SOA Recommendations for Disadvantaged Grids in the Tactical Domain" has identified a subset of the CES, which form a set of foundational core services that should be supported at the tactical level:

Messaging services: These services enable exchange of messages between systems, and are

The authors are with the Norwegian Defence Research Establishment (FFI).

[1] The Future Mission Network concept was approved by the Military Committee on 16.11.2012, and its name has since changed to Federated Mission Networking.

0163-6804/15/$25.00 © 2015 IEEE

needed to support basic functions such as blue force tracking, distribution of sensor information, and sharing of plans.

Collaboration services: These services enable communication between humans, and include functionality such as instant messaging, video conferencing, and document sharing. Coordination between units involved in the same mission requires that at least a subset of these services is available.

Service discovery: The inherent limitations of communication resources in the tactical domain mean that the availability of services will change over time, and service discovery is needed to handle this dynamicity.

Security services: Information exchange in the tactical domain must be protected to ensure confidentiality, integrity, availability, authenticity, and non-repudiation.

The CES subset listed above is presented in more detail with accompanied tests and evaluations according to our work on the subject in the following sections. We start by looking at messaging services where computer systems on a system level communicate using messages. This fundamental part of information exchange is a first step toward core services on a tactical level.

Messaging between Systems

Request/response is a message exchange pattern in which a requestor sends a request message to a service. The service will process the request and return a response to the requestor. Together with the publish/subscribe message pattern, request/response covers the vast majority of interaction patterns between computers.

Normally, the request/response pattern is implemented in a synchronous fashion, where the connection between the requestor and the service is kept open until either the response is returned or the request times out. However, it can also be done asynchronously, such that the connection is closed as soon as the request is delivered, and then the response is delivered at some later time, through a callback function. The latter is especially useful when the processing time of the service can be long.

In addition, the request/response pattern can be used for a push-based message delivery pattern, where the data is delivered in the request message, and the recipient only responds with an acknowledgment message.

As opposed to request/response, the publish/subscribe paradigm relieves the client of having to check for new data. Instead, the node simply sends a subscription request to the information provider, asking to be notified whenever new information is available. This has several advantages: The network traffic is reduced, since the client does not have to send periodic requests; the server load is reduced, since there are fewer requests to process; and the client will potentially receive new data sooner, although this is dependent on the request frequency in a request/response setting (which in turn will affect network and server load). For a given subscription, the notifications are normally always of the same type, independent of the actual information that is delivered (i.e., the payload of the notification). When a client wants to subscribe to a specific type of data, it therefore expresses the type of information it is interested in by including a topic in the subscription request.

Web Services

Web services are based on loose coupling between client and server, and instead of having to rely on application programming interfaces (APIs), the focus is on message formats. Thus, a web service can be used by any platform that supports exchange of messages, which conform to the format used by the service interface. Web services often use the XML-based Simple Object Access Protocol (SOAP) for information exchange, and are in widespread use on the Internet today, with civil and commercial products and development tools readily available. Both request/response and publish/subscribe are supported.

For publish/subscribe, the use of WS-Notification is specified, including all sub-specifications (WS-BaseNotification, WS-BrokeredNotification, and WS-Topics).

In a NATO context, there are two general requirements that must be met by any message exchange mechanism: interoperability and ability to function in DIL environments.

Web Services in DIL Networking Environments

Web services in general focus on environments with static networks and abundant data rates, which in a military context typically means strategic, operational, and deployed tactical levels. Consequently, the overhead associated with web services is not a problem in such environments.

However, in FMN the challenge is to enable users to exchange information with each other at all operational levels. This includes users in the field who may only communicate with others over radio systems with DIL characteristics. Radio systems such as HF or VHF may have a very low data transfer rate due to the need for long range signals and jamming resistance. In addition, some radio systems suffer from long turn times for directional changes, plus long setup times for connections.

Reducing the traffic generated is thus necessary. This can be done by both the application itself and the platform/communication system [4]. Filtering done by the application is typically based on message content (e.g., only send tracks within a certain radius from the user). On the platform level, filtering is typically based on criteria such as importance of the message, type of data (e.g. text or video), and priority.

In addition, a common way of reducing network traffic is through compression. Although verbose, XML-based messages are compression friendly, and the size can be reduced significantly, even with standard compression mechanisms like gzip [5].

As mentioned above, NATO has chosen the WS-Notification standard for publish/subscribe. This standard is well suited to strategic networks, but may require some adaptation for deployment in tactical networks. We have

Request/response is a message exchange pattern in which a requestor sends a request message to a service. The service will process the request, and return a response to the requestor. Together with the publish/subscribe message pattern, request/response covers the vast majority of interaction patterns between computers.

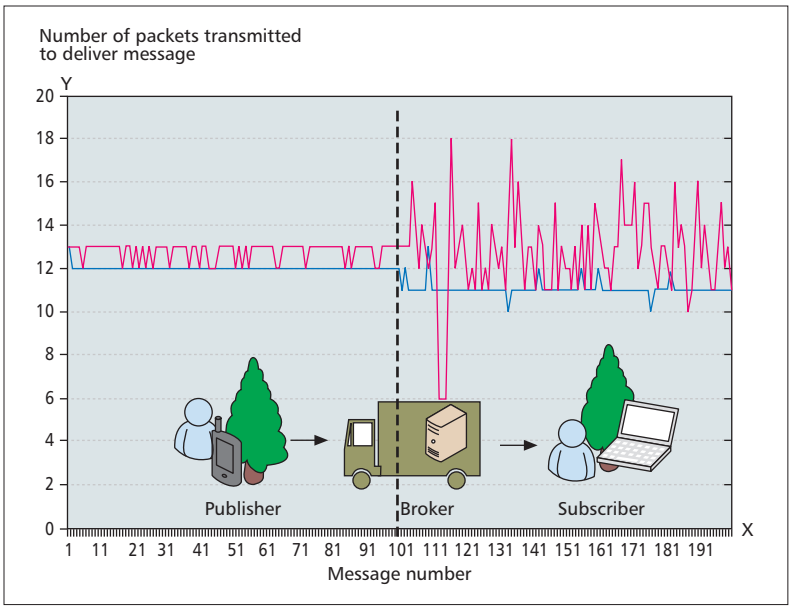


Figure 1. Publish/subscribe over tactical radio.

attempted to use WS-Notification over tactical broadband radios, and our results show that it functions, but that loss of messages must be expected under poor networking conditions. Figure 1 illustrates this using actual radios. On the x-axis the interval 1–100 is the total message count for the first half of both runs from publisher to broker. The interval 101–200 on the x-axis shows the second half of the experiment runs from broker to subscriber. The y-axis shows the packet count; thus, the fluctuation illustrates poor conditions leading to retransmissions. In good conditions 12 packets constitute a message sent, and at times more packets are sent in retransmission in order to try delivering the packets and other times the message is lost.

We performed two experiments, using Kongsberg WM600 radios and a network degradation tool (a matrix of attenuators) for emulation of poor link conditions, using NATO Friendly Force Information (NFFI) [6] over WS-Notification. Note that Fig. 1 shows some fluctuations in traffic, which can be attributed to signal loss and routing changes. If we were to simulate this, instead of using actual hardware, these results would have looked "cleaner." The scenario is simple: a deployed user periodically reports his/her position to a broker, which relays the information back to the tactical forward deployed at HQ. In the first run (depicted by the blue line), the radios are well within range and fully connected. One hundred messages consisting of multiple NFFI messages were sent, and all were received by HQ. In the second run (depicted by the red line), the conditions are good between publisher and broker, but poor on the link between broker and subscriber (i.e., HQ). HQ only received 85 percent of the issued messages due to packet loss and the fluctuation in the packet count resulting in retransmissions.

It should be noted that WS-Notification is a relatively simple standard. In principle, it is a reversed request/response service, in the sense that the server invokes a web service at the client side when delivering a notification. In addition, the standard is based on unicast message transmission only, which may have implications in limited capacity networks; even when multiple nodes in the same network want the same information, a WS-Notification broker sends one unicast message to each recipient rather than sending one multicast message that reaches all recipients. When subscribers unexpectedly leave the network, permanently or temporarily, WS-Notification is unable to deliver messages to them. NATO has created an add-on to the WS-Notification standard that allows caching of messages so that they can be saved for later delivery. In radio-based networks, where the transmission medium is shared, there is potential for a significant reduction in network load by switching from unicast to multicast. In this case, a reliable multicast mechanism would seem necessary. Note that making such a switch will require further functionality to be implemented by WS-Notification, that is, the ability to manage multicast group memberships.

Messaging services as described above enable information exchange between systems. At the next level we need to facilitate collaboration between humans. The following section discusses the parameters for enabling functionality like chat and video conferencing.

Collaboration

Collaboration services are part of the NATO CES, but differ from other services in that they are not pure middleware services as such, but provide functionality directly to the user. Examples of typical collaboration services include audio, video, and chat. The NNEC SOA Baseline covers collaboration services, but only points to a standard to use for chat (i.e., XMPP — the Extensible Messaging and Presence Protocol). In this section we summarize available collaboration services and their suitability for the tactical domain.

Commercially available collaboration services can enable collaboration between soldiers in different physical locations. Most current technologies are geared toward use across the Internet or within an enterprise network. The common denominator here is high data rate be and fairly stable network connections. We have identified three main challenges of using collaboration services in the tactical domain:

1. How applicable are the collaboration services to DIL environments, and to what extent can they be adapted to the tactical domain?
2. Interoperability is of paramount importance. The services must be able to interoperate with other systems.
3. Security must be upheld and compatible with the direction NATO is taking with FMN and the NFIP.

Text-Based Collaboration

Text-based collaboration services include chat, which allows users to exchange text messages. The messages can be delivered either directly between two participants (instant messaging), or between several participants (chat room). In

	Chat	Audio and video	Data-centric collaboration services
Adaptation to the tactical domain	Solutions are known and tested.	Known solutions work well in the civil domain, but must be tested in the tactical domain. A specific implementation must be explicitly tested for compliance.	New trends in the civil domain barely introduced in the military domain.
Interoperability	Agreement on XMPP, but tactical adaptations and security protocols are not standardized. Requires a gateway between proprietary solution and standardized XMPP to function seamlessly.	Several standards exist, but in practice interoperability is not always achievable.	Well established standards in the civil domain, but these have to be adapted to the tactical domain.
Security	Known mechanisms can be applied, but open issues exist related to tactical adaptations and interoperability.	Interoperability issues related to streaming and many-to-many communication.	No support for classified information. Largely based on network and transport layer security.

Table 1. Current state of collaboration services (adapted from [9]).

NATO, the XMPP-based JChat is being used. XMPP functions very well in stable, infrastructure-based networks. However, as our previous research has shown [7], it is not a protocol directly applicable in the tactical domain. When we talk about increased mobility here, we mean that the nodes' velocity (in meters per second) increases. The mobility model used was random waypoint. As mobility increases, XMPP gradually delivers fewer chat messages, our bespoke solution, Mist [8], delivers more than 99.5 percent of the messages in all experiments. Mist is an experimental middleware that implements application layer multicast. It is especially designed for use in mobile DIL environments. We use this middleware as a foundation for both experimental chat and service discovery applications. The bandwidth consumed by XMPP is approximately double that of Mist. For complete experiment details, see [7]. These aspects indicate that XMPP is best used in infrastructure-based static networks, whereas other solutions should be employed in the tactical domain where resources are scarce and nodes are mobile. Note that all communication with the XMPP server was compressed using zlib compression, and we used a single multi-user chat room.

Audio, Video, Application, and Data Sharing Collaboration

These kinds of services are well understood and much used in the civil domain, and can also be used in infrastructure-based networks. However, they are to a lesser degree employed in tactical networks (particularly those with high mobility and low data rate) where conventional solutions do not function.

For all of the above areas, the current state of the collaboration services is shown in Table 1. There, the colors of the cells indicate the state of available solutions and products. The green area in the table indicates that even though there may still be open issues, this area is well covered by current solutions. The yellow areas indicate that there are known solutions, but there are still some open issues. The magenta areas indicate that more research is needed. For further details, see the survey of existing applicable technologies given in [9].

Messaging and collaboration both play an important part toward FMN. In addition to these services, as we discuss in the next section, service discovery is important to facilitate an up-to-date view of the available services (e.g., weather, chat, email).

Service Discovery

Operating in a DIL environment puts an additional demand on service discovery. Services will change over time, and this dynamicity has to be handled accordingly.

A strategic-level fixed network can employ civilian standard service registries for discovery like UDDI or ebXML. UDDI and ebXML are central registries, constituting a single point of failure. The further one moves from fixed networks, the more one needs an interoperability gateway to mediate service discovery between levels [3].

Web Service Discovery in DIL Networking Environments

Deploying technology designed for fixed-infrastructure networks on tactical networks might not be feasible as resources can be scarce and there are no guarantees of connectivity at any given time. Web service discovery is an important part of the web service scheme, and in [10] we evaluate web service discovery in military tactical networks. Important criteria for service discovery in such environments are that discovery needs to be distributed and robust. All participants in an operation need to be able to both use and potentially share their services.

As mentioned above, standard web service registries are not suitable for DIL environments. WS-Discovery is somewhat better as it is a distributed solution, eliminating the problem of a single point of failure, but it is still designed for an office environment and does not take into consideration the specific challenges of a DIL environment.

Another service discovery parameter to consider is if the service discovery solution is reac-

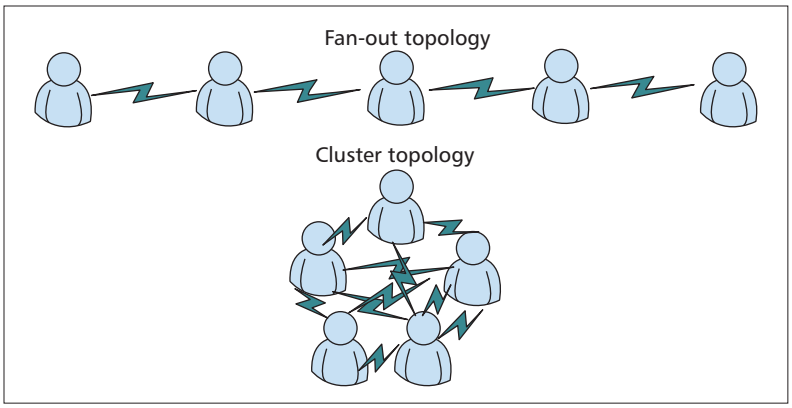


Figure 2. Fan-out-topology vs. cluster topology.

tive (the client probes for updates) or proactive (updates are sent automatically at given time intervals). A client will include a cache of discovered services, and either the client regularly probes to see what services are available at the moment and eliminates cache entries that are no longer responding to probes (reactive), or the services regularly issue advertisements and the client adds advertised services to the cache and declares services that are no longer being advertised to be stale (proactive). More frequent probing consumes more bandwidth, but allows clients' caches to be more accurate to conditions. Each client probing the network individually consumes more bandwidth than each service issuing regular advertisements. However, clients would need to time-out stale cache entries quickly if they have not heard an advertisement in a while. Less frequent probing or advertising-stale decisions saves bandwidth at the cost of client cache accuracy.

To evaluate web service discovery solutions in DIL environments we considered both reactive and proactive discovery protocols. Protocols designed for DIL environments are SAM and Mist. SAM is short for service advertisements in mobile ad hoc networks (MANETs) [3]. It is an experimental service discovery protocol developed especially for use in DIL environments. SAM service advertisements are sent at regular time periods using IP-multicast. Mist, on the other hand, does not rely on IP-multicast. The last solutions to be evaluated are the Service Location Protocol (SLP) originally designed for other purposes, but with some adaptation enabled to be used for web service discovery, and finally, WS-Discovery.

Testing and Evaluation of DIL Web Service Discovery

In order to illustrate the usefulness of different web service discovery protocols in DIL environments, we have evaluated them in specific node configurations according to bandwidth usage. The result shows there is a need for specialized protocols to efficiently operate in DIL environments.

The participating nodes are configured in different types of topologies during a mission, as shown in Fig. 2. There are in some cases several smaller groups where the physical topology of the groups vary from a cluster topology, where the nodes are all in reach of each other, to a fan-out topology, where the nodes form a line, and there might not be connectivity between all the nodes, but only between direct neighbors.

Table 2 describes the average data rate of cluster and fan-out topologies. In the evaluation we used an in-house developed test tool to make sure the four evaluated protocols were tested on equal terms.

The first part of Table 2 describes the results of evaluating a fan-out topology. WS-Discovery and SLP are reactive protocols. They use the most bandwidth in contrast to Mist and SAM. This is true for both central and edge nodes. In a fan-out topology it is important to notice the difference between average data rate of the central node of the topology and the nodes residing at the edge of the topology because the central nodes are responsible for forwarding on behalf of other nodes.

The second part of Table 2 describes the average data rate results for a cluster topology. In a cluster all the nodes are within reach of each other. If you look at the per node result, it is approximately the same as for the edge nodes in the fan-out topology as the topology enables direct requests to all nodes. The bandwidth per query is a calculation of bandwidth divided by queries per second, while the proactive ones have constant usage of bandwidth to maintain state.

The result of testing web service discovery in different military squad topologies revealed that WS-Discovery and SLP should not be used if bandwidth is a concern. The experimental solutions Mist and SAM showed a much more efficient solution in DIL environments.

Handling service discovery in a DIL environment is a challenge. Another challenge is to ensure the security of services. In the next section, securing web resources is discussed.

Security

In FMN, information exchange and collaboration between nations span more than one security domain. Inside a security domain the web resources are often guarded with security solutions, such as web authentication where participants are granted and denied access based on, for example, username and password. In a federated scenario we manage such authentication by setting up trust relations between the different systems enabling participants to access several security domains authenticating only to their local authentication authority in a single sign on (SSO) scheme.

However, in a military setting there are some potential limitations to take into consideration. In the civil arena there are more than enough data rate and stable connections, but in DIL environments the overhead created by the security solutions might be costly for the network and influence the stability of the resources available.

Web Authentication

Users often have to authenticate before they are granted access to a web resource, and it is the site owner that handles the authentication scheme (e.g., username and password). More

and more sites collaborate in order to give the users a complete solution to their specific problem. As an example, consider a user going on a trip and needing a flight and a rental car. The airline and car service might have separate sites, and if the user had to authenticate at both sites it could become troublesome and time consuming. Remembering several passwords is a challenge in itself. But if the airline and car service collaborate, and they agree to trust each other to authenticate users in an SSO setting, they allow a user to authenticate at one site and get access to the other site automatically.

In [11] we study the cost of adding an SSO solution using SAML 2.0 in a federated environment. These results are complementary to the work done in [12] where we measured and evaluated overhead of SOAP security and showed that the SOAP messages increased in size by a factor of five when compared to running with no security. The study exposed a potential problem of bringing such solutions to DIL environments. Traditionally, SSO is handled in browser cookies, where the cookies store access information. When an SSO solution is to include more than one Domain Name System (DNS), the cookie solution will not work, as a cookie may not be shared between DNS domains. The SSO scheme adopted in our experiments uses a central service handling web authentication. The entities handling security are an identity provider (IdP) and a service provider (SP). Digital certificates form the basis of trust between IdP and SP, so a functioning public key infrastructure (PKI) is a prerequisite for SSO.

SP is the security lock on the resource that can be opened by the security tokens. The security tokens are the responsibility of the IdP, which produces and distributes security tokens when the identity of a user is established.

Average data rate for fan-out topology				
Protocol	Mist	WS-Discovery	SAM	SLP
Central	0.06 kB/s	14.90 kB/s	0.28 KB/s	7.12 kB/s
Edge	0.05 kB/s	2.28 kB/s	0.02 KB/s	1.05 kB/s
Average data rate for cluster topology				
Protocol	Mist	WS-Discovery	SAM	SLP
Total	0.62 kB/s	27.30 kB/s	0.27 kB/s	12.57 kB/s
Per node	0.05 kB/s	2.27 kB/s	0.02 kB/s	1.05 kB/s
Per query	N/A	27.08 kB/q	N/A	12.59 kB/q
Per query/node	N/A	2.26 kB/q/n	N/A	1.05 kB/q/n

Table 2. Average data rate be for cluster and fan-out topologies (from [10]).

Enterprise vs. Federated SSO

In a military setting, a web resource federation scenario is very important as it enables collaboration between coalition partners, enabling them to easily share web resources. But when bringing SSO into a federated scenario, there are even more challenges implementing a SSO solution in a federated DIL environment.

The enterprise scenario requires all the participants to belong to the same enterprise, and Fig. 3 depicts two enterprises in a federation. The consumer has a direct trust relationship to the local IdP, and the IdPs of the different enterprises forms a trust relation across enterprise borders. The SP secures a web resource. When the consumer, in the remote domain, requests access to a web resource without an authentication token, the SP redirects to the consumer's local domain's IdP, and the consumer is prompted to authenticate. If the consumer successfully authenticates to the local domain's IdP, the IdP further requests a token from the remote domain's IdP authorizing access to the web resource.

Security Overhead in DIL

There are two ways of initiating web authentication, SP-initiated and IdP-initiated. This is defined as where the user goes to first. A user

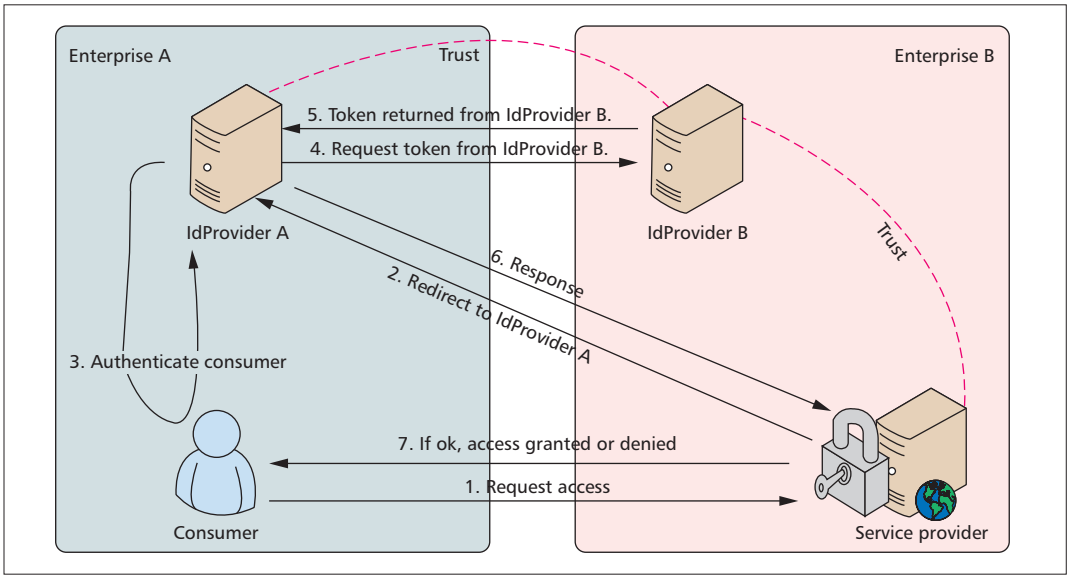


Figure 3. Federated web authentication.

Federated SSO	Network traffic in bytes		
	Network traffic	Payload	Overhead
SP-initiated SSO (not logged in)	7459	207	7252
SP-initiated SSO (logged in)	7089	207	6882
IdP-initiated SSO (not logged in)	5505	207	5298
IdP-initiated SSO (logged in)	5340	207	5133

Table 3. Federated web authentication [11].

can start by going to an SP, getting redirected to an IdP, and then directed back after authentication, or the user can go to the IdP first, authenticate, and then manually go to the SP to request access.

Table 3 shows the measured results of SP initiated SSO and IdP initiated SSO. The overhead ranges from 5233 to 7252 bytes. If the user already has a valid token, the overhead only diminishes between 165 and 370 bytes. This tells us that the evaluation of how long a token is valid is not that important when considering overhead. The result is rather that adding security is costly, but at the same time it is a necessity.

There are experimental approaches to supporting SSO in DIL environments with low overhead, an example being [13]. However, interoperability has largely been neglected in such experimental solutions. Hence, further research is necessary in order to bring interoperable SSO to the tactical domain.

Summary

The realization of the FMN concept rests on NATO's work on the NFIP. Future NFIP spirals include work on enabling interoperability in the tactical domain. NNEC also focuses on interoperability, and points to SOA and web services as an enabling technology. CES identified for this enablement are, among others, messaging, collaboration, discovery, and security. These CES were evaluated on the tactical level in this article. We found that further research is still needed in different areas in pursuit of fully realizing FMN in the tactical domain.

References

[1] Consultation, Command and Control Board (C3B), Core Enterprise Services Standard Recommendations: The SOA Baseline Profile Version 1.7, Enc. 1 to AC/322-N(2011)0205, NATO Unclassified releasable to EAPC/PFP, 11 Nov. 2011.

[2] NATO Allied Command Transformation, "NATO FMN Implementation Plan v. 3.0, Volume I," approved by North Atlantic Council, Aug. 6, be 2014.

[3] F. T. Johnsen *et al.*, "Web Services Discovery across Heterogenous Military Networks," *IEEE Commun. Mag.*, Oct. 2010, pp. 84–90.

[4] NATO STO, STO-TR-IST-090 — SOA Challenges for Real-Time and Disadvantaged Grids. STO-TR-IST-090 AC/323(IST-090)TP/520. Final Report of TR-IST-090, Apr. 2014.

[5] M. A. Teixeira *et al.*, New Approaches for XML Data Compression," *Proc. Int'l. Conf. Web Info. Sys. and Technologies*, 2012, pp. 233–37.

[6] NC3B Info. Sys. SC, Interim NFFI Standard for Interoperability of FTS, AC322(SC5)N(2006)0025, 16 (approved on 16 Dec. 2006).

[7] M. Skjegstad *et al.*, "Distributed Chat in Dynamic Networks," *IEEE MILCOM 2011*, 7–10 Nov. 2011, Baltimore, MD, pp. 1651–57.

[8] M. Skjegstad *et al.*, "Mist: A Reliable and Delay-Tolerant Publish/Subscribe Solution for Dynamic Networks," *2012 5th Int'l. Conf. New Technologies, Mobility and Security*, 2012

[9] E. Gjørven *et al.*, "Towards NNEC — Breaking the Interaction Barrier with Collaboration Services," FFI-Report 2014/00943, http://rapporter.ffi.no/rapporter/2014/00943.pdf

[10] M. Skjegstad, F. T. Johnsen, and T. Hafsøe, "An Evaluation of Web Services Discovery Protocols for the Network-Centric Battlefield," *Military Commun. and Info. Sys. Conf 2011*, Amsterdam, Netherlands, 17–18 Oct. 2011.

[11] M. R. Brannsten, "Federated Single Sign On in Disconnected Intermittent and Limited (DIL) Networks," *IEEE VTC 2015-Spring Int'l. Wksp. Service-Oriented Computing in Disconnected, Intermittent and Limited Networks*, Glasgow, Scotland, May 2015.

[12] Trude Hafsøe *et al.*, "Using Web Services and XML Security to Increase Agility in an Operational Experiment Featuring Cooperative ESM Operations," *14th Int'l. Command and Control Research and Tech. Symp.*, Washington, DC, June 2009.

[13] A. Fongen, "Federated Identity Management in a Tactical Multi-Domain Network," *Int'l. J. Advances Sys. and Measurements*, vol. 4, nos. 3 and 4, 2011

Biographies

Marianne R. Brannsten (Marianne-rustad.brannsten@ffi.no) is a research scientist at the Norwegian Defence Research Establishment (FFI), engaged in theoretical research and practical development in areas such as distributed systems and service oriented architecture. She received her Master's degree from the University of Oslo (UiO) in 2006, and has been working at FFI since then.

Frank T. Johnsen (frank-trethan.johnsen@ffi.no) received his Ph.D. from UiO. He started work as a scientist at FFI in 2006. At FFI he is currently working within the area of secure pervasive SOA. His research interests include web services, quality of service, and middleware. He also holds a position as a part-time associate professor at UiO.

Trude H. Bloebaum (trude-hafsoe.bloebaum@ffi.no) is a scientist at FFI, where she has been working since 2006. Before coming to FFI she worked with content distribution systems at UiO. She received her Cand.scient. degree from UiO. Her research interests are web services, quality of service, and network protocols.

Ketil Lund (ketil.lund@ffi.no) is a scientist at FFI, where he has been working since 2006. His research interests include service oriented architectures, web services, quality of service, and middleware. At FFI he is currently working within the area of secure pervasive SOA. He received his Ph.D. in informatics from UiO.

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.

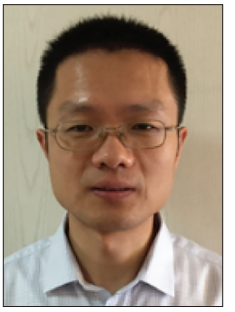
Articles should be submitted to the *IEEE Communications Magazine* submissions site at

http://mc.manuscriptcentral.com/commag-ieee

Select "Standards Supplement" from the drop-down menu of submission options.

FUTURE RAILWAY COMMUNICATIONS



David W. Matolak *Marion Berbineau* *David G. Michelson* *Chen Chen*

The growth of intelligent transportation systems (ITS) has recently accelerated as governments and regulators seek to use ITS to achieve important economic and policy goals. ITS must be efficient, environmentally friendly ("green"), and, of course, safe. The scope of ITS is broad, and all transportation modes will require communications, including future vehicular (automotive) systems and vehicle-to-vehicle/vehicle-to-infrastructure (V2V/V2I) communications, aeronautical communications, maritime communications, satellite communications, railway communications, and possibly others.

The growing use of railway communications for both passenger and freight transportation has yet to attract sufficient attention in the literature. With numerous high-speed rail systems being used and planned in China and Europe in particular, as well as the development of subways and tramway lines, reliable communications for both railway control/safety and passenger applications are of great current interest.

This Feature Topic on future railway communications should have broad appeal to this magazine's readership, as rail systems are used worldwide by millions of people daily. We have accepted eight papers for this Feature Topic, covering topics from channels and antennas to railway cellular and security.

"A Survey on Future Railway Radio Communications Services: Challenges and Opportunities" presents the main requirements and trends for the use of wireless systems in the high-speed rail, subway, and tramway contexts. The main operational services are described, and important key performance indicators (KPI) as well as safety requirements are given.

Of all the studies conducted in support of future railway communication systems, none are more fundamental than channel measurement. "Channel Sounding for High-Speed Railway Communication Systems" reviews the state of the art in radio channel sounding techniques for high-speed rail (HSR) and proposes a novel LTE-based HSR channel sounding scheme.

In the article "Future Railway Services Oriented Mobile Communications Network," the authors introduce the concept of the fifth generation (5G) for railways (5G-R), addressing 5G networks and key technical implementations in future railway development. Heterogeneous network architecture with Global System for Mobile Communications for Railway (GSM-R), Long Term Evolution for Railway (LTE-R), 5G-R, and integrated transportation are proposed.

In "WDM RoF-MMW and Linearly Located Distributed Antenna System for Future High-Speed Railway Communications," the authors propose a dual-hop architecture capable of providing high-speed communications for high-speed rail. Based on a fiber optic backhaul and millimeter-wave radio access network, the system uses conventional WLAN to provide service to users.

In "Providing Current and Future Cellular Services to High-Speed Trains," the authors provide an interesting comparison between using direct links from railway remote units to subscribers inside trains and using an intermediate relay unit located outside the train. Despite the high penetration losses incurred via propagation through train windows, the direct links provide a simpler and less expensive solution, albeit achieving lower system throughput. Other issues, both technical and operational, are also discussed, indicating directions for future work in this area.

In "Automatic Train Control over LTE: Design and Performance Evaluation," the authors propose a quality of service (QoS) management scheme for train control traffic based on methodologies used in conventional LTE and demonstrate how a QoS policy can be devised based on analysis of that traffic.

Reliable and secure transmission of train control data is of the utmost importance, as it allows safe operation of the train. In "Cybersecurity Analysis on Next Generation Train Control Systems," the authors present a detailed security analysis of the train operation control system under current and future scenarios. Based on their analy-

sis, a more robust cryptographic mechanism, a new key distribution, and a key storage scheme are proposed for railway systems.

Wireless systems are used for sensor applications, too. In "Ultra-Wide-Bandwidth Systems for the Surveillance of Railway Crossing Areas," the authors raise the possibility of using UWB wireless systems to detect, localize, and discriminate vehicles or obstacles that might be entrapped in a level crossing area so that an appropriate alarm can be raised or appropriate action taken.

BIOGRAPHIES

DAVID W. MATOLAK (MATOLAK@cec.sc.edu) received his B.S. degree from Pennsylvania State University, his M.S. degree from the University of Massachusetts, and his Ph.D. degree from the University of Virginia, all in electrical engineering. He has more than 20 years' experience in communication system R&D and deployment, with AT&T Bell Labs, L3 Communication Systems, MITRE, and Lockheed Martin. He has over 100 publications, eight patents, and expertise in wireless channel characterization, spread spectrum, and ad hoc networking.

MARION BERBINEAU (marion.berbineau@ifsttar.fr) received an Engineer degree from Polytech Lille, France, and a Ph.D. degree from the University of Lille, both in electrical engineering, respectively. She is a research director at IFSTTAR and deputy director of the Components and Systems Department. She is an expert in the field of radio wave propagation, wireless systems for telecommunications, and localization for ITS, particularly for railways. She is active as an expert in GSM-R and future systems (e.g., LTE-R, 5G).

DAVID G. MICHELSON (davem@ece.ubc.ca) received his B.A.Sc., M.A.Sc., and Ph.D., all in electrical engineering, from the University of British Columbia. He now leads the Radio Science Lab at UBC and serves as co-director of the AURORA Connected Vehicle Test Bed. His research interests include antenna design and channel modeling for railway communications and intelligent transportation systems. He is a member of the Boards of Governors of both the IEEE Communications and Vehicular Technology Societies.

CHEN CHEN (chen.chen@samsung.com) received his Ph.D. degree in communication engineering from Xidian University, Xi'an, China, in 2005. Since 2005 he has been with Samsung Telecom R&D Center, Beijing, China. His research experience has been on mobile communication system R&D for FDD-LTE eNodeB and TD-LTE eNodeB; he is now in charge of the Department of Advanced Technology Research, including enhanced UL CoMP, FD-MIMO, FeICIC, and so on. His research interests also include intelligent transportation systems and vehicular ad hoc networks.

A Survey on Future Railway Radio Communications Services: Challenges and Opportunities

Juan Moreno, José Manuel Riera, Leandro de Haro, and Carlos Rodríguez

ABSTRACT

Radio communications is one of the most disruptive technologies in railways, enabling a huge set of value-added services that greatly improve many aspects of railways, making them more efficient, safer, and profitable. Lately, some major technologies like ERTMS for high-speed railways and CBTC for subways have made possible a reduction of headway and increased safety never before seen in this field. The railway industry is now looking at wireless communications with great interest, and this can be seen in many projects around the world. Thus, railway radio communications is again a flourishing field, with a lot of research and many things to be done. This survey article explains both opportunities and challenges to be addressed by the railway sector in order to obtain all the possible benefits of the latest radio technologies.

INTRODUCTION

The amount of people who take a train every day in either subways, high-speed lines, or tramways is innumerable. In many of these trains, safety and non-safety services are provided through communications systems. The use of these systems is now a market trend in the industry, mostly in Europe, the United States, China, and Japan, but also in many emerging countries.

The value-added features that radio communications have provided to railways can be grouped into three types of services: safety-related, which are responsible for the safe movement of trains; operational non-safety services, including services for operators or stakeholders without safety implications, like CCTV, passenger information, remote maintenance, and sensing. Finally, the third group is devoted to providing Internet access to onboard passengers. The three of them have different exigencies due to the diverse nature of the requirements demanded of the network.

The maturity of some of these services is varied, but all of them show a lower degree of development than is desirable. For example, some operators provide Internet access for their passengers, but the available data throughputs are very poor. Moreover, the niche market condition of railways and the trend of having "one service, one radio" have driven cost steeply upward. Thus, the application of more advanced radio communication systems can bring about a lot of advantages for passengers and railway operators in all these fields: having safer and closer trains moving at higher speeds means transporting more people with better quality of service; providing customers with a good sense of security, especially in mass transit, where real-time CCTV is vital for almost every railway operator; and finally, giving a good onboard connection to the Internet implies higher incomes for operators. There are many more advantages, but those are the most relevant ones.

All these aspects are big challenges for various reasons: safety services have strong requirements in terms of reliability, availability, timing, and, of course, security. Also, operational radio telephony is a technology that belongs to the field of public safety (with all that means in terms of group calls, functional addressing, device-to-device communication, etc.). Some operational services (e.g., video streaming) are very demanding in terms of bandwidth, and providing Internet access is not a trivial task when the terminals are traveling at speeds close to 350 km/h. Finally, implementing vehicle-to-vehicle (V2V) communications could become a substantial upgrade for some services.

Today, Long Term Evolution (LTE) is a reality in public mobile communications, but it is still only a promising technology for railways. Its standardization group is trying to overcome some of the problems that made third generation (3G) technologies somewhat of a failure on railways, and as we will see, LTE is very likely to provide an excellent framework for the desired radio convergence that would be able to offer all these services over a single media, with quality of service and providing security mechanisms.

In this article we provide summarized insight into all of these challenges, mapping them to technological issues. We discuss far and near future aspects, highlighting some railway issues that are more relevant and trying to glimpse the future of the field of radio communications in railways. As far as we know, there is no survey of these characteristics in the literature.

Juan Moreno and Carlos Rodríguez are with the Metro de Madrid S.A.

Leandro de Haro and José M. Riera are with Universidad Politécnica de Madrid (UPM).

This line of work is supported in part by the Spanish Ministry of Economy and Competitiveness, through project TEC2014-57821-R.

0163-6804/15/$25.00 © 2015 IEEE

The structure of this article is as follows. The challenges for critical services (related to safety) are summarized. Value-added operational (non-safety) services are described, and then how to provide Internet access to onboard passengersis discussed. We provide insight on some general technological aspects related to all the previous types of services; and finally, conclusions are presented.

Safety Services

There are two types of critical services: those related to the safety of the train itself (railway signaling) and public safety ones (including operational voice among others). In this section we mention both of them. Typically, all safety-based services need the highest safety level (SIL4 [1]), low bandwidth (less than 1 kb/s per train), significant delay constraints (less than 500–800 ms in the worst case, usually even less) and the traffic pattern is usually real-time variable bit rate (RT-VBR). Voice calls need more kilobytes per second (i.e., 64 kb/s), but it depends on the codec; a good reference for maximum jitter could be 30 ms.

Signaling for High-Speed Trains

The state-of-the-art signaling system for mainline and high-speed rail is the European Rail Traffic Management System (ERTMS). In this system, both levels 2 and 3 strongly depend on the train-to-wayside communications (provided by the Global System for Mobile Communications — Railway, GSM-R, system) to work. ERTMS level 2 is a successful technology all around the world. However, level 3 (which implies the removal of track circuits; Fig. 1) shows a bit of reluctance to follow the path of level 2. The reason behind this (other than some technical issues) is the translation of costs from the infrastructure manager to the operators, because the train integrity is now guaranteed by the train itself.

ERTMS is a critical service for train safety, where trains periodically report their location, which the wayside equipment sends to train movement authorities, telling a train at each point how far it can go and how fast. This implies small-sized packets (most of the time less than 100 bytes), handovers to be completed in less than 300 ms, end-to-end delays lower than 500 ms, and connection establishment time below 8.5 s.

An industry trend for rolling stock is to assume more signaling functions at the expense of the wayside equipment, which means more importance for radio communications. As we see in the next subsection, this trend is shared by subway systems.

However, the capital expenditure (CAPEX) of an ERTMS system is quite high for some low-density traffic. Therefore, some alternatives to standard ERTMS may appear. The most relevant of them is the regional ERTMS, which divides the line into dark zones, where only one train can be inside each one. GSM-R coverage is punctual and does not require track circuits.

Carrying ERTMS data over a packet-switched network (instead of a circuit-switched one, like GSM-R) is another issue that has been discussed for a while [2]. This could overcome some of GSM-R's limitations [3], extending its life a little longer.

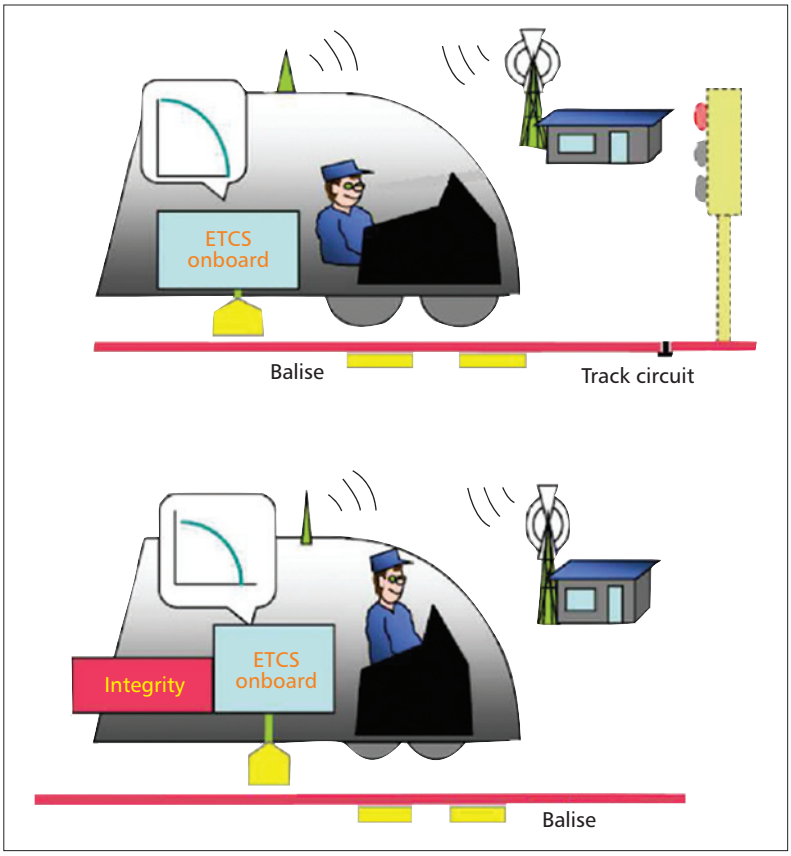


Figure 1. ERTMS level 2 (top) and level 3 (bottom). In level 3 there is no need for track circuits because train integrity is self-guaranteed by the train. Both levels 2 and 3 require GSM-R coverage and a network of balises.

Signaling for Subways: CBTC

Communications-based train control (CBTC) is the ERTMS counterpart for subway trains. It is also a very successful technology, but it is not as standardized as ERTMS. CBTC systems allow trains getting closer (below 80 s headway in some cases) and safer, so it has become a de facto standard for automatic lines, driverless trains, and almost every new line.

Every vendor follows its own implementation of the radio subsystem, but they are usually based on the IEEE 802.11 family of standards. Its technical requirements are very similar to those of ERTMS, and include end-to-end delay below 800 ms and short messages (64 bytes) exchanged more or less frequently (300–500 ms). Thus, both CBTC and ERTMS demand low transmission speeds but need very reliable radio systems.

CBTC [4] technology is well established and enjoys good health, but it can be improved. A reasonable way to reduce headway (i.e., get trains closer) is to downsize delays in the whole "command" chain (processing, transmission, processing again, reaction time, etc.), but sometimes this is hard to achieve. However, the solution may be in the communications channel. If we introduce direct communications between trains (without the intervention of the interlocking or any wayside equipment), the end-to-end delay is drastically reduced.

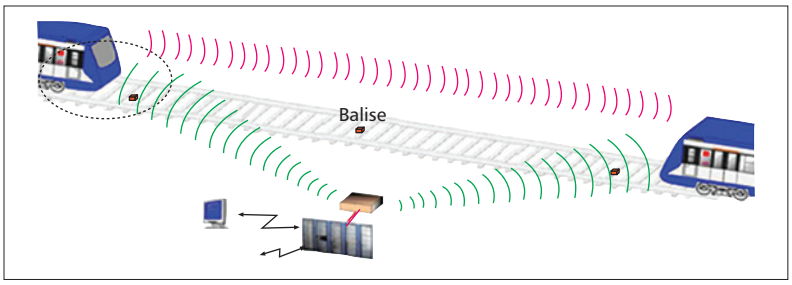


Figure 2. Depiction of a CBTC system with V2V capabilities.

This philosophy implies having a reliable onboard device-to-device (D2D) radio system, able to carry hard real-time information between two trains (V2V) in addition to the already available vehicle-to-infrastructure (V2I) communication (Fig. 2). This direct communication could be implemented now by using a public safety system like TETRA, or in the near future with LTE Release 13, for which direct communication capabilities will be defined. These V2V communications may be difficult to achieve if trains have obstacles between them, but if we realize that this V2V communication is meant only to get trains even closer, it makes sense. Anyway, this solution is not trouble-free, because it needs a great deal of effort in systems engineering before it comes into reality. Alstom, among other signaling vendors, is checking the feasibility of integrating this technology (called Alstom Urbalis Fluence) into its CBTC systems, but it has not released any results yet.

Tramways: Integration on Smart Cars Platforms

The tramway is the type of railway least likely to use automation systems for operational service because it shares its way with other vehicles, like cars, buses, and even pedestrians (not only other trains). Hence, they run on sight like a car or a bus, and they can suffer accidents as those do. In the near future, smart cars and smart highways will be very common, and in this kind of service, where vehicles send information to each other to avoid crashes or improve traffic efficiency, trams could be a major actor.

Some technologies like native V2V communication standard (IEEE 802.11p [5]) or, looking at lower levels of the stack, GeoNetworking protocol (meant for ad hoc routing) should allow trams to integrate into smart car platforms, thereby improving both safety and efficiency. This is a mid-term challenge for the industry, because even the automotive part of the system is not that mature. Some of these smart car services are time-critical (typically RT-VBR), where delays should be bounded (up to 50 ms) and have low bandwidth demands.

Signaling Data over Satellites

ERTMS makes use of GSM-R and a network of beacons, but for those lines where the traffic density is not worth such an expensive deployment, there could be other alternatives, like satellites. Satellites could be used for both locating trains (aided by some GNSS systems, like GPS or Galileo in the future) and communicating with the wayside equipment. Today, two major projects are trying to validate satellite technology for railway safety services. The first one is 3InSAT, led by Ansaldo STS and the European Space Agency, now being tested in Sardinia, intended to remove track circuits, beacons, and GSM-R infrastructure using satellites for both location and communication. It also incorporates a machine-to-machine (M2M) communication system to replace the satellite signal when it is not available. The second project is SATLOC, which is very similar to 3InSAT, but only uses satellites to locate trains, whereas the communications subsystem is based on 2G/3G. The real challenge here is the setup of a safety SIL4 service over a satellite system, and also some security issues have to be addressed. The LOCOPROL/LOCOLOC project has some interesting results in this field [6].

The Far Future of Freight Trains: Virtual Coupling

A true rail safety service for the far future to be implemented over a radio communications system is virtual coupling [7]. Freight trains could circulate separated by a distance even shorter than the braking distance, because the convoy of virtually coupled trains is linked by a ultra-reliable hard real-time communications radio, and each one of them shares the same data (speed, braking commands, etc.). Thus, they behave as one single train, but the coupling between them is only virtual. Of course, it is far from becoming reality and is only a concept, but in the far future it could help jammed freight lines achieve higher capacities, because the virtual train would only occupy one slot. In some ways, this is a generalization of the V2V communication for safety introduced above.

Public Safety in Railways

Public safety (PS) communications systems are widely used all over the world for law enforcement, emergency medical services, border security, environment protection, fire fighting, search and rescue, and railways. In railways they are mostly used for operational communication between the train and the control center. They are considered a wide sense safety service because their failure does not represent a problem for the safe movement of the train, but some operators decide to interrupt service if this system is not available.

The set of functionalities required for a PS communications system basically includes the extra requirements for GSM that incorporate GSM-R (plus, obviously, the voice service). They are the following:

- Direct communication between devices (D2D)
- Group call
- Push-to-talk (fast call initiation)
- Priorization of calls, including preemption
- Data, mostly messaging
- Location-based services, like functional addressing and location addressing

A high level of reliability and security (both authentication and integrity should be guaranteed) is also needed.

However, the two most relevant PS systems in railways (TETRA and GSM-R) have the same problem: their low capacity. TETRA is mostly used in subway environments, and GSM-R in mainlines and high-speed lines. Due to many factors (among them the support of TETRA Association) LTE is betting heavily on being the next PS communications system, but today Third Generation Partnership Project (3GPP) LTE hardly incorporates any public safety functionality. In its Releases 12 and 13 it will do so (Table 1), but this is a road paved with many difficulties, because all these PS functionalities are very challenging.

Among all these challenges, the most important of them are the following [8]:

- D2D: Node discovery, routing, radio resource management and security. It allows location-based services, and communication between nodes without a supporting infrastructure.
- Push-to-talk: PS communications need to be agile, so connection establishment time has to be very short (below 500 ms).
- Spectrum: PS channels are going to be unused most of the time, so their spectrum should be almost entirely shared, with perhaps a small part dedicated. Spectrum allocation, sharing, and management are also important issues, as always.
- Security: The degree of standardization of 3GPP LTE for PS security is very low as of 2014. This is a very relevant point because a failure here could break down the complete system.

Functionality	3GPP LTE Release	Scheduled date
D2D	12	March 2015
Group call	13	March 2016
Priorization calls	Already available	—
Security	12-13	March 2016
Push-to-talk	13	March 2016
Resilient EUTRAN	13	March 2016

Table 1. Public safety functionalities and the 3GPP Releases in which they are to be included.

Operational Services

In this section, we discuss the challenges and opportunities related to operational non-safety services (typically, SIL0).

Passenger Information/Infotainment

This family of services consists of providing passengers multimedia content related to the location of the train, its stops, and so on. It is a classic service, and its evolution goes toward integration with the signaling system, because it has the best knowledge of the train location. Location-based services (LBS) could provide some added value to this feature.

CCTV

The closed circuit TV (CCTV) system is another classic operational service in railways. Due to the nature of the data (video), it is very demanding in terms of bandwidth (1 Mpx camera implies more than 1 Mb/s), and if it has a real-time basis (e.g., video streaming from the control center) it also requires bounded delays and jitter (125 ms and 25 ms, respectively). Another usual service is on-demand recording download (sometimes stored onboard). In driverless or unattended subways, it is a key pillar of the operational process.

The Internet of Things

The Internet of Things (IoT) is one of the latest game changers in the IT industry. This paradigm implies sensing, communicating, and aggregating all the information to obtain knowledge. Railways are more ready than other sectors to embrace it for the following reasons: modern trains have sensors almost everywhere, train-to-wayside radios are now more popular than ever, control centers are a key part of every railway network, and railway operations need to know the location of every single train. All of these are reasons for the adoption of the IoT philosophy.

However, there are some challenges and difficulties: the onboard sensors do not follow an open architecture and usually do not get out beyond the train; the railway is a very hostile environment for all this hardware; security issues should be addressed properly; the massive scale of protocols, data volume, and architectures; and finally, the conservative nature of railways does not help a new paradigm that still has many unknowns to clarify to be embraced. Anyway, if we focus on aiding operation and maintenance, it is very likely that IoT will have a lot to offer for railways.

Services for Passengers: Internet Access

In this section, we discuss the challenges and opportunities related to services for passengers (SIL0); usually, the philosophy here is best effort. Providing reliable and competitive Internet access to onboard passengers is not a trivial task. It has been an open problem in recent years, and due to many difficulties, it is still open. Among these complications, we find the following:

- A hostile environment, with high temperatures, vibrations, electromagnetic interference, and limited access for maintenance
- Vehicle penetration loss (VPL), usually 15–25 dB, depending on the frequency and type of vehicle
- Cyber-security
- Development of an attractive business model that offsets the high CAPEX required
- Frequent handovers
- The presence of tunnels

There are several solutions with more or less maturity [9] and some (relative) success stories, like that implemented by Thalys on its high-speed trains. However, they lack the desired quality that could make this service a reference and a draw for customers (in both high-speed lines and subways). Among all current and near

Relays are used to improve cell coverage, backhaul, and to increase spectral efficiency. There are many types, subtypes and classification criteria, but all mobile relays have one thing in common: they are onboard. Devices of this kind have been standardized by 3GPP LTE, WiMAX, and many others.

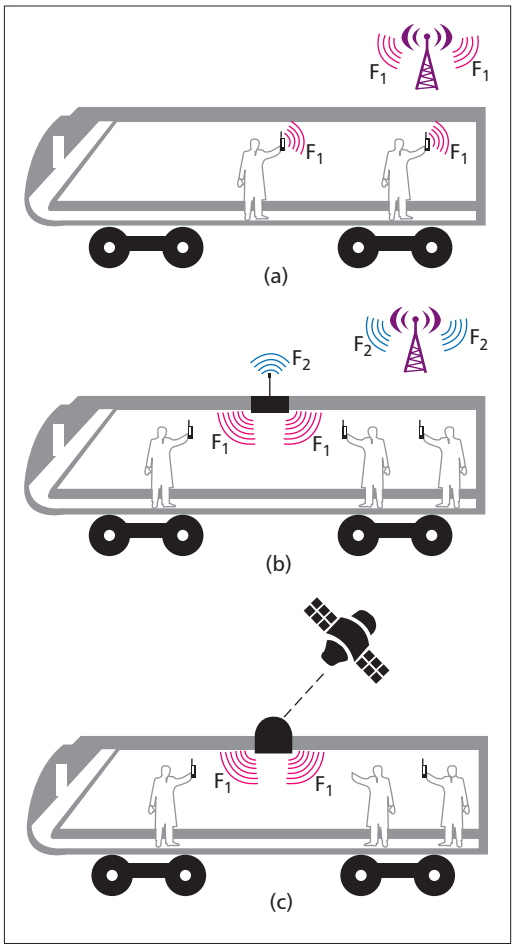


Figure 3. Methods for providing Internet access to on-board passengers: a) by a mobile relay and a cellular network; b) by a cellular network; and c) by a satellite system.

future solutions, two stand out from the rest (Fig. 3):

- Mobile relay (MR)-based [10]
- Satellite-based

Mobile Relays

Relays are used to improve cell coverage and backhaul, and increase spectral efficiency. There are many types, subtypes, and classification criteria, but all MRs have one thing in common: they are onboard. Devices of this kind have been standardized by 3GPP LTE (starting from Release 11), WiMAX, and many others.

Its function is to split the direct link between the onboard user equipment (UE) and the base stations (BSs) into two segments. Using an MR, there are several links between the UEs and the MR, but only one between the MR and the BS. The main advantages are:

- Placing the MR antenna on the roof of the train, VPL could be overcome (allowing the use of more efficient modulation schemes and therefore higher bit rates).
- Doppler, multipath, and other undesired effects could be abridged using digital signal processing (DSP) techniques in the MR, techniques that are too expensive to be implemented in cheaper UEs.
- Group handover decreases signaling traffic.
- A toehold for operators to develop a business model.
- Larger UE battery lives.

And the major drawbacks are the following:

- Hostile environment and difficult maintenance conditions
- Increases end-to-end latency
- "All or nothing" group handover (if it fails, affects everybody)
- Need to handle interference/scheduling issues with fixed base stations
- Integration with train's systems (TCMS, etc.)

Satellites

Satellites were the first solution put into practice (Thalys' case), and they have many advantages and limitations. The advantages are their simplicity (no need for a terrestrial network, so it is a good solution for railway operators that do not have an agreement with infrastructure administrators) and low CAPEX. The first and most important limitation is the need for a backup technology when trains are in tunnels, and the second is the high operational expenditure (OPEX) implied (other solutions like MR also have significant costs). To properly communicate with satellites, it is usually necessary to place huge antennas on trains' roofs, which has an impact on gauge and aerodynamics. Many research projects have explored the possibility of providing Internet access to onboard passengers [11], but most of them have failed, due to either the technical complexity or the lack of an adequate business plan.

Cyber Security

Increasingly, cyber security is a major issue on every communication service, not only in those related to providing Internet access. It is true that cyber attacks on railways have either not happened or not leaked out. The only exception (but very few details were given) is the Shenzhen Subway incident in 2012, where 3G service was shut off for a day after trains unexpectedly stopped. It is clear that security aspects cannot be ignored anymore. To investigate this type of issue, the European Project SECRET [12] arose, which intends to study all the electromagnetic risks and threats related to the railway environment. Finally, the coexistence of all these different types of services (safety-related, operational, and Internet access) is also a risk that should be addressed.

Technological Issues

Here we discuss some challenges that are either not related to any of the three types of services previously discussed or related to all of them.

Radio Convergence

A very common scenario in these services that need train-to-wayside communications is "one service, one radio." Thus, legacy radio telephony, TETRA or GSM-R, multipurpose radios based on IEEE 802.11 b/g, signaling radios for subway trains, and so on imply that in one single train there could be as many as four different

radio systems. This is due to the different nature of each one (narrow/wideband, analog/digital, critical/non-critical, trunking, IP, etc.) and the different timing for their commissioning.

However, this means very high CAPEX and OPEX, which could be avoided by using a single convergent radio, a system that aggregates all these traffic flows (Fig. 4), handling them with proper QoS and security policies. Today, the best candidate for such a convergence is LTE, above all if the public safety features explained earlier succeed in being incorporated in the 3GPP LTE standard. Moreover, betting on LTE implies IP convergence.

Another issue to take into account is the obsolescence of some of these systems (especially GSM-R, which has faced some serious limitations from its very beginning).

Withdrawal of Onboard Wiring

Years ago, some operators started transmitting some car-to-car data through an IEEE 802.11g dedicated link, avoiding mechanical coupler connectors and increasing the available bandwidth. This was the kick-off for the wire-removal trend in the rolling stock scenario. On a modern train we have hundreds of meters with all kinds of wires (supply, train buses, car buses, RF cables, network, etc.) with their associated connectors. All these wires cause a lot of breakdowns and huge setup and maintenance costs, so their replacement by wireless links would be a significant improvement for railway operators. This is the idea behind one of the packages of the European R&D Project Roll2Rail, which is scheduled to be launched this year.

High-Speed Scenarios

The train speed record is 574.8 km/h, and it was achieved by an Alstom train in France in 2007. However, high-speed trains usually have velocity peaks of 300–350 km/h, and Maglevs rarely run faster than 430 km/h (at least in passenger service). At such fast speeds, the Doppler effect and spreading are much more demanding than at pedestrian or car speeds, causing multipath and spreading.

Another problem could be the estimation of the channel, because if reference signals take samples at a period higher than the coherence time of the channel, system performance may decay. This channel's coherence time is inversely proportional to the Doppler shift, which depends on both carrier frequency and vehicle speed. For example, when a train travels at 300 km/h and the carrier frequency is 2.4 GHz, the Doppler shift is 1.3 kHz, and the coherence time is 317 μs. Thus, to let reference signals follow channel variations properly, we should consider pilot patterns that estimate the channel with a period no longer than 317 μs.

Channel Modelling

Having an accurate channel model is key to deploying efficient communication systems. However, in V2I and V2V scenarios this is far from being taken for granted, because there are many aspects that have to be addressed. In a recent research work, Bo *et al.* [13] identify some of the pending issues and associated difficulties. Some examples are:

- In high-speed rail, the variety of scenarios (cuttings, viaducts, etc.) reaches 12 types and 18 subtypes, and most of them still have to be measured.
- In tunnels, the breakpoint that separates the near zone and far zone has not been properly calculated yet. Hrovat *et al.* [14] provide an exhaustive survey of channel modeling in tunnels.
- Train vehicle influence has to be incorporated properly to models, including VPL impact.
- In the V2V field, there is still more work to do: non-stationary channel modeling, models that consider elevation angles, and so forth.

All these pending issues face the same difficulties to succeed: the high cost of carrying out intensive tests in railway environment.

Parameter	Satellite	Mobile relay
High-speed performance	Very good	Good
Coverage	Medium	Good
Security	Good	Good
Maturity	High	Low
Bitrate	Medium–low	High
Delay	Very high	Low
QoS support	No	Good
IP	No	Yes
Cost	High OPEX Low CAPEX	Low OPEX Medium–low CAPEX

Table 2. Comparison between mobile relays and communication satellites to provide Internet access to onboard customers.

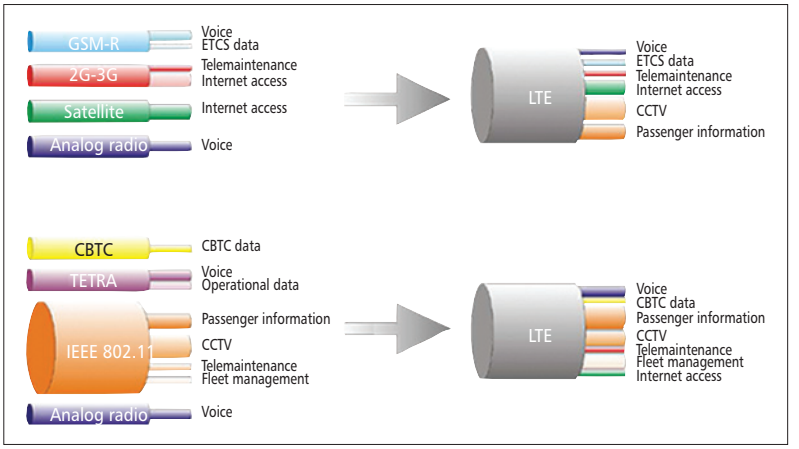


Figure 4. Non-convergent and convergent radio for (top) high-speed railways and (bottom) subways.

We leave open a question: Is LTE the next trend in mobile communications for railways? Some aspects are favorable to this idea (3GPP LTE group has put the focus on railways and also on public safety; LTE has received some attention from the railway environment, etc.) but today it is still a promising technology, not a reality.

Conclusion

In this article we have summarized all the major challenges and opportunities related to radio communications that railways will meet in both the near and far future. As most usually depicted, railway services are divided into three main categories: safety-related, operational, and passenger-centric. We have outlined the most important requirements for each of them, as well as the related functionalities and the challenges behind them. Also, we take a look at other aspects like security, channel modeling, and radio convergence.

Finally, we leave open a question: Is LTE the next trend in mobile communications for railways? Some aspects are favorable to this idea (the 3GPP LTE group has put the focus on railways and also on public safety; LTE has received some attention from the railway environment, etc.) but today it is still a promising technology, not a reality (few metro lines around the world have an LTE train-to-ground radio). Of course, other options are possible, like the future 5G standard, a cognitive radio system, or even a technologically neutral one. A recent ERA report [15] provides insight on the current situation and future opportunities for operational services. Also, the recently launched European project Roll2Rail will provide some outputs in this direction.

So the big question of LTE and railways is still open. However, it is crystal clear that the future of railways will use a lot of radio-based services.

In Memory of Leandro de Haro

Leandro de Haro passed away on July 30, 2015. The three remaining co-authors of this article wish to express our gratitude to Leandro for his wonderful contributions, his research work in antennas and communications, and also for being a fantastic person.

References

[1] IEC 61508:2010 "Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems" (parts 1–7), 2010.

[2] S. F. Ruesche, J. Steuer, and K. Jobmann, "The European Switch: A Packet-Switched Approach to A Train Control System," *IEEE Vehic. Tech. Mag.*, vol. 3, no. 3, Sept. 2008, pp. 37-46.

[3] A. Sniady, J. Soler, "An Overview of GSM-R Technology and Its Shortcomings," *2012 12th Int'l. Conf. ITS Telecommun.*, Taipei, 5–8 Nov. 2012, pp. 626–29.

[4] IEEE Std 1474.1-2004, "Communications-Based Train Control (CBTC) Performance and Functional Requirements," 2004.

[5] IEEE Std 802.11p-2010, "Information technology — Local and Metropolitan Area Networks— Specific Requirements — Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 6: Wireless Access in Vehicular Environments," 2010.

[6] E. Lemaire, F. Lesne, and M. Bayart, "Using Dataflow Traceability Between Functions in the Safety Evaluation Process," *IMACS Multiconf. Computational Engineering in Systems Applications*, Beijing, China, 4–6 Oct. 2006, pp. 1095–1102.

[7] U. Bock and J. U. Varchmin, "Enhancement of the Occupancy of Railroads Using Virtually Coupled Train Formations," World Congress on Railway Research, 1999, Tokyo, Japan, 1999.

[8] G. Baldini *et al.*, "Survey of Wireless Communication Technologies for Public Safety," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 2, 2nd qtr. 2014, pp. 619–41.

[9] L. Chen *et al.*, "Mobile Relay in LTE-Advanced Systems," *IEEE Commun. Mag.*, vol. 51, issue 11, Nov. 2013, pp. 144–51.

[10] A. Papadogiannis *et al.*, "Pass It On: Advanced Relaying Concepts and Challenges for Networks Beyond 4G," *IEEE Vehic. Tech. Mag.*, vol. 9, issue 2, June 2014, pp. 29–37.

[11] D. T. Fokum and V. S. Frost, "A Survey on Methods for Broadband Internet Access on Trains," *IEEE Commun. Surveys & Tutorials*, vol. 12, no. 2, 2nd qtr. 2010, pp. 171–85.

[12] M. Heddebaut *et al.*, "Towards a Resilient Railway Communication Network against Electromagnetic Attacks," Transport Research Arena 2014, Paris, France, Apr. 2014.

[13] Bo Ai *et al.*, "Challenges toward Wireless Communications for High-Speed Railway," *IEEE Trans. Intelligent Transportation Sys.*, vol. 15, issue 5, Oct. 2014, pp. 2143–58.

[14] A. Hrovat, G. Kandus, and T. Javornik, "A Survey of Radio Propagation Modeling for Tunnels," *IEEE Commun. Surveys & Tutorials*, vol. 16, no 2, 2nd qtr. 2014, pp. 658–69.

[15] F. Pujol and J. S. Marcus, "Final Report for the European Railway Agency, Evolution of GSM-R," Ref: ERA/2014/04/ERTMS/OP, Apr. 2015.

Biographies

Juan Moreno García-Loygorri (juan.moreno@metromadrid.es) received his M.Sc. degree in telecommunication engineering from Universidad Carlos III de Madrid in 2006. He is also a Ph.D. candidate at Universidad Politécnica de Madrid (UPM), Spain, where he has recently completed his Ph.D. thesis, scheduled to be defended in November 2015. Since 2007, he has worked in the railway sector, first at High Speed Railways and, since October 2008 at Metro de Madrid, where he currently works in the Rolling Stock Engineering Department. He participates in many R&D projects like Roll2Rail and Tecrail. His research interests include mobile communications, telecommunication systems in railways, propagation in tunnels, and MIMO.

José Manuel Riera [M'91, SM'13] (jm.riera@upm.es) received his M.S. and Ph.D. degrees in telecommunication engineering from UPM in 1987 and 1991, respectively. Since 1993 he has been an associate professor of radio communications at UPM. His research interests are in the areas of radiowave propagation and wireless communication systems. He has directed more than 40 research projects in these fields, funded by private companies, public agencies, and national and international research programmes, including UE, ESA, and COST. He is the author of more than 130 technical papers, 100 of them published in international journals and conference proceedings or as book chapters.

Leandro de Haro (leandro@gr.ssr.upm.es) received his ingeniero de telecomunicación degree in 1986 and his Doctor Ingeniero de Telecomunicación degree (Apto cum laude) in 1992, both from E.T.S.I. Telecomunicación, Departamento de Señales, Sistemas y Radiocomunicaciones, UPM. From 1990 on, he developed his professional career in the Departamento de Señales, Sistemas y Radiocomunicaciones as professor titular de universidad in the signal theory and communications area. In 2012 he became a full professor in the Departamento de Teoría de la Señal y Comunicaciones of UPM. His research activity covered the following topics: antenna design for satellite communications (Earth stations and satellite onboard); study and design of satellite communication systems; and study and design of digital TV communication systems. He was actively involved in several official projects and with private companies (national and international). He was also involved in several European projects (RACE, ACTS, COST). The results of his research activity may be found in several presentations in national and international conferences as well as in published papers. Prof. Leandro de Haro passed away on July 30, 2015.

Carlos Rodríguez Sánchez [M'06] (carlos@metromadrid.es) received his B.Sc. degree in industrial engineering (electronics and microprocessors) from the Universidad Pontificia de Comillas, Madrid, Spain, his M.Sc. degree in industrial engineering (electronics and automation) from UPM, and his Ph.D. degree in economics from the University of Alcala, Spain. He received a Ph.D. degree in electrical, electronics and system engineering from the Spanish University for Distance Education (UNED), Madrid. His industry experience includes several positions as a rolling stock and trackside engineer in the railway industry. He is currently head of the Engineering Department, Metro de Madrid S. A.. His research interests are related to railway signaling and safety in software development within industrial environments.

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

Channel Sounding for High-Speed Railway Communication Systems

Tao Zhou, Cheng Tao, Sana Salous, Liu Liu, and Zhenhui Tan

ABSTRACT

High-speed railway (HSR) communications have recently attracted much attention due to some specific railway needs, such as passenger experience services, business process support services, and operational data and voice services. The HSR radio channel, as a basis for the design of HSR communication systems and the evaluation of HSR communication technologies, has not yet been sufficiently investigated. This article focuses on radio channel sounding techniques for future HSR communication systems. Emphasis is placed on the fundamental features of measuring the HSR channel, and a review of the state of the art in HSR channel measurements is given. We also propose a novel LTE-based HSR channel sounding scheme and present results from our own measurement campaigns.

INTRODUCTION

The rapid development of high-speed railway (HSR) transport has led to a growth in demand for broadband wireless communication services from both train operators and passengers. In Europe and China, train operators employ the Global System for Mobile Communications for railway (GSM-R), a narrowband communication system, to maintain reliable train-to-ground communication. To guarantee train operational safety, in addition to standard train control services involving dispatching, shunting, and maintenance, broadband services such as onboard video surveillance and track monitoring are becoming requirements, and it is necessary to enhance the comfort and satisfaction of passengers who expect high-quality in-journey services such as mobile services, online TV, and Internet access. Considering rail operational demands and passengers' communications needs, the next-generation HSR communication system [1] and wireless network architecture [2] need to be developed to provide a variety of services for safety and infotainment.

The potential key communication technologies in an HSR system involve cell combination, mobile relay [3], multiple-input multiple-output (MIMO), and coordinated multipoint (CoMP) transmission [4]. Typical features of these techniques are the following.

Cell combination: The high mobility of the receiver yields frequent handovers; for example, when a train is running at 300 km/h and intersite distance along the rail is 1.2 km, handovers occur every 7 s. Such frequent handovers lead to a large amount of signaling or even to a signaling storm. Cell combination of several base stations (BSs) into a big virtual cell reduces the handover frequency. However, in this architecture the receiver could suffer from interference among neighboring BSs.

Mobile relay: Mobile relay is employed for multihop coverage from BS to mobile relay station (MRS), and to user equipment or other onboard devices, such as cameras, TVs, and access points. The mobile relay station, which can support one or many technology types, connects to the access network using antennas mounted on the outside of the train car. In addition, in-train networks can be organized to support various communication services. The mobile relay is expected to avoid penetration loss from the train body, but the train roof could cause distorted antenna radiation, affecting the received power.

MIMO: MIMO is an important technology that exploits the spatial domain of mobile fading to enhance capacity. Due to the special features of the HSR channel, which is dominated by the line of sight (LOS) component, HSR scenarios are not favorable for the application of MIMO technology. However, there is sufficient space on top of the train for distributed MRS antennas, which can be deployed at appropriate spacing to improve MIMO performance.

CoMP: CoMP is an emerging innovative technology for improving transmission efficiency. It transforms the interference from cooperative BSs into useful signals by coordinated transmission and reception. Applying CoMP in HSR communication systems mitigates the interference caused by cell combination and exploits the diversity gain to enhance capacity. The BSs in the combined cell should be connected to a backhaul network, which is used to exchange the information involving data, control information, and channel state information.

Since the radio channel determines the performance of wireless communication systems, detailed knowledge and accurate characterization of its parameters in realistic HSR propagation

Tao Zhou, Liu Liu, and Zhenhui Tan are with Beijing Jiaotong University.

Cheng Tao is with Beijing Jiaotong University and Southeast University.

Sana Salous is with Durham University.

0163-6804/15/$25.00 © 2015 IEEE

scenarios is crucial. The majority of radio channel models used for system simulation are based on extensive channel measurements. Therefore, channel sounding is a precondition for the design of HSR communication systems and the evaluation of HSR communication technologies. Although measurement campaigns on HSR are expensive, time-consuming, and difficult to perform, a few HSR channel measurements have been reported using radio channel sounders [5–7] or railway-network-based measurements [8, 9]. As future HSR communication systems will apply MIMO and CoMP, channel measurements should consider the correlation properties of the individual links and multiple links that significantly affect the performance of MIMO and CoMP. In addition, the impact of using cell combination and mobile relay needs to be evaluated using channel measurements in real-world HSR networks. As of now, to the best of our knowledge, there is no appropriate HSR channel sounding method that addresses these issues.

Motivated by this observation, this article gives an overview of HSR channel sounding techniques and describes a novel technique based on Long Term Evolution (LTE) that meets the requirements and challenges of HSR sounding and measures the HSR channel comprehensively, reliably, and efficiently. We also present results from our latest field measurement results.

HSR Channel Sounding

In radio channel measurements, a known signal that repeats at a rate twice the highest expected Doppler shift is transmitted, and the received signal is analyzed to evaluate the effects of the channel on its transmission. In channel sounding we observe how many echoes of the signal are received, and their amplitude and phase variations along the travel route. These are measured using a channel sounder that detects the electromagnetic wave transmitted via a particular communication channel to determine the statistics of either the channel impulse response (CIR) or channel frequency response (CFR). Classical channel sounding techniques can be loosely classified into two categories: narrowband, using a continuous waveform (CW), or wideband. Narrowband sounding only provides signal fading characteristics but does not provide information regarding the multipath components. These issues can be resolved using wideband techniques, which involve periodic pulse sounding, and pulse compression waveforms, which avoid the need for high peak transmitted power and provide processing gain. The two commonly used techniques are either coded sequence transmission or chirp sounding [10]. Another technique uses an orthogonal frequency-division multiplexing (OFDM) signal to probe the channel [11].

Requirements in HSR Channel Measurements

The main channel measurement requirements are determined by the time delay resolution, which is inversely proportional to the transmitted bandwidth and the maximum expected Doppler shift, which determines the waveform repetition frequency (WRF) [10]. The time delay resolution indicates the smallest difference in time delay between resolvable multipath components. In [1], the bandwidth of future HSR communication systems is recommended to be higher than 10 MHz. This sets the time delay resolution to 100 ns, corresponding to a distinguishable 30 m difference in propagation distance. The WRF, which should be at least twice the maximum expected Doppler shift, depends on the carrier frequency and vehicular speed. In [1], it is suggested that for HSR dedicated communications, high-priority service be in the 800 MHz frequency band and low-priority service in the 1.8 GHz frequency band. At 1.8 GHz, the maximum expected Doppler shift can be accommodated within 600 Hz for a maximum train speed of 360 km/h, and thus the WRF should exceed 1.2 kHz. The WRF also determines the overall time delay window, which is related to spread of multipath and the farthest distance that can be measured. The time delay window, however, has to be in excess of the extent of multipath to allow for travel distances away from the transmitter. For instance, when the multipath components are assumed to extend over 1 μs and the distance of two BSs in HSR communication systems is 1.2 km, which corresponds to 4 μs, the time delay window has to be in excess of 5 μs to ensure that the movement of the multipath components is still within the observable window.

As for MIMO measurements, these requirements are essentially the same with the added requirement of multiple transmissions and multiple receptions. MIMO channel sounding can be achieved in three possible architectures. The most popular one is a full sequential architecture using time-division multiplexing mode with switching between antennas at the transmitter and the receiver. To enable Doppler measurement, the scanning of all antennas should be completed within the coherent time of the channel. For the 8×8 HSR MIMO channel system, the required WRF should be as high as 76.8 kHz. In contrast, a full parallel architecture employing a number of orthogonal techniques, such as code-division multiplexing and frequency-division multiplexing, transmits and receives simultaneously from all antennas, which only requires the same WRF as single-antenna systems. An alternative to these two architectures is a semi-sequential architecture where the transmitter is switched between the different antennas with a number of parallel receiver channels [10]. For the example of eight antennas, the needed WRF corresponds to 9.6 kHz.

Challenges in HSR Channel Measurements

While current channel sounding techniques are quite mature, and extensive measurement campaigns have been reported for terrestrial cellular communication systems, there are still some challenges in HSR channel measurements that should be addressed.

The first challenge is how to employ the commonly used channel sounding techniques in HSR environments. Generally, to guarantee the secu-

While current channel sounding techniques have been quite mature and extensive measurement campaigns have been reported for terrestrial cellular communication systems, there are still some challenges in HSR channel measurements that should be addressed.

LTE supporting both frequency division duplexing and time division duplexing modes adopts OFDM and MIMO technologies. Typical deployed carrier frequencies are in the range of 400 MHz to 4 GHz, with scalable carrier bandwidths from 1.4 MHz to 20 MHz.

rity of a train control network, the wireless frequency bands are under surveillance by the railway administration. In addition to the existing network signals, such as GSM-R, wideband code-division multiple access (WCDMA), and LTE, any other wireless test signals are prohibited from transmission along the rail. Since all conventional channel sounding techniques need to transmit a particular waveform, it is difficult to directly apply these methods in HSR channel measurements.

Another challenge is how to measure MIMO and CoMP channels. The requirement of WRF in HSR scenarios makes HSR MIMO channel measurement very challenging. Most of the MIMO channel sounders with the full sequential architecture cannot meet such a high requirement. Although the WRF can be decreased by reducing the number of transmit or receive antennas, this would degrade the angular measurement capability. To enable the CoMP channel measurement, there is a kind of static multi-link measurement where single-link measurements are used to form a static multi-link scenario [12]. However, such an approach is not valid if parts of the channel change between measurements at different locations. Thus, multi-link channel measurements should be carefully designed and conducted using advanced channel sounding equipment and techniques that can record the observations of two or more channels simultaneously.

The final important issue to address is the measurement efficiency, which is extremely low when the commonly used channel sounders are employed. For instance, if the train travels at a speed of 360 km/h and the coverage radius of a channel sounder is 1 km, the recording time of the sounder is approximately 20 s. The obtained experimental data in this short period may not be adequate to extract the statistical properties of the HSR channel. In addition, one HSR line could cover a wide variety of scenarios, such as urban, suburban, rural, hilly, as well as some special scenarios like cutting, viaduct, tunnel, and station. Measuring all these scenarios is meaningful to establish a complete and accurate HSR channel model. However, it is extremely challenging when using universal channel sounders with low measurement efficiency to achieve this task.

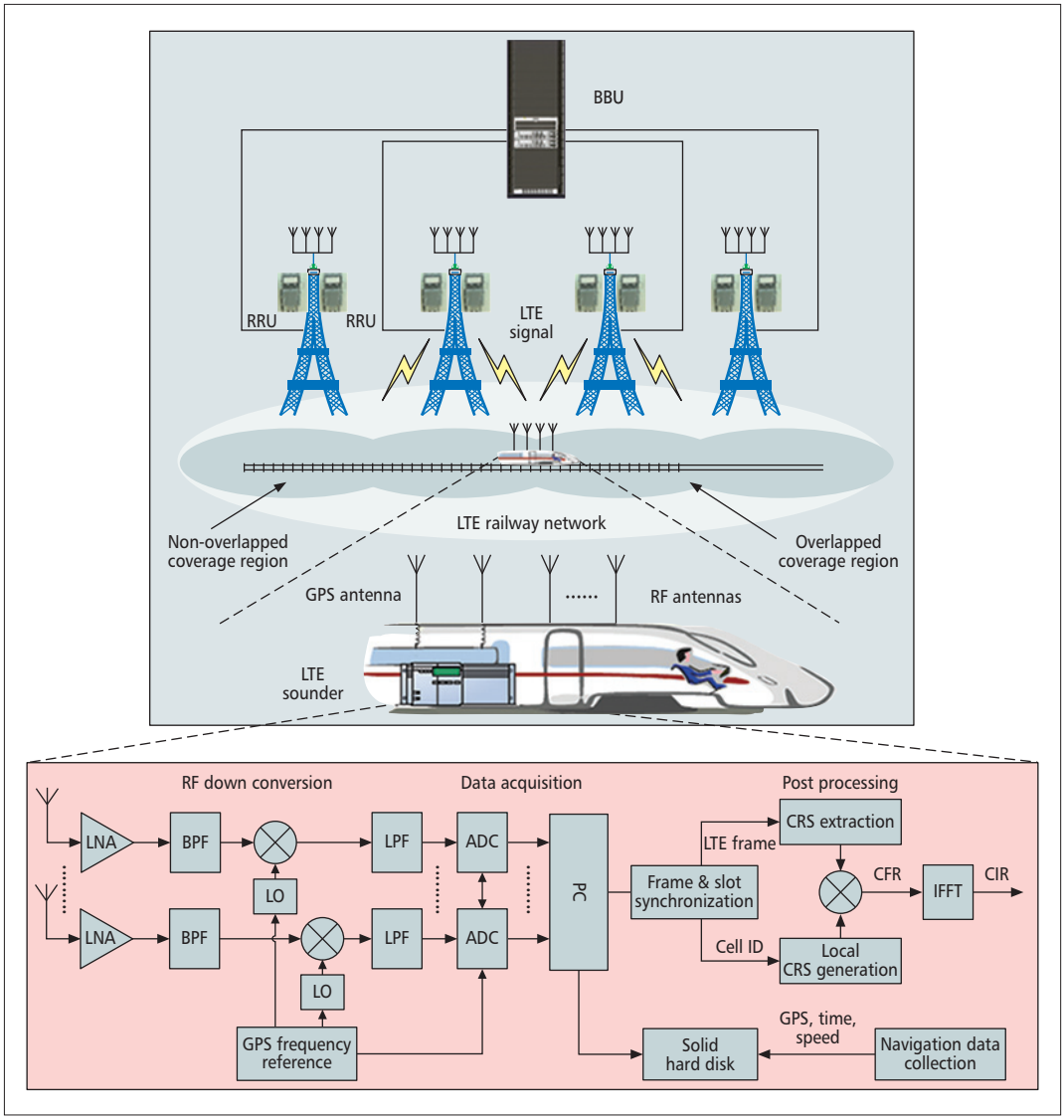


Figure 1. LTE-based HSR channel sounding scheme.

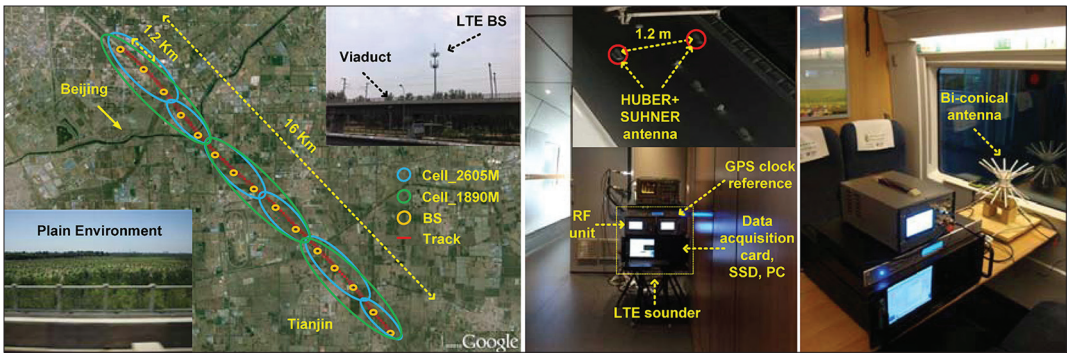


Figure 2. LTE-based HSR channel measurements in DC and RC scenarios.

Recent Advances in HSR Channel Measurements

In this section, we briefly review some recent typical HSR radio channel measurement campaigns, which can be simply classified into two categories.

Channel-sounder-based measurements: As of now, there are several measurement campaigns taken under high mobility conditions, using standard commercial multidimensional channel sounders with full sequential architectures. One of the first reported HSR measurements employed the RUSK sounder to measure the single-input multiple-output mobile relay channel in Germany [5]. The used multi-carrier spread spectrum signal is a kind of OFDM signal, which concentrates the energy in the band of interest. HSR channel measurements in rural and hilly scenarios in Taiwan used Propsound, where the direct link from the BS to the user equipment (UE) was considered to measure the single-input multiple-output and multiple-input single-output channel [6]. Furthermore, single-input single-output (SISO) mobile relay channel measurements using Propsound were carried out in viaduct scenarios on HSR in China [7]. However, there are still no reported MIMO channel measurements in a specific HSR scenario utilizing standard commercial channel sounders.

Railway-network-based measurements: Due to the restrictions imposed by using traditional channel sounders on HSR, since 2011 some researchers have resorted to the railway-network-based channel sounding method. The basic idea of the method is to exploit the signal transmitted from the railway network to enable continuous measurements along the whole rail. Obvious advantages of this method are the high measurement efficiency and low measurement limitation. A series of GSM-R channel measurements were conducted in viaduct scenarios on HSR in China [8]. For channel characterization purposes, the GSM-R signal is regarded as a narrowband CW signal and hence unsuitable for wideband measurements. To enable wideband channel characterization, the common pilot channel signal in the dedicated WCDMA network was collected and analyzed to extract the multipath properties [9]. The common pilot channel signal is related to a cell-specific scrambling Gold code. Unfortunately, the measurement bandwidth of this method does not meet the requirements of time delay resolution for the HSR channel, and it lacks the spatial sounding ability.

LTE-Based HSR Channel Sounding Scheme

Existing HSR channel measurement data are insufficient for the design of future HSR communication systems. Reported measurement campaigns do not provide a comprehensive and efficient HSR channel sounding approach. Inspired by [8, 9], we have investigated a novel HSR channel sounding scheme which employs LTE railway networks to implement the excitation of the time-frequency-space HSR channel [13], as shown in Fig. 1. In the following we briefly describe this scheme from three aspects involving the LTE railway network, LTE signal, and LTE sounder.

LTE Railway Network

LTE aims to support a wide variety of scenarios involving indoor, urban, suburban, and rural areas covering both low and high mobility conditions. In recent years, there have been large-scale deployments of LTE networks all over the world. It is reported that the LTE railway network covered all HSRs in China, for a total of 15,000 km by 2014, and as HSRs continue to grow, LTE railway networks that grow with them will exceed 30,000 km in 2020. The railway network is totally different from a conventional cellular network, adopting a narrow strip coverage mode and a building baseband unit (BBU) plus remote radio unit (RRU) structure. One BS site has two RRUs applying directional antennas to transmit RF signals in opposite directions along the track. Cell combination is employed in this network where the RRUs in one cell are connected together via optical fiber and then to a BBU that is in charge of RF signal processing. The entire network can be classified into two categories: a non-overlapped coverage region and an overlapped coverage region. In the non-overlapped coverage area, only one signal is received from one BS. In the overlapped area, the two same signals from neighboring BSs could arrive at the receiver simultaneously, but they can be distinguished in the delay domain due to the different propagation delays.

LTE Signal

LTE supporting both frequency-division duplexing and time-division duplexing modes adopts OFDM and MIMO technologies. Typical

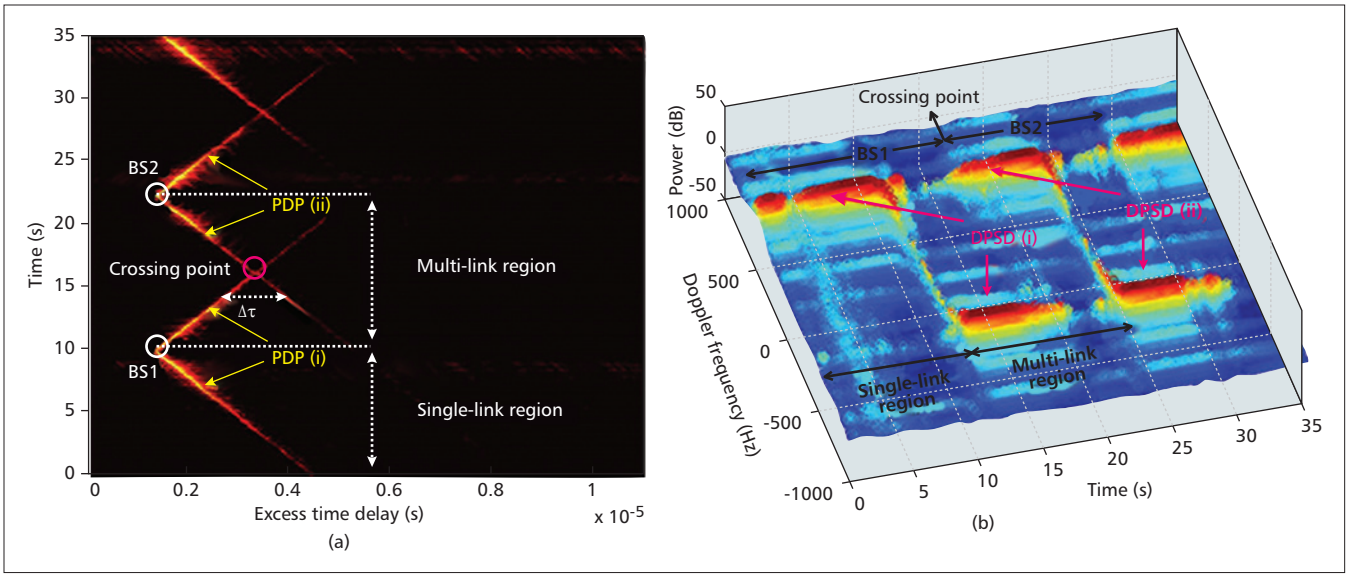


Figure 3. Results of cell combination measurements in the RC scenario: a) time-variant PDP; b) time-variant DPSD.

deployed carrier frequencies are in the range of 400 MHz to 4 GHz, with scalable carrier bandwidths from 1.4 to 20 MHz. In [14], it was shown that cell-specific reference signals (CRSs) play a key role in LTE-based channel sounding, and completely determine the measurement capability. The CRSs are embedded in the LTE time-frequency-space frame structure with a diamond shape. In the time direction, the maximum repetition period of the CRSs is 0.5 ms, corresponding to 2 kHz WRF. In the frequency direction, there is one CRS every six subcarriers with a spacing of 90 kHz on each OFDM symbol. This spacing allows a maximum time delay window of 11 μs. In the case of 20 MHz LTE, the total number of CRSs on each OFDM symbol is 200, which corresponds to 18 MHz measurement bandwidth and 56 ns time delay resolution. In the space direction, there are four ports carrying the orthogonal CRSs. This implies a full parallel architecture with a maximum of four transmit antennas can be used. Thus, the LTE signal meets the HSR channel sounding requirements outlined in the previous section.

LTE Sounder

The LTE sounder is used to collect the channel data in the whole LTE railway network. Different from conventional channel sounders, the LTE sounder only has a receiver. The LTE sounder with basic functions, such as RF down conversion and data acquisition, employs the same full parallel architecture as the transmitter. In particular, GPS frequency reference is used to enable the frequency consistency between the network and the LTE sounder. Also, baseband data and navigation data are stored together in a solid state disk (SSD) for post processing. Frame and slot synchronization should be implemented to acquire LTE frames and determine the cell identity (ID) for extracting the received CRS and generating the local CRS, respectively. Since the CRS is OFDM modulated, frequency domain correlation is used to estimate the CFR, which can subsequently be transformed to the CIR by inverse fast Fourier transform (IFFT) operation, as shown in the simulations presented in [14].

LTE-Based HSR Channel Sounding Results

Based on the LTE railway network from the Beijing to Tianjin HSR in China, we carried out LTE-based HSR channel measurements on a high-speed test train, as shown in Fig. 2. Two types of scenarios, direct coverage (DC) and relay coverage (RC), are considered in our measurements. We used two networks: the 2605 MHz network with 6 cells for SISO DC measurements, and the 1890 MHz network with 3 bigger cells for 2 × 2 MIMO RC measurements. The measured environment is a typical plain viaduct. The BS is less than 20 m away from the viaduct, and the BS antenna is generally 10, 20, or 35 m higher than the rail. Detailed measurement parameters are listed in Table 1. In the following we present the measurement results from the cell combination, mobile relay, MIMO, and CoMP perspectives.

Cell Combination Measurement Results

Figure 3 shows results of the cell combination measurement in the RC scenario, focusing on power delay profile (PDP) and Doppler power spectral density (DPSD). In Fig. 3a, the time-variant PDP during the first 35 s period in a certain cell is plotted. Two obvious PDP transitions regarding BS1 and BS2 with positions identified by the white circles are highlighted. When the train enters into the non-overlapping area, we define a single-link region where one propagation link between BS1 and the train, indicated by PDP (i), almost occupies the whole time delay window. However, once the train moves into the overlapping area between BS1 and BS2, another propagation link between BS2 and the train, denoted by PDP (ii), appears in the time delay window as well. We regard this area as a multi-link region where the time delay window covers

two links simultaneously. In the multi-link region, the time delay window can be divided into two parts, one for PDP (i) and another for PDP (ii). As the train moves away from BS1, the delay difference $\Delta\tau$ between PDP (i) and PDP (ii) is gradually shortening. If the train travels at the crossing point marked with a red circle in Fig. 3a, PDP (i) and PDP (ii) would be indistinguishable. Since $\Delta\tau$ determines the time delay window of PDP (i), a threshold for $\Delta\tau$ should be set to enable the coverage of most multipath components. Figure 3b plots the corresponding time-variant DPSD result. When the train passes through the coverage of BS1 and BS2, two typical Doppler transitions from the maximum positive frequency to the minimum negative frequency are observed and marked as DPSD (i) and DPSD (ii). In the single-link region only DPSD (i) exists, while in the multi-link area DPSD (i) and DPSD (ii) appear at the same time. The two DPSDs have the same maximum Doppler shift but opposite angles of arrival.

The results in Fig. 3 confirm that cell combination causes a sort of artificial multipath interference and Doppler interference. From the engineering perspective, to cope with such interference, the BSs are normally deployed with a reasonable spacing, which ensures that the maximum propagation delay between neighboring BSs is less than the period of the cyclic prefix (CP). For instance, the distance of two BSs should be no more than 1.41 km, corresponding to 4.69 μs CP length in the case of 20 MHz LTE. From the technical perspective, however, this interference can be turned into useful signals according to the CoMP technology.

Mobile Relay Measurement Results

Based on the measurement data in the single-link region, we highlight the RC measurement results, concentrating on path loss (PL) and root mean square (RMS) delay spread (DS). For comparison, we also show the DC measurement results. Figure 4 illustrates the ensemble PL results in all cells and PL models for the DC and RC cases. The PL in the DC scenario is found to be 20–30 dB higher than that in the RC scenario. This value corresponds to the penetration loss of the train car. On the other hand, the resulting PL exponents in the DC and RC scenarios are 3.16 and 3.76, respectively, which are all much higher than that in free space propagation. For the DC case it is understandable that the outdoor to indoor propagation condition has an impact on the PL exponent; however, for the RC case the main reason, which is explained in [5, 7], is that the train carriage roof would act as a ground plane and thus affect the antenna radiation pattern by causing a null in a certain incidence angle area of the radiation pattern. The cumulative distributed functions (CDFs) of the RMS DS are also computed, and the results for 50 and 90 percent are estimated as 101.2 and 184.3 ns for the DC scenario, and 27.6 and 92.6 ns for the RC scenario. Since the in-train propagation environment causes additional scattering and reflecting components, the RMS DS for the DC case is higher than that for the RC case.

According to these results, we conclude that

Item	Value	
Transmitter		
Measurement scenario	DC	RC
Measurement frequency	2605 MHz	1890 MHz
Measurement bandwidth	18 MHz	18 MHz
CRS power	12.2 dBm	12.2 dBm
Antenna type	±45° cross-polarized directional	±45° cross-polarized directional
Antenna gain	18.6 dBi	17.4 dBi
Horizontal beamwidth	60°	67°
Vertical beamwidth	4.9°	6.6°
Electric tilted angle	3°	3°
Receiver		
Antenna type	Bi-conical	HUBER+SUHNER
Antenna gain	0 dBi	8.5 dBi
Antenna number	1	2
Antenna spacing	/	1.2 m (7.6 wavelengths)
Train velocity	285 km/h	285 km/h

Table 1. Measurement parameters.

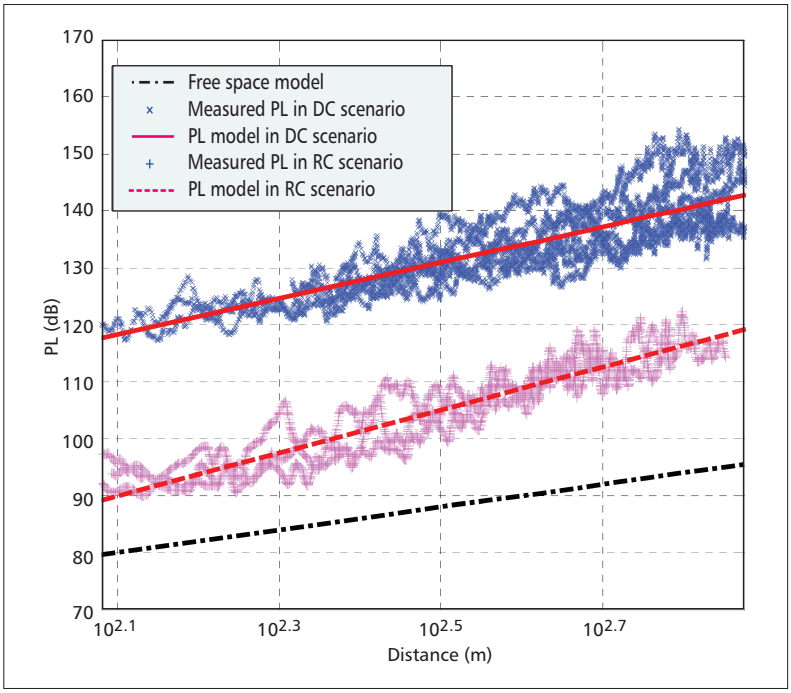


Figure 4. Results of path loss in the DC and RC scenarios.

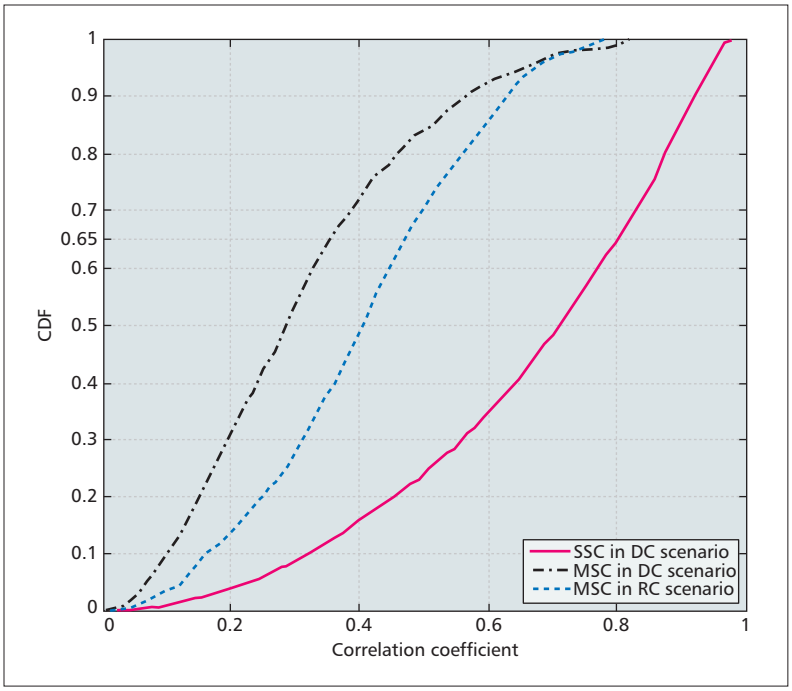


Figure 5. Results of the correlation coefficient in the DC and RC scenarios.

the mobile relay does not only enhance coverage but also avoids the impact of the in-train environment. However, the problem of using mobile relay is that the received power would attenuate faster due to the distorted antenna radiation pattern caused by the train carriage roof.

MIMO and CoMP Measurement Results

For MIMO and CoMP channel characterization, we mainly focus on the single-link spatial correlation (SSC), single-link channel capacity (CC), and multi-link spatial correlation (MSC). The MSC exists due to the environment similarity arising from common scatterers contributing to different links [15]. We denote the measured CIR matrix in the single-link region as H_{pq}^{S}, and the CIR matrices extracted from BS1 and BS2 in the multi-link region as $\mathrm{H}_{pq}^{M,1}$ and $\mathrm{H}_{pq}^{M,2}$, where p and q are the indices of the antenna elements at the transmitter and receiver, respectively. Based on the measurement data in the single-link region, the SSC between two sub-channels (e. g., H_{11}^{S} and H_{22}^{S}) are derived. Similarly, according to the effective measurement data (setting the threshold of $\Delta\tau$ as 1 μs) in the multi-link region, the MSC between $\mathrm{H}_{11}^{M,1}$ and $\mathrm{H}_{11}^{M,2}$ is estimated. Figure 5 illustrates the results of SSC in the RC scenario, and MSC in the DC and RC scenarios. It is observed that almost 65 percent of SSC values are less than 0.8. For the MSC, the results are optimistic in both DC and MRC scenarios where the majority of the correlation coefficient values are below 0.8. Moreover, since rich scatterers in the in-train environment lead to a low degree of environment similarity, the DC case has lower MSC than the RC case. The corresponding CC in the RC scenario is also calculated and approximates 9.9 b/s/Hz in the case of 20 dB SNR in comparison to 11.2 b/s/Hz for the independent and identically distributed complex Gaussian channel with zero mean and unit variance.

From the above observations, we can infer that the cross polarized antenna configuration at the BS side and the large antenna spacing at the train side are beneficial to improve MIMO performance in the LOS-dominant HSR scenario. In addition, CoMP will have good performance in terms of micro-diversity due to the small MSC.

Conclusions

In this article, we discuss channel sounding for HSR communication systems. After summarizing the requirements, challenges, and recent advances in HSR channel measurements, we describe a novel LTE-based channel sounding scheme. Furthermore, we show typical results from LTE-based measurement campaigns, concentrating on the evaluation of cell combination and mobile relay effects, and the analysis of MIMO and CoMP performance. These field test results confirm the viability of the proposed scheme, and provide useful information for the optimization of current HSR LTE systems and the design of future HSR communication systems.

Acknowledgment

The authors would like to thank Huisheng Wang and Yuzheng Zhang from China Academy of Railway Sciences for their help with performing the channel measurements, and Long Sun from Huawei for useful discussion of LTE railway network structure and configuration. The research was supported in part by the NSFC projects under grant No. 61371070, No. 61471030, Beijing Natural Science Foundation (4142041, 4152043), and the open research fund of National Mobile Communications Research Laboratory, Southeast University (No.2014D05).

References

[1] B. Ai *et al.*, "Challenges Toward Wireless Communications for High-Speed Railway," *IEEE Trans. Intelligent Transportation*, vol. 15, no. 5, Oct. 2014, pp. 2143–58.
[2] L. Yan, X. M. Fang, and Y. G. Fang, "Control and Data Signaling Decoupled Architecture for Railway Wireless Networks," *IEEE Wireless Commun.*, vol. 22, no. 1, Feb. 2015, pp. 103–11.
[3] D. Soldani and S. Dixit, "Wireless Relays for Broadband Access," *IEEE Commun. Mag.*, vol. 46, no. 3, Mar. 2008, pp. 58–66.
[4] L. Zhu *et al.*, "Design and Performance Enhancements in Communication-based Train Control (CBTC) Systems with Coordinated Multi-Point Transmission and Reception (CoMP)," *IEEE Trans. Intelligent Transportation*, vol. 15, no. 3, June 2014, pp. 1258–72.
[5] P. Kyösti *et al.*, "WINNER II Channel Models," IST-WINNER II D1.1.2, Nov. 2007.
[6] R. Parviainen, P. Kyösti, and Y. Hsieh, "Results of High Speed Train Channel Measurements," *Proc. COST 2100 TD(08) 646*, Oct. 2008, pp. 1–6.
[7] L. Liu *et al.*, "Position-Based Modeling for Wireless Channel on High-Speed Railway under a Viaduct at 2.35 GHz," *IEEE JSAC*, vol. 30, no. 4, May 2012, pp. 834–45.
[8] R. S. He *et al.*, "An Empirical Path Loss Model and Fading Analysis for High-Speed Railway Viaduct Scenarios," *IEEE Antennas Wireless Propagation Lett.*, vol. 10, Aug. 2011, pp. 808–12.
[9] J. H. Qiu *et al.*, "Broadband Channel Measurement for the High-Speed Railway Based on WCDMA," *Proc. IEEE VTC-Spring '12*, Yokohama, Japan, May 2012, pp. 1–5.
[10] S. Salous, *Radio Propagation Measurement and Channel Modelling*, Wiley, 2013.
[11] A. F. Molisch, *Wireless Communications*, 2nd ed., Wiley, 2011.

[12] J. S. Jiang, M. F. Demirkol, and M. A. Ingram, "Measured Capacities at 5.8 GHz of Indoor MIMO Systems with MIMO Interference," *Proc. IEEE VTC-Fall '03*, Orlando, FL, Oct. 2003, pp. 388–93.
[13] L. Liu *et al.*, "A Highly Efficient Channel Sounding Method Based on Cellular Communications for High-Speed Railway Scenarios," *EURASIP J. Wireless Commun. Net.*, vol. 2012, Article ID: 307, 2012
[14] T. Zhou *et al.*, "A Study on a LTE-Based Channel Sounding Scheme for High-Speed Railway Scenarios," *Proc. IEEE VTC-Fall '03*, Las Vegas, NV, Sept. 2013, pp. 1–5.
[15] X. Cheng *et al.*, "Cooperative MIMO Channel Modeling and Multi-Link Spatial Correlation Properties," *IEEE JSAC*, vol. 30, no. 2, Feb. 2012, pp. 388–96.

BIOGRAPHIES

TAO ZHOU (taozhou.china@gmail.com) received his B.E. degree in communication engineering from Changchun University of Science and Technology, China in 2009. He is currently working toward his Ph.D. degree in communication and information system within the Institute of Broadband Wireless Mobile Communications, School of Electronics and Information Engineering, Beijing Jiaotong University, China. His current research interests include high-speed railway communications, channel sounding technologies, and channel modeling for high-speed railway scenarios.

CHENG TAO received his M.S. degree in telecommunication and electronic systems from Xidian University, Xian, China, in 1989 and his Ph.D. degree in telecommunication and electronic system from Southeast University, Nanjing, China, in 1992. He is currently a professor and director of the Institute of Broadband Wireless Mobile Communications, Beijing Jiaotong University. His research interests include mobile communications, radio channel measurement and modeling, and signal processing for communications.

SANA SALOUS received her B.E.E. degree from the American University of Beirut, Lebanon, in 1978, and her M.Sc. and Ph.D. degrees from Birmingham University, United Kingdom, in 1979 and 1984, respectively. Since 2003 she holds the Chair in Communications Engineering and is director of the Centre for Communication Systems, Durham University, United Kingdom. Her research interests include channel characterization in various frequency bands and radio channel sounders and radar systems for radio imaging.

LIU LIU received his B.E. degree in communication engineering and Ph.D degree in communication and information system from Beijing Jiaotong University in 2004 and 2010, respectively. He is currently an associate professor at the Institute of Broadband Wireless Mobile Communications, School of Electronics and Information Engineering, Beijing Jiaotong University. His general research interests include channel measurement and modeling for different propagation environments, and signal processing of wireless communication in time-varying channel.

ZHENHUI TAN received his M.S. and Ph.D. degrees in communications and information systems from Beijing Jiaotong University and Southeast University, Nanjing, China, in 1982 and 1987, respectively. From 1995 to 2008 he was a president of Beijing Jiaotong University. He is the author of two books and more than 100 papers. His current research interests include digital mobile communications networks, spread spectrum communications, broadband wireless access, and adaptive filtering algorithms.

Future Railway Services-Oriented Mobile Communications Network

Bo Ai, Ke Guan, Markus Rupp, Thomas Kürner, Xiang Cheng, Xue-Feng Yin, Qi Wang, Guo-Yu Ma, Yan Li, Lei Xiong, and Jian-Wen Ding

ABSTRACT

The future development of the railway is highly desired to evolve into a new era where infrastructure, trains, travelers, and goods will be increasingly interconnected to provide high comfort, with optimized door-to-door mobility at higher safety. For this vision, it is required to realize seamless high data rate wireless connectivity for railways. To improve the safety and comfort of future railways, wireless communications for railways are required to evolve from only voice and traditional train control signaling services to various high data rate services including critical high-definition (HD) video and other more bandwidth-intensive passenger services, such as onboard and wayside HD video surveillance, onboard real-time high data rate services, train multimedia dispatching video streaming, railway mobile ticketing, and the Internet of Things for railways. Corresponding mobile communications network architecture under various railway scenarios including inter-car, intra-car, inside station, train-to-infrastructure and infrastructure-to-infrastructure are proposed in this article. Wireless coverage based on massive MIMO for railway stations and train cars is proposed to fulfill the requirement of high-data-rate and high spectrum efficiency. The technical challenges brought by the massive MIMO technique are discussed as well.

Bo Ai, Ke Guan, Qi Wang, Guo-Yu Ma, Yan Li, Lei Xiong and Jian-Wen Ding are with Beijing Jiaotong University.

Bo Ai (corresponding author) is also a visiting professor with the Department of Electrical Engineering, Stanford University.

Markus Rupp is with TU Wien.

Thomas Kürner is with Technical University Braunschweig.

Xiang Cheng is with Peking University.

Xue-Feng Yin is with TongJi University.

INTRODUCTION

High-speed railway (HSR), intercity railway, subway, light rail, and other rail traffic systems have brought much convenience for people's travel with less energy consumption and air pollution compared to cars and airplanes. Moreover, HSR moves very fast with high comfort and high punctuality. To ensure safe and reliable operation of railways, the train operation control system acts as a nerve center. To make such a nerve center work well, a reliable bidirectional communication link between the train and the ground is of great importance. Global System for Mobile Communications for Railway (GSM-R) plays a key role in realizing such bidirectional communications. Train timetable information, the driving license, train speed, train location, and other train control signals can be transmitted through a GSM-R network.

With the increasing demand for new railway services such as railway multimedia dispatching communication and railway emergency communication, Long Term Evolution for Railway (LTE-R) is now under discussion [1, 2]. Such a broadband communication system has the capability of 100 MHz data transmission rate in high mobility with 20 MHz bandwidth. However, just as those topics specified in the European "Shift2Rail" project (http://www.shift2rail.org), intelligent rail infrastructure, intelligent mobility management, smart rail services, and a new generation of rail vehicles ultimately form the requirements of a seamless high data rate wireless connectivity for future rail development. Thus, higher-frequency-band techniques such as millimeter-wave (mmWave), the fifth generation (5G), and the corresponding mobile communication network should be designed accordingly to provide high capacity and high data rate for future railway services. As far as the authors know, although plenty of literature has discussed GSM-R communication networks [3] and 5G techniques [4], no literature deals with the communications network regarding future railway services.

The new services for railways may pose special requirements for the new mobile communications network architecture. Future railway services and typical communications scenarios are described. The heterogeneous mobile network architecture for future railway system, related promising key technologies, and technical challenges are discussed. Massive multiple-input multiple-output (MIMO)-based wireless coverage for railway stations, inside cars, and other railway hotspot areas is discussed in detail. Finally, conclusions are drawn.

FUTURE RAILWAY SERVICES AND BANDWIDTH REQUIREMENT

With numerous HSRs being operated and planned in China and Europe in particular, reliable communications for railway control and safety are of great current and future interest. In addition, high efficiency, environment friendli-

0163-6804/15/$25.00 © 2015 IEEE

ness, and passenger convenience are also goals of future railway developments. In 2010, the E-train project (http://www.uic.org/etf/publication/publication-detail.php?code_pub=190_14) from the International Union of Railways (UIC) summarized over 200 railway services including train dispatching, train control, train operation communication, train state monitoring, and so on. In 2014, Horizon 2020 [5] emphasized that there should be more railway services focused on realizing the objective of "smart, green and integrated transport." To sum up, even though the LTE-R network is designed to bear many railway services, real-time HD video transmissions supporting automatic driving, security closed circuit television (CCTV) in the train, remote maintenance of trains, and other high data rate railway services still challenge the current mobile communications network for railways.

Services for the Future Railway

In order to realize the above-mentioned vision, communications for future railways are required to evolve from only critical signaling applications to various high data rate services: onboard and wayside HD video surveillance, onboard real-time high data rate services, train multimedia dispatching video streams, railway mobile ticketing, and the Internet of Things for railways.

Onboard and wayside HD video surveillance: Video surveillance services capture live 720p or 1080p true HD video images from high-definition television (HDTV) IP cameras and high-definition serial digital interface (HD-SDI) cameras located on trains (for train operation) and the wayside along rail tracks (for asset condition monitoring). In light of security concerns (terrorist attacks, riots, and emergencies), the HD video images are required to be both stored locally and delivered in real time from trains/wayside monitors to local and central train control centers (TCCs).

Onboard real-time high data rate services: One of the most attractive future railway services is the wireless Internet inside train cars. Not only surfing the wireless Internet, having a pleasant chat through Facebook or Twitter social networks [6], passengers onboard also expect to access real-time HD video for business and entertainment such as on-vehicle video conference and live broadcast.

Train multimedia dispatching video stream: Comprehensive dispatching information including written words, data, voices, and images is provided by a multimedia dispatching command communication system to the dispatcher. For instance, regarding automatic driving, an intensive train multimedia dispatching video stream of doorways for driverless trains is required by dispatchers from remote TCCs to ensure that the train doorways are clear before a train sets off.

Railway mobile ticketing: To ensure the security and reliability of the passenger ticketing system, a dedicated railway communication network will be used for the ticketing system, and a handheld ticketing terminal, through which identity identification, check-in, and ticket selling will be realized. This will alleviate the crowdedness of the ticket hall and ease passenger traveling.

Internet of Things for railways: In addition to real-time query and tracking the whole process of the location of the train and goods, the Internet of Things for railways can be developed to integrate the sensing information of rail infrastructures including bridges, viaducts, tunnels, leaky feeders, rail gaps, frozen soil, and slope protection through various sensing measures such as infrared, sound sensors, and temperature sensors. The information is collected and sent back to the computing center. On-time forecasting and management decisions can be made through big data or cloud computing platforms to ensure safe operation of the train.

Bandwidth Requirement of the Future Railway

The determination of the bandwidth depends mainly on the services and the number of users. The services consuming most of the bandwidth should be analyzed. For an analysis of the number of users, the rules of simultaneous communication users within certain scenarios and different service types should be followed.

Taking the real-time HD video broadband connections inside a car in future railway services into consideration, data rates of up to 3 Gb/s (allowing 1920 × 1080 @ 60 Hz @ 24 bits) in a 40 MHz channel can be supported by the wireless home display interface (WHDI). With around 130–180 passengers/seats (in recent high-speed and intercity trains and some double-deck trains), up to 3.6 GHz total bandwidth is required for one car if 50 percent of the passengers want to have real-time HD video service for business or entertainment. If bidirectional HD video streaming is expected (for video conference), the bandwidth requirement may be double (up to 7.2 GHz required). Therefore, an LTE-R system with 20 MHz bandwidth cannot support such services, and we have to resort to the mmWave frequency bands and the 5G communication system for railway (5G-R), which provide large bandwidth and high data rate transmission capability. The METIS project group regards high mobility scenarios as typical of 5G. The 5G promotion group of IMT-2020 in China released a white paper on 5G vision and requirements, in which the railway and subway are defined as two typical scenarios for 5G. The subway represents the 5G typical character of super high density with over 6 persons/m^2, and the high-speed railway represents the typical character of high mobility with speeds up to 350 km/h or above.

In general, the requirement of future railway services and typical communication scenarios calls for large bandwidth and high data rate transmission capabilities.

Mobile Communications Network Architecture

To represent the above service space, five communication scenarios can be defined: train-to-infrastructure, inter-car, intra-car, inside station, and infrastructure-to-infrastructure. Although technologies for the train-to-infrastructure sce-

The determination of the bandwidth depends mainly on the services and the number of users. The services consuming most of the bandwidth should be analyzed. For an analysis of the number of the users, the rules of simultaneous communication users within certain scenarios and different service types should be followed.

Considering a variety of railway services and scenarios, the corresponding mobile communications network architecture should be heterogeneous including various types of access networks working at different frequency bands, satisfying multi-band, multi-scenarios and various requirements of wireless coverage.

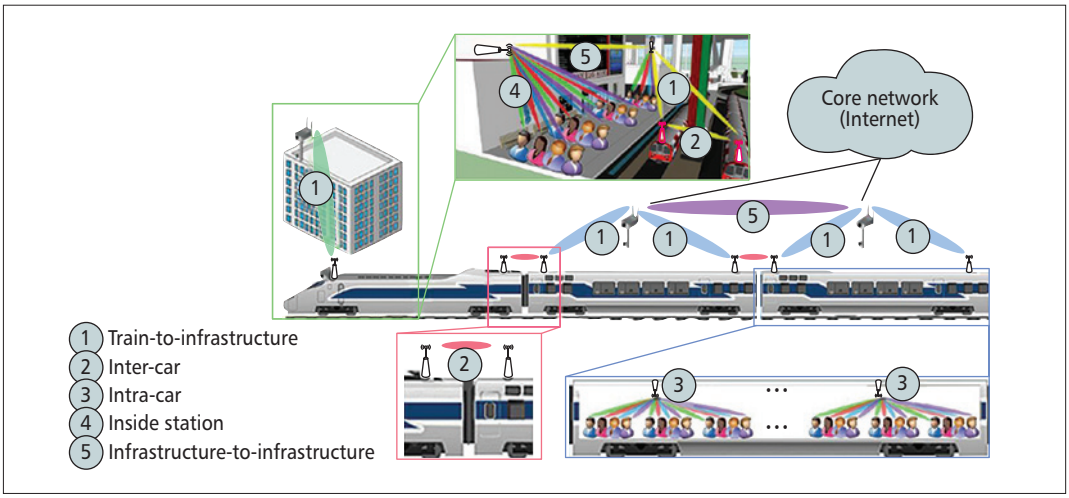


Figure 1. Panorama of seamless high data rate wireless connectivity for railways.

nario have been investigated, very few technologies have been examined for the inter-car and intra-car scenarios. This makes it very difficult to build a seamless network supporting all five communication scenarios for railways with the current technologies.

Railway Communication Scenarios

The five communication scenarios for future railways are described in detail as follows.

Train-to-infrastructure: Two kinds of links are required between the access points (APs)/transceivers of the train and the infrastructures of fixed networks. The links provide bidirectional streams with high data rates and low latencies, as well as robust communication links with latencies lower than 100 ms together with an availability of 98–99 percent, while moving at speeds up to 350 km/h or above.

Inter-car: A wireless network runs between cars to avoid the high expense of wiring a train for network access and the inconvenience of rewiring when a train is reconfigured. This scenario requires a high data rate and low latency because the APs are arranged in each car such that each AP serves as a client station for the AP in the previous car, while also serving as an AP for all the stations within its car.

Intra-car: The links provide wireless access between the APs in the car and the passengers or sensors of equipment inside the car. In this scenario, real-time HD videos need to be accessed with low latencies.

Inside the station: The links provide wireless access between the APs and the user equipment in railway stations. Users are strongly interested in access to mobile broadband communication services. The stations provide a fixed/wireless communication infrastructure to support general commercial (e.g., cash desks) as well as operational services (e.g., automatic doors, surveillance, fire protection).

Infrastructure-to-infrastructure: HD video and other information is transmitted in real time among multiple HDTV IP/HD-SDI cameras, and the APs deployed on the trains, on station platforms, and the wayside along rail tracks, as a high date rate wireless backhaul or the Internet of Things. Infrastructures are real-time connected and interactive, supported by bidirectional data streams with very high data rate and low latencies.

Mobile Communications Network Architecture

Considering a variety of railway services and scenarios, the corresponding mobile communications network architecture should be heterogeneous including various types of access networks working at different frequency bands, satisfying multiple bands, multiple scenarios, and various requirements of wireless coverage. Apart from satellite communications, which can provide medium- and high-capacity services for railways and will become more important in the future, the architecture of a heterogeneous land mobile communication network for future railway systems is shown in Fig. 2. Such a network will be composed of a macro base station used for macrocell coverage, and some micro base stations as backhaul to offer hotspot area coverage. Transferring from one network mode to another requires much time to adapt even with a self-organizing network (SON) [1]. Therefore, how to realize *seamless switching between different network modes with low latencies* is really a challenging task.

The dedicated mobile communication networks, including GSM-R, LTE-R, and 5G-R, and the public mobile communication network in such heterogeneous architecture are represented as ① and ② in Fig. 2, respectively. The dedicated networks cover the railway lines, railway station, freight station, marshaling station, railway hub, and other railway areas for train operation control purposes. The communication performance should satisfy the reliability, availability, maintenance, and safety (RAMS) from a UIC standard or specification. Moreover, HSR operates at extremely high speeds, introducing high Doppler spreads and frequent handovers. Thus, how to plan the *distance between two adjacent base stations efficiently* is another challenging task. It should guarantee both enough time for handover under high mobility and good cell edge coverage to avoid call drops.

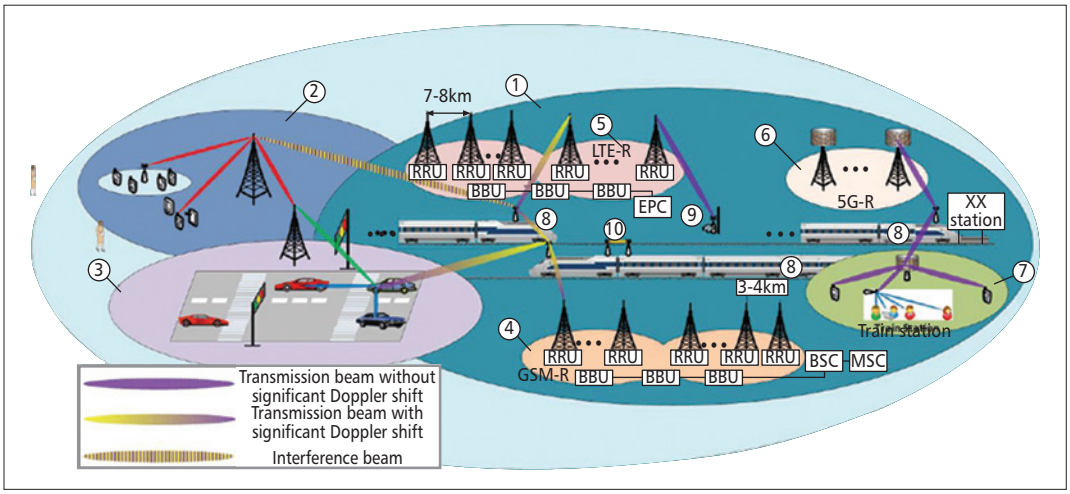


Figure 2. Heterogeneous mobile communication network for future railway systems.

One of the challenging tasks for 5G-R are propagation characteristic and channel models for massive MIMO dynamic channels under various railway scenarios at 6 GHz, 28 GHz, 38 GHz, 60 GHz and 300 GHz frequency bands.

Adjacent base stations of a GSM-R network, denoted by ④ in Fig. 2, are 3–4 km apart to guarantee good cell edge coverage. The traffic data is transmitted to a base station controller (BSC) from a mobile service switching center (MSC) of the core network and finally to the radio remote unit (RRU) from the building baseband unit (BBU). The GSM-R system is nearly mature. However, *the channel models for GSM-R under various railway scenarios are not perfect. Moreover, GSM-R should be compatible and interconnected with an LTE-R network*. LTE-R, denoted by ⑤, will probably adopt 450 or 800 MHz frequency bands in China. The distance between two adjacent base stations may be 7–8 km for 450 MHz and 3–4 km for 800 MHz to ensure both enough handover time and good quality of cell edge coverage. LTE-R uses the flat network structure based on IP. The traffic data is transmitted to a BBU from the Evolved Packet Core (EPC). EPC is the architecture for the convergence of voice in an LTE network. The BBU then transmits data to the RRU. The unsolved issues for LTE-R are *propagation characteristics and channel models at 450 and 800 MHz, the key techniques adaptive to high mobility*.

A 5G-R network, represented by ⑥ in Fig. 2, is developed to cater for large bandwidth and high data rate services. Due to its characteristic high frequency bands and reduced coverage capability, 5G-R is not appropriate for railway line coverage, but for railway hotspot areas such as railway stations and intra-car. As for the implementation of 5G-R, the best candidate technique is massive MIMO, which is regarded as one of the key techniques for 5G. However, the majority of the current research work on massive MIMO has focused on static channel conditions. One of the challenging tasks for 5G-R is propagation characteristic and channel models for *massive MIMO dynamic channels under various railway scenarios at 6, 28, 38, 60, and 300 GHz frequency bands*.

The railway station scenario denoted by ⑦ in Fig. 2 can be equipped with massive MIMO to provide high capacity and high data rate transmission for passengers inside the station. There can be one macro base station with several micro base stations installed in the corners inside the railway station, forming a wireless backhaul. The key technique that remains to be solved is *the propagation characteristics and channel models for massive-MIMO-based coverage under the railway station scenario*. The main duty of the public mobile communications network is recreation such as wireless Internet access for passengers. A challenging task for the coexistence of public and dedicated mobile communication networks is *how to avoid the serious adjacent channel interference*. In the future, the railway, subway, and other rail systems may be combined with road traffic to form an integrated transportation network, where not only vehicle-to-vehicle (V2V), but also train-to-train (T2T), train-to-infrastructure (T2I), and train-to-vehicle (T2V) communications will be included. T2V communication is useful when the train passes through the intersection of a rail track and a road. T2T communication, represented by ⑧, means the direct communication between two trains without infrastructure and any other APs. The most challenging tasks for integrated transportation are *channel characteristics and channel models under the T2T scenario, the categorizing of T2T transmitted messages, the media access control (MAC) layer routing mechanism, and the key techniques for long-range communications with high mobility*. ⑨ represents the ground-to-ground communications of the infrastructures along the rail tracks. The infrastructure along the rail tracks refers to the access points or webcams.

Massive-MIMO-Based Wireless Coverage and Technical Challenges

In the railway station and inside the car, where the transmission rate and system capacity need to be improved, massive MIMO can be used, which has been proved theoretically to achieve high data rate, high spectral efficiency, and high energy efficiency [7]. These gains come from its diversity and beamforming. Transmission and receiving modes can be adjusted to be adaptable

to intensive users simultaneously inside the car with flexible grouping of hundreds of antennas. Figure 3 shows the diagram of wireless coverage for railways based on massive MIMO techniques.

Figure 3 denotes the signal access mode into the car. The blue line denotes the signal at frequency bands below 6 GHz used for wide area coverage. However, 7.2 GHz bandwidth may be demanded for HD video services. MmWave or sub-mmWave bands such as 28 or 300 GHz offers orders of magnitude greater spectrum than current wireless allocations. They enable high-dimensional antenna arrays for further gains via beamforming and spatial multiplexing. Moreover, the antenna size can be dramatically reduced compared to that of the low frequency bands. The red lines denote mmWave or sub-mmWave.

There are mainly three access modes for the coverage inside the cars. Just as depicted in Fig. 3 (1-1), the signal from the base station penetrates directly into the car with penetration loss up to 24 dB, posing much higher requirements for the transmission power of the base station and the receiver sensitivity. When we use other access modes, straight penetration of the signal is unavoidable, which may cause interference to useful signals. It is a challenge to figure out *how to shield the signal directly into the car when using other access modes*. Vehicle repeaters can be adopted for high mobility environments, shown as (1-2), the signals received from the on-vehicle transceiver are modulated and forwarded to the micro base or WiFi signal repeater. However, the typical unresolved problem is that with a repeater, an AP inside the car may receive a weak signal through the train body and a strong one from the repeater, but with considerable delay. *How to design transmission schemes for good reception of a repeater at high moving speeds* is a challenging task. Figure 3 (1-3) represents a two-hop access mode. One hop is from the base station to the antennas on the top of the train, and another is from antennas on top of the train to receiver antennas inside the car. This may avoid large penetration losses introduced by the train body. However, as we know, higher frequency bands have large attenuation and path

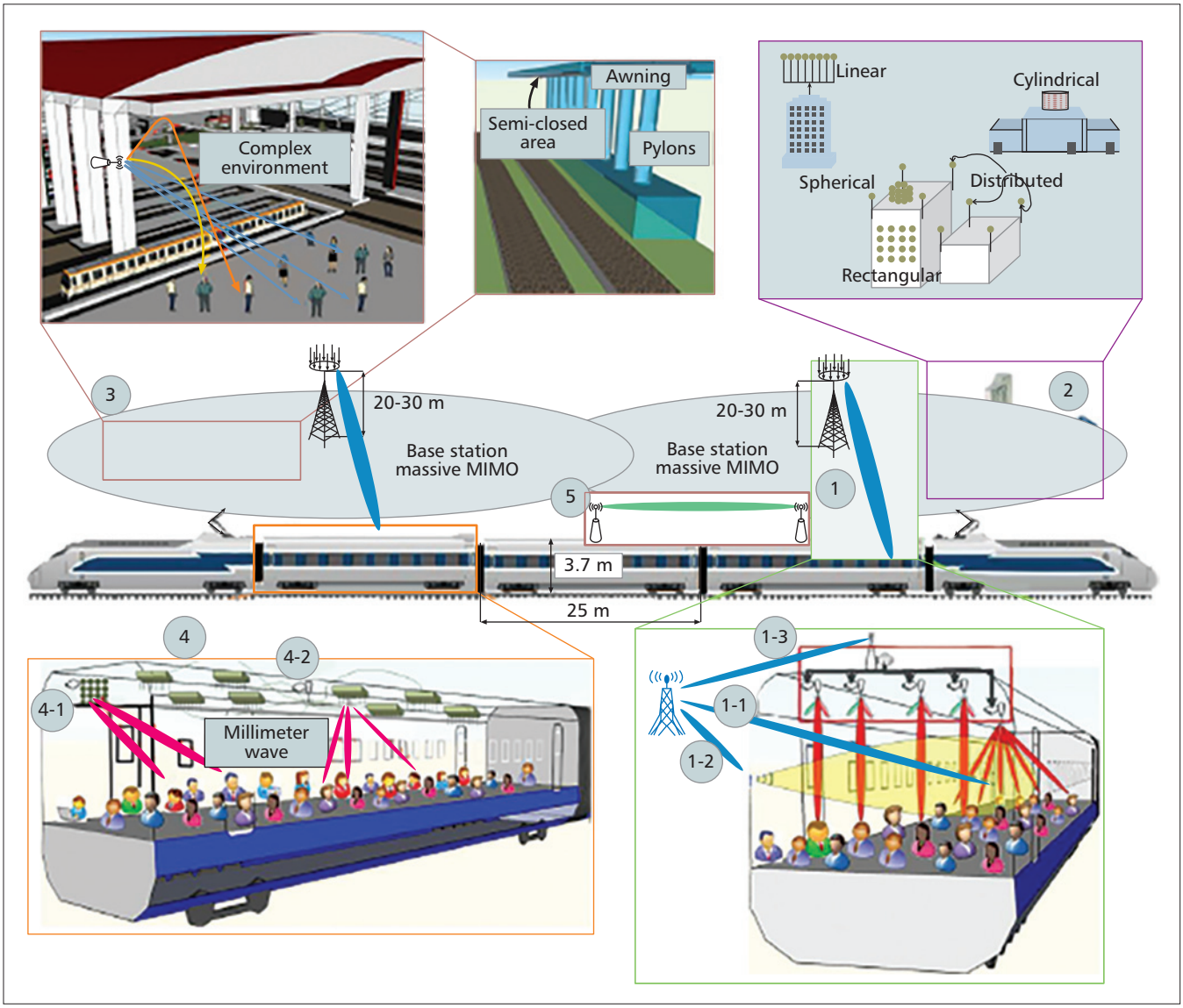


Figure 3. Massive-MIMO-based wireless coverage for railway systems.

loss, resulting in very limited transmission range. Therefore, the communication between the base stations to the train (i.e., train-to-ground communication) may use the working frequency bands below 6 GHz. By using the multihop technique at different frequency bands [8], higher frequency bands can be used inside the car for large bandwidth wireless coverage.

Massive MIMO used in a dense urban area is denoted by ② in Fig. 3. There may be many users under this scenario, where massive MIMO can be utilized to enhance the system capacity. Since it is inconvenient to set up base stations with large-scale antennas, we can deploy an alternative form of massive MIMO antennas at the base station side utilizing nearby high buildings. The shape of the antennas can be diverse such as linear array, cylindrical array, spherical array, plane array, and distributed antenna arrays. The task here is that *the channel models and receiver design should be appropriate for various shapes of antennas arrays*.

Figure 3 represents the scenario in the railway station with the characteristics of the semi-closed scene, dense crowd, and complicated environment. The deployment of the antennas is restricted by the scenes; for example, the linear array will probably not be equipped under such a scenario. Therefore, *how to design the appropriate antenna array types, including the irregular antenna array* is an interesting task.

Figure 3 represents the scenario inside the car. It needs to provide high data rate services for many users simultaneously. The car is a closed scene of rectangular shape, likely to cause multiple backscattering, resulting in large attenuations and losses, especially at mmWave or sub-mmWave frequency bands. A 3D beamforming technique can be utilized to overcome these shortcomings. As for the detailed deployment of a massive MIMO antenna array, the fundamental infrastructure such as the goods and luggage racks inside the cars and the user distributions should be considered as well, because they may have effects on the shadow fading. The conventional shadowing loss caused by the users for indoor mmWve coverage can come to 4–5.5 dB. Because the passengers are generally uniformly distributed, it is the best choice to deploy antennas with low power in a distributed form, equipped in the center of the top of the car, vertical to the train running direction. This may either be applicable to the passenger distribution or avoid the obstruction to the signals to a great limit. In this way, the diversity gain can be fully exploited. *How to deploy the antennas array including the antenna numbers, the shapes, and the pitch angles efficiently to satisfy various scenarios and requirements* is a research direction for massive MIMO.

In fact, the antenna can be either centrally configured in a base station to form a centralized antenna array, as depicted by (4-1) in Fig. 3, or configured among multiple nodes to form the distributed antenna array, as depicted by (4-2) in Fig. 3. The centralized antenna array provides larger array gain and more spatial degree of freedom to suppress intra- and inter-cell interference. The centralized antenna array may be equipped on the top of the train, forming strong propagation beams, to track the departure and arrival of the train with the centralized power. The installation of the centralized antenna can use the directional mode, similar to the splayed shape, to reduce the effect of Doppler shift. The research on *centralized massive MIMO applications at high mobility* is a most challenging task. The correlation coefficient of massive MIMO is large, especially for centralized antennas. The correlation and coupling effects may degrade its performance dramatically. One research direction is analysis of *the correlation and coupling mechanisms for massive MIMO*.

For distributed massive MIMO, there are many antennas sufficiently close within a very limited area, and strong beamforming can be achieved with the coordination of different antennas. Because of the cuboid shape of the train car and very dense crowds, the advantages of the distributed antenna can be fully released compared to the centralized antenna. Reasonable deployment of the distributed antennas inside the car should be determined according to specific conditions of the cars and the passenger distribution. Note that distributed massive MIMO is different from a distributed antenna system (DAS). There is only one antenna at the remote access unit (RAU) for DAS, but tens to hundreds of antennas for distributed massive MIMO. In the latter case, one user terminal can see all the distributed antennas simultaneously with beamforming, while there is no beamforming effect for DAS.

Inter-car communication is denoted by ⑤ in Fig. 3. Optical fibers and wireless communications can both be adopted for inter-car communications. However, optical fibers are not recommended to connect communication nodes because it may be expensive to wire a train for network access, and rewiring may be needed every time the train is reconfigured. Currently, the main possible wireless connection forms for inter-cars are WiFi, WiMAX, and dedicated short-range communications (DSRC).

Optical fibers are not recommended to connect communication nodes because it may be expensive to wire a train for network access, and rewiring may be needed every time the train is reconfigured. Currently, the main possible wireless connection forms for inter-cars are WiFi, WiMAX, and dedicated short-range communications.

Channel Modeling and System-Level Modeling

Whether for GSM-R, LTE-R, 5G-R, or mmWave applications, the major prerequisite condition is a thorough knowledge of the propagation characteristics of the wireless channel. Wireless channel modeling is the important basis and essential means for communication network planning and optimization, transmitter and receiver design, and physical and upper layer key techniques selection. Recently proposed new 3-D channel models such as the TR36.873 model [11] are a step into the right direction. The new model allows the physically correct modeling of channels by clustering scatterers, supporting 3D; to realistically model line-of-sight (LOS) and non-LOS (NLOS) depending on distance; and to include antenna arrays even of large dimensions. Nevertheless, particular aspects of train connections are not yet supported or still need to be sufficiently parameterized. As for massive MIMO channel measurement, most of the literature now focuses on the virtual antenna array

With the implementation of massive MIMO techniques and higher working frequency bands into the railway systems, even higher transmission rates and more reliable wireless transmissions can be realized at extremely high speeds, aided by careful designs that preserve the high quality of user experience.

measurement in static condition, *the fast channel sounding techniques and dynamic channel parameters extraction for massive MIMO* are indeed very challenging works.

Moreover, a very difficult problem remains for system-level modeling, in which abstraction models are being derived in mathematical form to offer simulation results in acceptable time or even allow for explicit analytical solutions. For such methods to work, it is important to validate the abstraction steps by smaller transmission units that are compared to more detailed link level descriptions. Once the models agree, the system-level models can easily be scaled to hundreds of users, higher bandwidth and longer transmit durations with high likelihood to correctly resemble such behavior [10]. The unresolved issues are handovers as high-speed trains contain many users (an inter-city-express, ICE, carries typically more than 500) that need to be handed over to the next station in a very short time. Currently, no accurate channel models are set up for fast handovers of so many users. Classical hybrid automatic repeat request (HARQ) methods [11] may not work as expected due to high speeds. Theoretically, sufficient (quantized) feedback information is required in order to guarantee high data rate transmissions [12]. However, at high speeds such channel state information is quickly outdated, and once the speed exceeds 100 km/h there is no gain in feedback information. While users are being moved from one antenna port to the next, they typically observe a U-shaped attenuation profile in between, in which the received power drops to a minimum once the center between two antenna ports is crossed. Smart scheduling can compensate for such effects offering a high data rate on average per user [13].

Above all, we should develop the appropriate communication network architectures, channel and system-level models, and different layer key technologies such as massive MIMO to meet the needs of high data rate, high spectrum, and high energy efficiency of mobile communication system for railways.

Conclusions

HSR is developing very quickly in many parts of the world. Safety, reliability, high efficiency, environmental friendliness, comfort and humanization are the goals of future HSR developments. Against this application background, we propose network architectures and some challenging research directions including propagation characteristics, channel models, antenna designs, and seamless network modes switching. With the implementation of massive MIMO techniques and higher working frequency bands into the railway systems, even higher transmission rates and more reliable wireless transmissions can be realized at extremely high speeds, aided by careful designs that preserve the high quality of user experience.

Acknowledgments

This work was supported in part by the NSFC under Grant 61222105, National 863 Project under Grant 2014AA01A706, the State Key Lab project under Grant RCS2014ZT11 and RCS2014ZZ03, the Key Project of Chinese Ministry of Education under Grant 313006, the NSFC under Grant U1334202, and the Natural Science Base Research Plan in Shaanxi Province of China under Grant 2015JM6320.

References

[1] B. Ai *et al.*, "Challenges Toward Wireless Communications for High-Speed Railway," *IEEE Trans. Intelligent Transportation Systems*, vol. 15, no. 5, Oct. 2014, pp. 2143–58.
[2] K. Guan, Z. Zhong, and B. Ai,"Assessment of LTE-R Using High Speed Railway Channel Model," *3rd Int'l. Conf. Commun. and Mobile Computing*, Qingdao, China, Apr. 2011, pp. 461–64.
[3] Z. Zhong *et al.*, *Fundamental Theory of GSM-R Wireless Networks*, Beijing Jiaotong Univ. Publishing House, June 2009.
[4] METIS project, Deliv. D1.1, "Scenarios, Requirements and KPIs for 5G Mobile and Wireless System," tech. rep., Apr. 2013.
[5] HORIZON 2020 Work Programme 2014–2015 11. Smart, Green and Integrated Transport Revised, EC Decision C(2014)4995, July 2014.
[6] B. Ai *et al.*, "Social Network Services for Rail Traffic Applications," *IEEE Intelligent Systems*, vol. 29, no. 6, Dec. 2014, pp. 63–69.
[7] T. L. Marzetta, "Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, Nov. 2010, pp. 3590–3600.
[8] T. Kikkawa and Y. Sanada, "Wireless Relay Networks Using Multiple Frequency Bands," *2010 Int'l. Symp. Info. Theory and Its Applications*, Taichung, Taiwan, Oct. 2010, pp. 203–08.
[9] 3GPP, Release 12, V12.0.0, "Technical Specification Group Radio Access Network: Study on 3D Channel Model for LTE," tech. rep., Sept. 2014.
[10] S. Schwarz, *et al.*, "Pushing the Limits of LTE: A Survey on Research Enhancing the Standard," *IEEE Access*, vol. 1, May 2013, pp. 51–62.
[11] J. Ikuno, C. Mehlfuhrer, and M. Rupp, "A Novel Link Error Prediction Model for OFDM Systems with HARQ," *IEEE ICC*, Kyoto, Japan, June 2011, pp. 1–5.
[12] S. Schwarz, and M. Rupp, "Predictive Quantization on the Stiefel Manifold," *IEEE Signal Processing Lett.*, vol. 22, no. 2, Feb. 2015, pp. 234–38.
[13] S. Schwarz, C. Mehlfuhrer, and M. Rupp, "Throughput Maximizing Multiuser Scheduling with Adjustable Fairness," *IEEE ICC*, Kyoto, Japan, June 2011, pp. 1–5.

Biographies

Bo Ai [SM] (aibo@ieee.org) received his Ph.D. from Xidian University. He is a professor and Ph.D. advisor at Beijing Jiaotong University. He is deputy director of the State Key Lab of Rail Traffic Control and Safety in China. He has authored and co-authored six books and more than 220 papers. He is an IET Fellow. His research interests are focused on rail traffic mobile communications and channel modeling.

Ke Guan [M] received his Ph.D. from Beijing Jiaotong University. He is an associate professor at Beijing Jiaotong University. He received the International Union of Radio Science Young Scientist Award in 2014. He is a member of the IC1004 initiative. He has authored and co-authored more than 60 papers. His research interests include the measurement and modeling of wireless propagation channels, rail traffic communications, and future terahertz communication systems.

Markus Rupp [F] (mrupp@nt.tuwien.ac.at) received his Dr.-Ing. degree at the Technische Universität Darmstadt, Germany. Since 2001 he is a full professor at the Vienna University of Technology. He served as dean from 2005–2007 and as head of the Institute from 2014–2015. He has authored and co-authored more than 450 scientific papers and patents on adaptive filtering, wireless communications, and rapid prototyping, as well as automatic design methods.

Thomas Kürner [SM] (kuerner@ifn.ing.tu-bs.de) received his Dr.-Ing. Degree from Universität Karlsruhe, Germany. Since

2003, he has been a professor of mobile radio systems at TU Braunschweig, Germany. His work areas are propagation, self-organization of cellular networks, car-to-x communications, and channel characterization for future terahertz communication systems. He chairs the IEEE802.15 TG 100G and the WG of the European Association on Antennas and Propagation.

XIANG CHENG [SM] (xiangcheng@pku.edu.cn) received his Ph.D. degree from Heriot-Watt University and the University of Edinburgh, United Kingdom. He is an associate professor at Peking University. He has authored and co-authored over 80 papers. His research interests include mobile propagation channel modeling and simulation, next-generation mobile cellular systems, intelligent transport systems, and hardware prototype development and practical experiments.

XUE-FENG YIN [M] (yinxuefeng@tongji.edu.cn) received his Ph.D. degree in wireless communications from Aalborg University, Denmark, in 2006. Since 2008 he has been an associate professor at Tongji University. He has published more than 60 technical papers and co-authored a book on channel characterization and modeling. His research interests are in parameter estimation for radio channels, channel characterization, and stochastic modeling.

QI WANG [S] received his B.S. degree in communication engineering from Beijing Jiaotong University. He is currently working toward his Ph.D. degree in the State Key Lab of Rail Traffic Control and Safety, Beijing Jiaotong University. His research interests mainly include channel modeling for massive MIMO and vehicular-to-vehicular communications.

GUO-YU MA received his B.S. degree in electrical engineering from Beijing Jiaotong University, China, in 2012. Now he is working toward his Ph.D. degree at the State Key Lab of Rail Traffic Control and Safety, Beijing Jiaotong University. His current research interests are focused on massive MIMO and non-orthogonal multiple access techniques for 5G.

YAN LI [S] received her B.S. degree in electrical engineering in 2011 from Beijing Jiaotong University, where she is currently working toward her Ph.D. degree with the State Key Lab of Rail Traffic Control and Safety. She is now studying at the University of British Columbia, Canada, as an international exchange Ph.D. student. Her research interests are in the field of measurement and modeling of wireless propagation channels.

LEI XIONG received his Ph.D. from Beijing Jiaotong University in 2007. He is an associate professor at Beijing Jiaotong University. He has authored and co-authored two books and more than 30 papers. He is an expert on railway communications in China. His research interests are focused on rail mobile communications, channel simulation, and software defined radio.

JIAN-WEN DING received his M.S. degree from Beijing Jiaotong University. He is a lecturer at Beijing Jiaotong University. He is deputy director the of Laboratory of Rail Traffic Mobile Communication of Beijing Jiaotong University. He has authored and co-authored four books and more than 20 papers. His research interests are focused on rail traffic mobile communications and channel modeling.

WDM RoF-MMW and Linearly Located Distributed Antenna System for Future High-Speed Railway Communications

Pham Tien Dat, Atsushi Kanno, Naokatsu Yamamoto, and Testuya Kawanishi

ABSTRACT

In this article, we propose dual-hop network architecture capable of providing high-speed communications to high-speed trains (HSTs). The system uses a seamless fiber-millimeter-wave system for backhaul transmission from a central station to antennas on trains, and a high-speed in-train Wi-Fi network. The system can be combined with signal processing and network control technologies to compensate for interference and the Doppler effect, and to reduce the number of handovers. It can realize seamless connectivity between the inside and outside of trains to avoid penetration loss and help organize the in-train network optimally to increase coverage and data rate. We present and discuss the possible network architecture and technologies that can help realize the proposed network. We also present a proof-of-concept demonstration on a high-performance seamless fiber-MMW system that can be applied for applications in backhaul networks. The proposed network can be an attractive solution to provide broadband services such as video on demand and high-speed mobile signals to users on HSTs.

INTRODUCTION

High-speed trains (HSTs) have been developed rapidly worldwide, and are considered a fast, convenient, and green public transportation system. Communication demands of users on HSTs have also rapidly increased together with the emergence of broadband services such as video on demand, high-definition television, and new mobile services. However, these communication demands have not been paid much attention. Fast-moving users still need to transfer to large cells, which considerably reduces the available bandwidth [1]. In order to provide high data rate services to trains, gigabit-level communications should be realized and new broadband wireless communication systems should be developed. Nevertheless, there are many challenges to construct such a high-speed network for HSTs, including high penetration loss, fading and Doppler effects (DEs), and especially frequent handover.

CURRENT TECHNOLOGIES

To serve the increasing communication demand of users on moving vehicles, the Third Generation Partnership Project (3GPP) recently proposed several techniques [2]. Among the solutions, a dedicated moving relay node is considered cost effective to serve data-intensive users on HSTs [3]. It uses a hierarchical approach with separated backhaul links and access networks. With this configuration, vehicular penetration loss and shadowing effects can be avoided, and the number of handovers can be reduced. However, it is difficult to increase transmission throughput to end users because of capacity bottleneck of the backhaul links [3].

Recently, high-speed systems using a combination of radio over fiber (RoF) and distributed antenna systems (DASs) were proposed for HSTs [4, 5]. In these systems, by installing many small cells, remote antenna units (RAUs), along railway tracks, the distance of wireless backhaul links between the ground and the trains can be reduced, thus helping to increase transmission capacity. The RAUs can be connected to a central station (CS) by a wavelength-division multiplexing (WDM) RoF network. The use of RoF technology is very promising to simplify the networks, especially the RAUs. However, transmission latency and power consumption would be the main concerns because fiber and wireless links are connected via media converters. The use of wireless backhaul in the microwave (MW) band also limits the transmission capacity.

To further increase the capacity, a system with a fast handover mechanism using free-space optical (FSO) technology was proposed [6]. In this system, instead of using radio wave technology, high-speed FSO communication is used for backhaul links between HSTs and networks on the ground. Many lasers are installed along the railway track to communicate and keep track with a laser installed at the end of the trains. Each laser has a tracking mechanism to steer its laser beam toward the opposite transceiver.

Pham Tien Dat, Atsushi Kanno, and Naokatsu Yamamoto are with the National Institute of Information and Communication Technology.

Testuya Kawanishi is with the National Institute of Information and Communication Technology and Waseda University.

0163-6804/15/$25.00 © 2015 IEEE

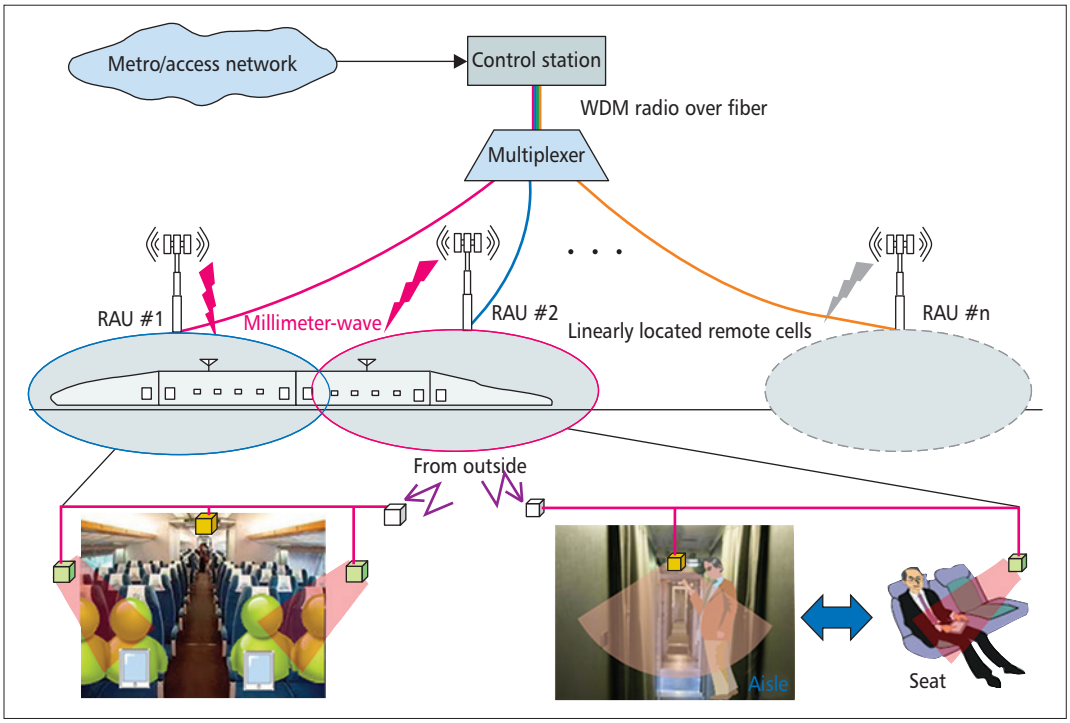


Figure 1. WDM RoF-MMW and a linearly located cell system for high-speed railway communications.

In order to provide high-data rate services to the trains, gigabit level communications should be realized and new broadband wireless communication systems should be developed. Nevertheless, there are many challenges to construct such a high-speed network for HSTs, including high penetration loss, fading and Doppler effects, and especially frequent handover.

Because the bandwidth of laser beams is much wider than those of radio waves, transmission capacity to HSTs can be greatly increased. Nevertheless, keeping stable communication for a train running at a very high speed would be a challenge because of the narrow size of the laser beam. The performance of FSO systems is also largely dependent on weather conditions, especially rain and atmospheric turbulence.

Our Contributions

To provide a stable and high-speed communication to HSTs, in this article, we propose a new dual-hop system using a seamless combination of fiber and millimeter-wave (MMW) links. The system comprises a high-capacity backhaul connected with an in-train wireless network via train antenna units (TAUs). The backhaul network, which consists of a WDM optical network and MMW links, connects a CS with many TAUs. The WDM optical network is to connect the CS with many RAUs installed along the railway track, and the MMW links are to connect RAUs with TAUs. The in-train network can consist of many access points (APs), and can be adaptively organized to increase coverage and data rate. This system is a part of an ongoing project called "MMW backhaul" funded by the Japanese government to develop a high-capacity communication system for HSTs. The system can help to increase transmission capacity compared to current solutions using communications in the MW bands. It can also help to increase communication reliability compared to FSO-based systems. The aim of this article is to provide a detailed discussion of our vision, general concepts, and underlying technologies for developing such a system. The performance results of a seamless fiber-MMW link, which is one of the most important underlying technologies in the proposed system, are also presented to show the potential of the system.

The remainder of this article is organized as follows. First, we introduce the MMW backhaul system. Next, we describe various underlying technologies in the MMW backhaul system. We then present a performance evaluation of a seamless fiber-MMW system. We also discuss the challenges and open issues of the system. Finally, we draw our conclusion at the end of this article.

System Description

Figure 1 shows the concept of our proposed system using a WDM RoF and linearly located DAS (LL-DAS) using MMW radio links. In this configuration, the WDM RoF-MMW and LL-DAS system serves as a backhaul network for distributing signals from a CS to antennas on trains. At the CS, all wireless services in the MW bands can be encapsulated into one or digitized to baseband (BB) signals before being converted to optical signals. The signals are then transmitted to the RAUs located along the railway track via the WDM RoF network. Specific wavelengths can be assigned to each RAU by the CS to identify the remote cell position and optimize the signal distribution. At the RAUs, received optical signals are converted to radio signals in the MMW band and transmitted via MMW links to TAUs without any further processing. In this manner, expensive and intelligent signal processing equipment can be located at the CSs and shared by many RAUs.

To avoid interference and enhance the signal transmission quality, high-directivity antennas should be installed at RAUs and TAUs. MMW signals received by TAUs are down-converted to the signals in the MW or BB bands, and pro-

The radio signal can be emitted into the air directly without any signal processing. By this seamless combination, RAUs can be very simple. This can help significantly simplify the systems, especially the RAUs, and considerably reduce the cost, power consumption, and latency.

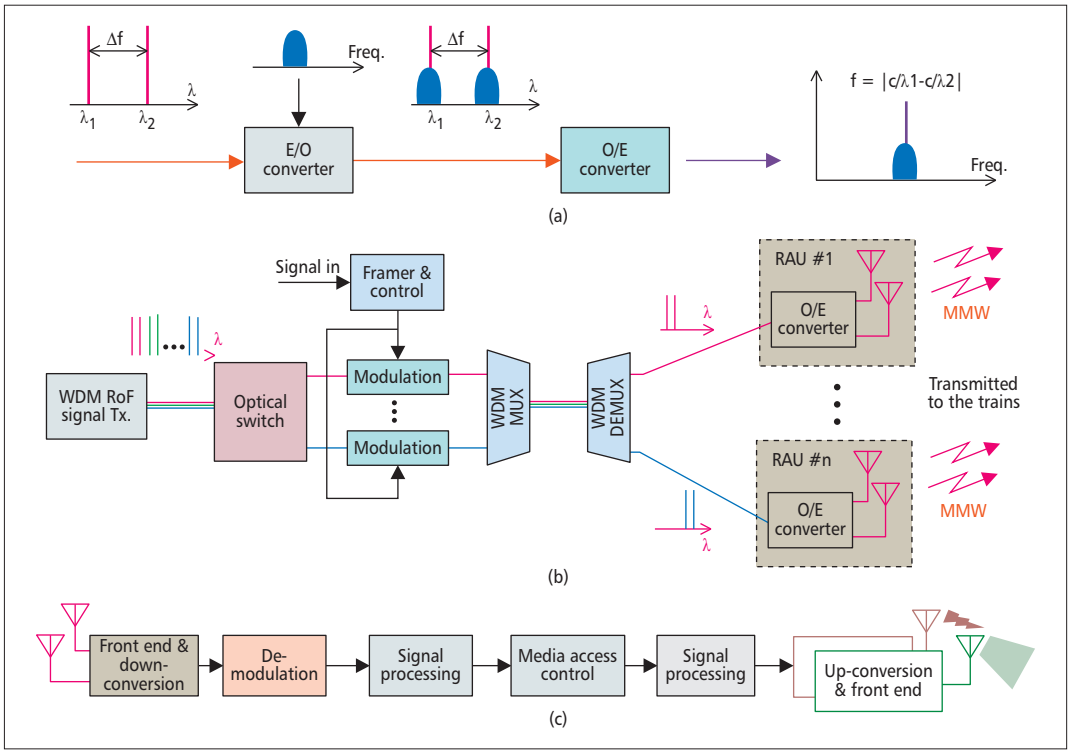


Figure 2. a) Principle of the seamless fiber-wireless combination; b) WDM RoF-MMW system for linearly located cells; c) TAU configuration.

cessed by signal processing units at TAUs to compensate for interference and fading effects. A media converter can convert the signals from the backhaul network to an in-train mobile network such as Wi-Fi. Finally, the signals are transmitted to end users via the in-train networks. These networks can be optimally organized to enhance both throughput and coverage to users. Advanced wireless technologies, such as small cells and beamforming, can be used in high-demand areas. In addition, the network can be combined with signal processing and network control technologies to compensate for the deterioration effects of interference, vibration, and fading, and to reduce the number of handovers. By using this network architecture, a high-speed communication system with simplified remote cell sites and low handover frequency can be realized for applications in high-speed railways. Details of the technologies for developing such a system are presented in the next subsections.

Key Techniques in the MMW Backhaul System

Seamless Fiber-MMW Links

In our proposed network, the combined fiber optic and MMW system forms a flexible and high-capacity backhaul system for connecting CSs on the ground to trains. There are two different methods for the combination:

- Connections via wired-wireless media converters (WWMCs) at RAUs [7]
- Seamless connections using RoF technology and direct optical-to-electrical (O/E) conversion at RAUs [8]

In the former method, wireless signals are carried from a CS to each RAU by an optical carrier signal. At the RAUs, the optical signal is converted to a radio signal in the MMW band via a WWMC. This WWMC may consist of radio front-ends, digital signal processing (DSP), and electrical local oscillators [7]. Because many equipment units and functions are needed at RAUs, the system is complicated, resulting in high cost, high power consumption, and long transmission delay. In our system, we adopt a seamless combination of fiber and MMW links using RoF technology [8]. Figure 2a shows the concept of this seamless combination. A RoF signal consisting of two optical signals is first generated. These two optical signals should have a frequency difference equal to the frequency of the MMW signal to be generated at the O/E converter. The generated RoF signal is modulated by electrical signals at an electrical-to-optical (E/O) converter and sent to the RAU via a fiber cable. At the RAU, a high-speed O/E converter converts the RoF signal to a radio signal. The radio signal can be emitted into the air directly without any signal processing. By this seamless combination, RAUs can be very simple. This can help significantly simplify the systems, especially the RAUs, and considerably reduce the cost, power consumption, and latency.

Furthermore, because the system is transparent to the transmitted signals from CSs, any signals can be transmitted over the system without having to reconfigure the RAUs. It also enables multiple radio signals, such as multiple services, multiple operators, multiple-band signals, and multi-radio access technologies, to coexist in the same system. However, a frequency- and phase-

stabilized MMW carrier signal should be generated to achieve a high-quality signal transmission. An optical modulation technology based method, as presented in our previous work [8], is suitable for generating a high-quality RoF and subsequent MMW signal. For converting a RoF signal to an MMW signal at RAUs, a uni-traveling-carrier photodiode is a promising candidate [9].

WDM Optical Distribution Network

In our system, signals from a CS should be distributed to many RAUs located along the rail track via an optical distribution network. To perform this function, similar to [4], we use a WDM RoF network. However, the network should be designed for optimal transmission of MMW signals. The configuration of the distribution network from the CS to RAUs using WDM technology is shown in Fig. 2b. At the CS, a multi-wavelength signal consisting of multiple RoF signals is first generated. A photonic frequency comb as presented in [10] can be a promising candidate to generate a frequency- and phase-stabilized multi-wavelength signal. An optical switch can be used to select appropriate RoF signals from the generated multi-wavelength signal and send to corresponding optical modulators for data modulation. In general, two modulation methods can be applied: simple intensity modulation and a high optical spectrum efficiency in-phase/quadrature (I/Q) modulation. The use of I/Q modulation can help increase the transmission capacity over the optical network.

The modulated RoF signals are then combined by a WDM multiplexer (MUX). The combined signals are transmitted via a single fiber cable to an intermediate point, where they are separated by a demultiplexing device (DEMUX). The separated signals are finally transmitted to corresponding RAUs by different fiber cables. In this configuration, a pair of fixed optical wavelengths can be assigned to each RAU.

Millimeter-Wave Backhaul Links

In our proposed system, the carrier frequency of radio links from RAUs to TAUs is an important issue. The 60 GHz band radio is a possible candidate because of its broad bandwidth. However, high atmospheric attenuation at this frequency band would significantly limit the transmission distance. The E-band (60–90 GHz) and W-band (75–110 GHz) are more promising for a wireless backhaul network due to its large available bandwidth and low atmospheric attenuation. To extend the radio transmission distance between RAUs and TAUs, the use of high-gain and high-directivity antennas and high-output-power amplifiers is important.

In-Train Communication Network

In our proposed dual-hop system, signals, after being received from the backhaul network, can be processed by DSP at TAUs and fed into an in-train network via a media converter. The process of recovering signals from the backhaul fiber-MMW network and distributing them into the in-train network is shown in Fig. 2c. The received MMW signals from the backhaul network are down-converted and demodulated before being input into a DSP to compensate for deterioration effects on the backhaul. It is then distributed into the in-train network via a media converter and DSP blocks.

The in-train network can be optimally organized using recently advanced wireless technologies such as small cells and/or beamforming and MMW to increase both throughput and coverage to end users. Wi-Fi networks at conventional frequency bands of 2.4 and 5 GHz can be good candidates. Many APs can be installed at different places inside the trains and connected to the TAUs via fiber links. To increase transmission throughput, high-speed WLANs such as 60 GHz IEEE 802.11ad can be further exploited in high-demand areas such as passengers' seats. Thus, the combination of Wi-Fi networks at 2.4/5 GHz and 60 GHz is very promising. A switching mechanism is necessary to switch between the two networks when users move inside the cars.

Network Control Technology

Handover is a major issue in high-speed railway communications, especially when using small-cell-based networks such as MMW systems for backhauling to trains. Handover occurs each time a TAU exceeds the cell boundary of one RAU and reconnects to the next RAU. However, because users on the cars need to perform a handover procedure at the same time, frequent handovers can be avoided using a moving cell concept. This is a cell that can move with the train along the railway track so that the train can communicate on the same frequency during the whole connection [4]. In this context, instead of having a fixed cell pattern with a fixed frequency for each cell, a dynamic cell pattern in which frequencies of the cells that are being communicated with trains can be dynamically assigned together with the movement of the trains.

There are two methods of performing the moving-cell concept in our system. The first one is similar to the system in [4] using an optical switch to change the optical wavelengths before data modulation at the CS. In our system, we use another method, presented in Fig. 3. In this configuration, the CS comprises several signal processing units (SPUs) to communicate with core networks and perform some signal processing functions, a switch for connecting the SPUs and RAUs, a switch controller for controlling the switch, and a train position detector (TPD) for detecting a train's position. The TPD can measure the received power from the uplink signals, and can detect the position and direction of the train. It can also realize an RAU the train is approaching on the basis of the absolute values of the received power. From the train position information from the TPD, the switch controller can control the switch to change the connections between SPUs and RAUs accordingly.

By using this network configuration, even when a train crosses the boundary of an RAU, it is not necessary to perform a handover procedure because signals from the same SPUs are transmitted to the trains. In addition, because the SPUs can be used for many RAUs, the circuit scale at the CS can be greatly reduced. Furthermore, the system power consumption can be greatly reduced by deactivating the RAUs the train is not approaching.

Handover occurs each time a TAU exceeds the cell boundary of one RAU and reconnects to the next RAU. However, because users on the cars need to perform a handover procedure at the same time, frequent handovers can be avoided using a moving cell concept.

For the MMW links, frequency-division multiplexing can be used to transmit and receive uplink and downlink signals at TAUs and RAUs. For the optical networks, wavelengths different from those in the downlink direction can be used to transmit uplink signals from RAUs to the CS.

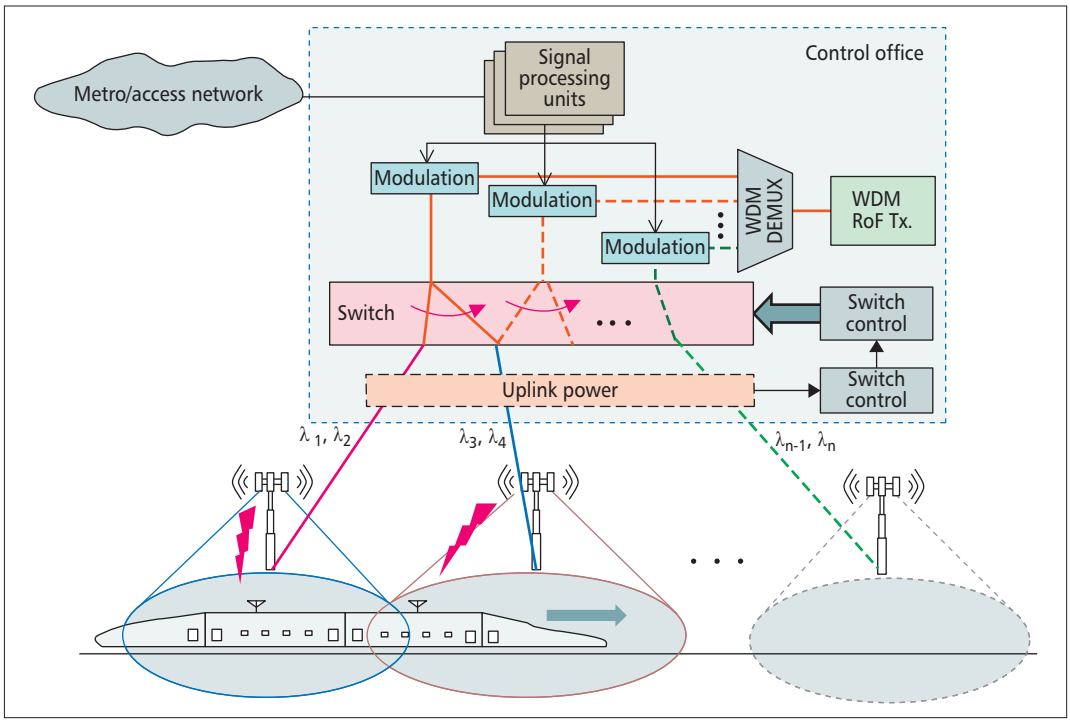


Figure 3. Network control and the moving cell concept.

Uplink Communications

The system can also be used for transmission of uplink (UL) signals from users on the trains to the networks on the ground. Signals emitted from user terminals can be captured by APs on the trains. The signals are then transmitted to TAUs where they are converted to BB signals via media converters, and processed by DSP before being up-converted to MMW signals to transmit over the backhaul system. At RAUs, the received MMW signals can be converted directly to optical signals using high-speed E/O converters and transmitted back to the CS via RoF links. Alternatively, the received signals are down-converted to low-frequency signals such as in the MW or BB bands before feeding into common RoF links to transmit to the CS. At the CS, the signals are converted back to electrical signals, and processed to recover the original signals correctly. For the MMW links, frequency-division multiplexing can be used to transmit and receive uplink and downlink signals at TAUs and RAUs. For the optical networks, wavelengths different from those in the downlink direction can be used to transmit uplink signals from RAUs to the CS.

Energy-Efficient Network

A highly energy-efficient network is very important for communications to HSTs because a large number of RAUs should be installed along the rail track. On the basis of our previous work [11], we can model the power consumption for each track length unit of the proposed system as

$$P = \frac{K}{D} + A \times (2^{\frac{C}{B}} - 1) \times D,$$

where K and A are the constants that can be determined from the power consumption of components at the CS and RAUs, B is the system bandwidth, C is the maximum achievable transmission rate target per cell, and D is the length of each RAU. It is clear from this model that there is always an optimal value of D so that the power consumption of the system is minimized. This value is determined as

$$D_{\min} = \sqrt{\frac{K}{A(2^{\frac{C}{B}} - 1)}}.$$

When D is smaller than this value, the system power consumption considerably increases because of the domination of power consumption at the CS. When increasing D over the optimal value, power consumption also increases significantly because of the higher power consumption of the RAUs. The cell size of RAUs therefore should be selected to optimize the power consumption of the system.

Seamless Fiber-MMW System Performance

The backhaul network using a combination of fiber optic and MMW links is the key element in our proposed system. The system should have stable and high performance in order to realize high-quality signal transmission from the ground to trains. In the following subsections, we present a proof-of-concept demonstration of transmission of radio signals over fiber-wireless systems in both the downlink (DL) and UL directions.

Downlink System Performance

In the DL direction, a stable and high-performance fiber-MMW system can be realized using a phase- and frequency-stabilized RoF signal

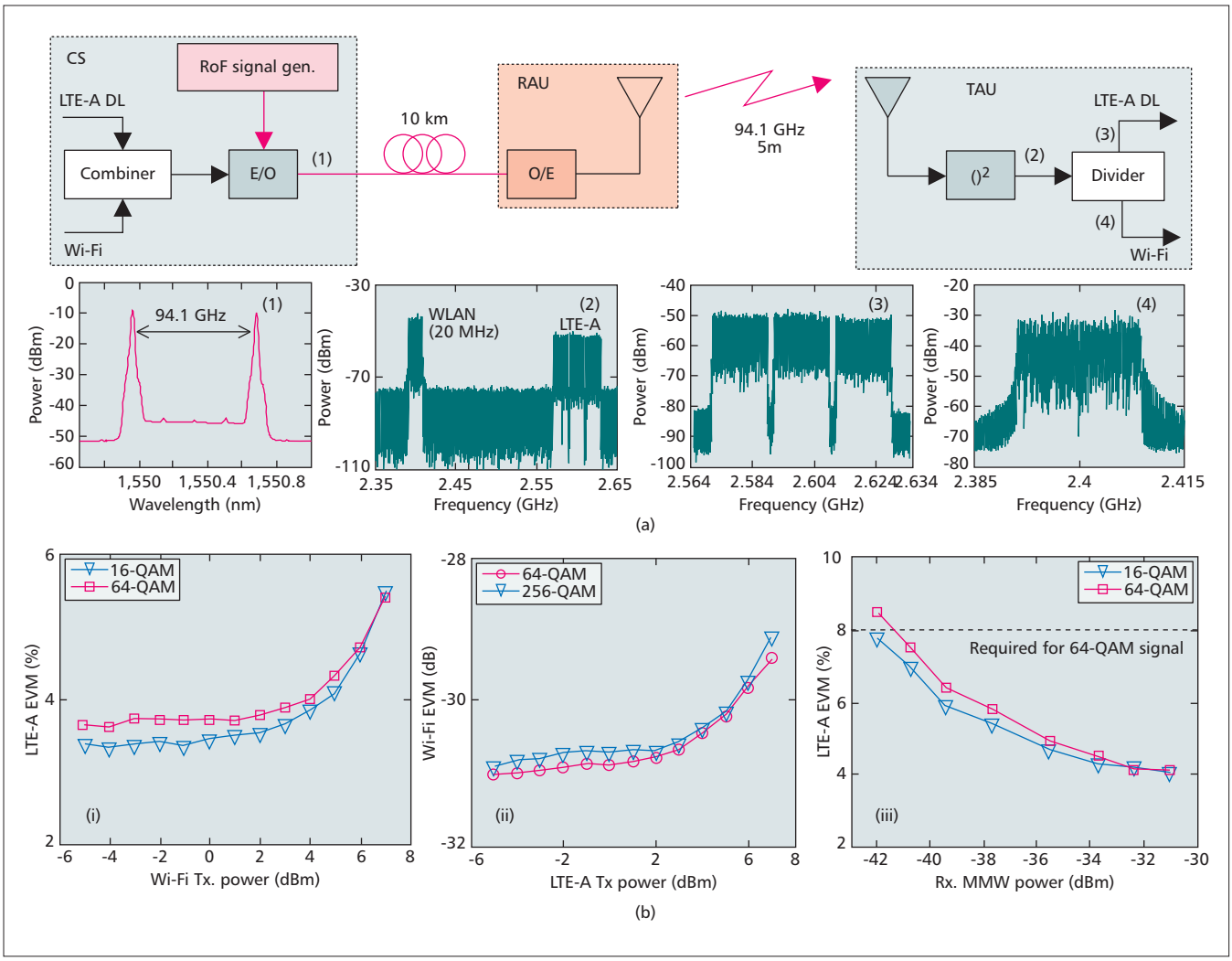


Figure 4. a) Experimental setup; b) performance of downlink signal transmission over a seamless RoF-MMW system.

generator at the transmitter [12] and a low phase noise MMW signal detector at the receiver, as shown in Fig. 4a [8]. We can transmit simultaneously multiple wireless signals over the system using subcarrier multiplexing. Shown in the figure is a simple case of simultaneous transmission of two wireless signals. Standard-compliant wireless signals, including a very-high-throughput WLAN 802.11ac signal and a high-speed DL LTE-A signal, are combined and transmitted over a combined 10 km fiber and 5 m, 94.1 GHz system. First, a RoF signal consisting of two optical signals with a frequency difference of 94.1 GHz is generated by a RoF signal generator [12]. The generated signal is then modulated by the wireless signals at an E/O converter and is transmitted to an optical receiver located at an RAU. At the RAU, the signal is directly converted to an MMW signal at 94.1 GHz by a high-bandwidth O/E converter. The MMW signal is emitted into free space by a horn antenna. After transmission over a 5 m link in the air, the signal is received by another horn antenna at a TAU, and down-converted to the original wireless signals by a square-law envelope detector (ED). The recovered signal is separated by a power divider and finally demodulated by commercially available software. The spectra of the transmitted signals at different points along the system are also presented in the figure.

We evaluate the signal performance using the root mean square (rms) error vector magnitude (EVM) parameter. The measurement follows the standard test procedure recommended by the IEEE 802.11 and 3GPP standards [13]. Figures 4b(i) and 4b(ii) show the results of Long Term Evolution-Advanced (LTE-A) and 802.11ac signals for different transmission powers of the co-transmission signal. The required EVM values of 64-quadrature amplitude modulation (QAM) LTE-A and 256-QAM 802.11ac signals are 8 percent and –30 dB, respectively. All signals are successfully transmitted. The figures show that increasing the transmit power of the coexisting signal has some impact on the performance of the other signal. However, the increase in the measured EVM is not very significant, confirming that satisfactory performance can be achieved even with the coexistence of other wireless signals. The performance of the LTE-A signal for different received powers of the MMW signal is shown in Fig. 4b(iii). It is confirmed that a minimum power of approximately –41 dBm should be received to success-

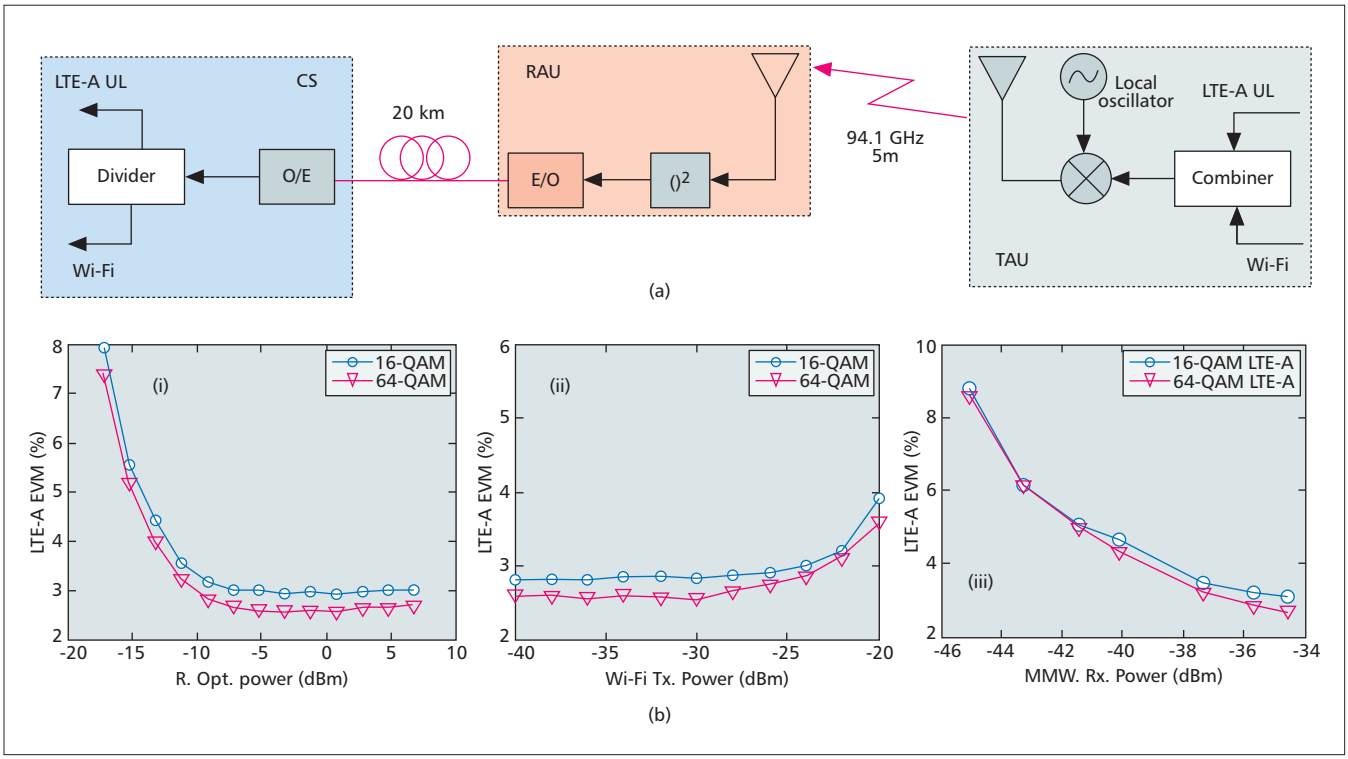


Figure 5. a) Experimental setup; b) performance of the UL LTE-A signal after transmission over the uplink MMW-fiber system.

fully recover a 64-QAM LTE-A signal. Using a similar approach as in [13], we can estimate that a maximum wireless transmission distance can be increased up to approximately 1 km using high-gain antennas and a high-output power amplifier.

We should note that a high-capacity BB signal transmission over the system can also be realized using optical IQ modulation and coherent detection [12] or a broadband orthogonal frequency-division multiplexing (OFDM) method [14]. In the case of using OFDM, signal parameters should be optimized according to transmission conditions such as fiber cable range, fading and attenuation on the radio links, and environment vibration and pressure.

Uplink System Performance

In the UL direction, a cascade of an MMW link and a RoF system in the MW band can be an attractive choice [13]. Similar to the DL direction, simultaneous transmission of multiple wireless signals can be realized. Figure 5a shows an example of simultaneous transmission of standards-compliant UL LTE-A and 802.11ac signals over a combined 94.1 GHz and 20 km fiber system. In this system, the wireless signals are combined and up-converted to an MMW signal at 94.1 GHz by an electrical mixer at the TAU. The signal is then emitted into free space by a horn antenna. After transmission over a 5 m free-space link as in the DL direction, it is received by another horn antenna at the RAU. The signal is then coupled into an ED to down-convert to the original wireless signals. Subsequently, it is converted to an optical signal and transmitted to the CS via a 20 km SMF cable. The received optical signal is converted back to the original wireless signal and finally analyzed by the software.

The performance of the UL LTE-A signal in the UL direction is shown in Figs. 5b(i) and 5b(ii) for different received optical powers and transmission powers of the coexisting 802.11ac signal, respectively. In the UL direction, a range of received optical power is essential because it defines an attenuation range for various fiber lengths and splitting ratio in the WDM distribution network. Figure 5b(i) shows that a range of approximately 22 dB can be achieved for a 64-QAM LTE-A signal, confirming the possibility of using widely used WDM passive optical networks in our backhaul system [15]. The effect of simultaneous transmission on UL LTE-A signal performance is shown in Fig. 5b(ii). Similar to the DL direction, increasing transmit power of the co-transmitting signal has some influence on the performance of the LTE-A signal. However, the effect is quite small, and the performance is still much better than the standard requirement. Figure 5b(iii) shows the results for different received power of the MMW signal at the RAU. From the figure, a minimum MMW power of approximately –44.5 dBm should be received to recover a 64-QAM UL LTE-A. Similar to the DL direction, we can estimate that the maximum transmission range for the UL MMW link can be extended to longer than 1 km using high-gain antennas and high-output power amplifiers.

Challenges and Open Issues

There are several challenges to develop a stable communication system for HSTs using MMW links for the backhaul network. First, radio-wave links between RAUs and TAUs can be affected by vibration and wind pressure when trains move in and out of tunnels. To overcome this chal-

lenge, in our system, antenna transceivers and radio circuits with vibration and weather resistance that enable stable communications in the railway environment will be developed.

DE is another important issue that should be considered in the proposed system. It is well known that DE considerably increases for fast moving objects such as HSTs and for a high-frequency radio network. To compensate for this effect, in our system, a specially designed DSP algorithm and circuit will be developed. It can be based on solutions being used in some MMW wireless standards such as IEEE 802.11ad and IEEE 802.15.3c but suitably developed for fast-moving users. Because of the dual-hop network architecture, the algorithm and circuits can be placed at TAUs that are normally not limited by size, power consumption, and complexity compared to user terminals in a conventional one-hop network.

Interference between direct and ground-reflected waves is another issue in the wireless backhaul network. To compensate for this effect, the use of spatial diversity techniques will be considered. Installing antennas of different heights at RAUs, transmission diversity by space-time coding, or using receive diversity at TAUs can be effective methods.

Interference of radio signals from different remote cells when the trains move inside an overlap area of two adjacent RAUs is another challenge of the proposed system. To overcome this issue, signals from many TAUs can be combined and processed before recovering back to the transmitted signals. Normally, overlap areas between adjacent RAUs should be organized so that a packet signal can be successfully received without having to change to a new RAU. For example, with a packet length of 10 ms and train speed of 500 km/h, the length of overlap areas should be at least 1.4 m. The distance between TAUs on trains should be much larger than this length. Thus, even if one TAU moves inside the overlap area, the others will be outside. By combining signals from these TAUs, DSP can identify the signal coming from an old or a new RAU, and can suppress the interference.

Conclusion

We propose a dual-hop high-speed network for future high-speed railway communication. The system uses a WDM MMW RoF and LL-DAS system for the backhaul network and an adaptively optimized Wi-Fi in-train network. Using this two-hop architecture, seamless connectivity can be realized between the inside and outside of trains. The seamless combination of fiber optic and MMW links is important for simplifying the backhaul network, especially the remote small cells. By combining with a network control and moving cell concept, we can greatly reduce the number of handovers when trains move inside an area controlled by a CS. The network structure can also enable the activation and deactivation of remote cells adaptively according to the position of the trains, and help reduce power consumption.

We also present a proof-of-concept demonstration on a high-performance fiber-MMW system that can be used for the backhaul network. The measured results show that satisfactory performance can be achieved for simultaneous transmission of multiple radio signals in both downlink and uplink directions. It is estimated that a maximum distance of up to 1 km can be obtained for the MMW link transmission, which is sufficiently long for the small-cell DAS system. However, the size of each RAU should be optimized, depending on the required capacity, to minimize the system power consumption.

It is estimated that a maximum distance of up to one kilometer can obtained for the MMW link transmission, which is sufficiently long for the small cells DAS system. However, the size of each RAU should be optimized, depending on the required capacity, to minimize the system power consumption.

Acknowledgments

This work was conducted as a part of the "Research and development for expansion of radio wave resources," supported by the Ministry of Internal Affairs and Communications (MIC), Japan.

References

[1]. F. De Greve *et al.*, "FAMOUS: A Network Architecture for Delivering Multimedia Services to Fast Moving Users," *Wireless Pers. Commun. J*, vol. 33, June 2005, pp. 281–304.

[2] 3GPP TR 36.836, "Technical Specification Group Radio Access Network; Mobile Relay for Evolved Universal Terrestrial Radio Access (E-UTRA)," tech. rep., accessed Nov. 20, 2012.

[3] Y. Sui *et al.*, "Moving Cells: A Promising Solution to Boost Performance for Vehicular Users," *IEEE Commun. Mag.*, vol. 51, no. 6, June 2013, pp. 62–68.

[4] Lannoo *et al.*, "Radio-over-Fiber-Based Solution to Provide Broadband Internet Access to Train Passengers," *IEEE Commun. Mag.*, vol. 45, no. 2, Feb. 2007, pp. 56–62.

[5] J. Wang *et al.*, "Distributed Antenna Systems for Mobile Communications in High Speed Trains, " *IEEE JSAC*, vol. 30, no. 4, May 2012, pp. 675–83.

[6] S. Haruyyama *et al.*, "New Ground-to-Train High-Speed Free-Space Optical Communication System with Fast Handover Mechanism," *Proc. IEEE OFC*, 2011.

[7] APT Report on "Wired and Wireless Seamless Connections Using Millimeter-Wave Radio over Fiber Technology for Resilient Access Networks," APT/ASTAP/REPT-11, Mar. 2014.

[8] P. T. Dat *et al.*, "Performance of a 90-GHz Radio-on-Radio-over-Fiber System Suitable for Communications in High-Speed Railways," *Proc. IEEE IMS*, 2014.

[9] H. Ito *et al.*, "W-band Uni-Travelling-Carrier Photodiode Module for High Power Photonic Millimeter-wave Generation," *Elect. Lett.*, vol. 38, iss. 22, Oct. 2002, pp. 1376–77.

[10] A. Kanno *et al.*, "Coherent MMW Terahertz Signal Transmission with Frequency-Reconfigurable RoF Transmitter Based on an Optical Frequency Comb," *Proc. IEEE GLOBECOM*, 2013.

[11] P. T. Dat *et al.*, "Energy and Deployment Efficiency of a Millimeter-Wave Radio-on-Radio-over-Fiber System for Railways," *Proc. IEEE OFC*, 2013.

[12] A. Kanno *et al.*, "Coherent Radio-over-Fiber and Millimeter-Wave Radio Seamless Transmission System for Resilient Access Networks," *IEEE Photon. J.*, vol. 4, no. 6, Dec. 2012, pp. 2196–204.

[13] P. T. Dat *et al.*, "High-Capacity Wireless Backhaul Network Using Seamless Convergence of Radio-over-Fiber and 90-GHz Millimeter-Wave," *J. Lightwave Tech.*, vol. 32, no. 20, Oct. 2014, pp. 0733–8724.

[14] T. L. Thanh *et al.*, "10-Gb/s Wireless Signal Transmission over a Seamless IM/DD Fiber-MMW System at 92.5 GHz," *Proc. IEEE ICC*, 2015.

[15] ITU-T Rec. G.987.2, "10-Gigabit-Capable Passive Optical Networks (XG-PON): Physical Media Dependent (PMD) Layer Specification," Oct. 2010.

Biographies

Pham Tien Dat [M'12] (ptdat@nict.go.jp) received his B.Eng. degree in electronics and telecommunication engineering from the Posts and Telecommunications Institute of Technology, Vietnam, in 2003, and his M.Sc. and Ph.D. degrees in science of global information and telecommunication studies from Waseda University, Japan, in 2008 and 2011, respectively. In 2011, he joined the National Institute of

Information and Communications Technology, Japan. His research interests are in microwave/millimeter-wave photonics, radio over fiber, and optical wireless systems.

ATSUSHI KANNO [M'11] received B.S., M.S., and Ph.D. degrees in science from the University of Tsukuba, Japan, in 1999, 2001, and 2005, respectively. In 2005, he was with the Venture Business Laboratory of the Institute of Science and Engineering, University of Tsukuba. In 2006, he joined the National Institute of Information and Communications Technology. His research interests are microwave/millimeter-wave/terahertz photonics, ultrafast optical communication systems, and lithium niobate optical modulators. He is a member of the Institute of Electronics, Information and Communication Engineers and the Japan Society of Applied Physics (JSAP).

NAOKATSU YAMAMOTO received his Ph.D. degrees in electrical engineering from Tokyo Denki University, Japan, in 2000. In April 2001, he joined the Communications Research Laboratory (CRL) (now the National Institute of Information and Communications Technology, NICT), Tokyo, Japan, where he is currently the director of the Lightwave Devices Lab. He was also with Tokyo Denki University as a visiting professor beginning in December 2013, and the Ministry of Internal Affairs and Communications as a deputy director from July 2012 to September 2013. His research interests include nanostructured materials and III-V semiconductor QD and their photonic device applications in photonic transport systems. He has proposed many types of novel crystal growth techniques, and successfully developed a QD optical frequency comb laser, an ultra-broadband QD light source, and a wavelength-tunable QD laser. He successfully demonstrated a high-speed and ultra-broadband photonic transport system constructed with novel nanostructured photonic devices. Recently, he proposed the use of 1.0-μmwave band photonic transport systems to develop a novel optical frequency resource for optical communications.

TETSUYA KAWANISHI [F] received his B.E., M.E., and Ph.D. degrees in electronics from Kyoto University, Japan, in 1992, 1994, and 1997, respectively. From 1994 to 1995, he was with the Production Engineering Laboratory of Panasonic. During 1997, he was with the Venture Business Laboratory, Kyoto University, where he was engaged in research on electromagnetic scattering and near-field optics. In 1998, he joined the Communications Research Laboratory, Ministry of Posts and Telecommunications (now the National Institute of Information and Communications Technology), Tokyo, Japan. During 2004, he was a visiting scholar in the Department of Electrical and Computer Engineering, University of California at San Diego. Since April 2015, he has been a professor at Waseda University, Tokyo, Japan. His current research interests include high-speed optical modulators and RF photonics.

Call for Papers

IEEE Communications Magazine

Bio-Inspired Cyber Security for Communications and Networking

Background

Nature is Earth's most amazing invention machine for solving problems and adapting to significant environmental changes. Its ability to address complex, large-scale problems with robust, adaptable, and efficient solutions results from many years of selection, genetic drift, and mutations. Thus, it is not surprising that inventors and researchers often look to natural systems for inspiration and methods for solving problems in human-created artificial environments. This has resulted in the development of evolutionary algorithms including genetic algorithms and swarm algorithms, and of classifier and pattern detection algorithms, such as neural networks, for solving hard computational problems.

A natural evolutionary driver is to survive long enough to create a next generation of descendants and ensure their survival. One factor in survival is an organism's ability to defend against attackers, both predators and parasites, and against rapid changes in environmental conditions. Analogously, networks and communications systems use cyber security to defend their assets against cyber criminals, hostile organizations, hackers, activists, and sudden changes in the network environment (e.g., DDoS attacks). Many of the defense methods used by natural organisms may be mapped to cyber space to implement effective cyber security. Some examples include immune systems, invader detection, friend vs. foe, camouflage, mimicry, evasion, and so on. Many cyber security technologies and systems in common use today have their roots in bio-inspired methods, including anti-virus, intrusion detection, threat behavior analysis, attribution, honeypots, counterattack, and the like. As the threats evolve to evade current cyber security technologies, similarly the bio-inspired security and defense technologies evolve to counter the threat.

The goal of this Feature Topic is twofold: (1) to survey the current academic and industry research in bio-inspired cyber security for communications and networking so that the ComSoc community can understand the current evolutionary state of cyber threats, defenses, and intelligence, and can plan for future transitions of the research into practical implementations; and (2) to survey current academic and industry system projects, prototypes, and deployed products and services (including threat intelligence services) that implement the next generation of bio-inspired methods. Please note that we recognize that in some cases, details may be limited or obscured for security reasons. Topics of interests include, but are not limited to:

- Bio-inspired anomaly and intrusion detection
- Adaptation algorithms for cyber security and networking
- Biometrics related to cyber security and networking
- Bio-inspired security and networking algorithms and technologies
- Biomimetics related to cyber security and networking
- Bio-inspired cyber threat intelligence methods and systems
- Moving-target techniques
- Network artificial immune systems
- Adaptive and evolvable systems
- Neural networks, evolutionary algorithms, and genetic algorithms for cyber security and networking
- Prediction techniques for cyber security and networking
- Information hiding solutions (steganography, watermarking) and detection for network traffic
- Cooperative defense systems
- Bio-inspired algorithms for dependable networks

Submissions

Articles should be tutorial in nature and written in a style comprehensible and accessible to readers outside the specialty of the article. Authors must follow *IEEE Communications Magazine*'s guidelines for preparation of the manuscript. Complete guidelines for prospective authors can be found at http://www.comsoc.org/commag/paper-submission-guidelines.

It is important to note that *IEEE Communications Magazine* strongly limits mathematical content, and the number of figures and tables. Paper length should not exceed 4500 words. All articles to be considered for publication must be submitted through the IEEE Manuscript Central site (http://mc.manuscriptcentral.com/commag-ieee) by the deadline. Submit articles to the "June 2016/Bio-inspired cyber security for communication and networking" category.

Schedule for Submissions

- Submission Deadline: November 1, 2015
- Notification Due Date: February 1, 2016
- Final Version Due Date: April 1, 2016
- Feature Topic Publication Date: June 2016

Guest Editors

Wojciech Mazurczyk
Warsaw University of Technology
Poland
wmazurczyk@tele.pw.edu.pl

Sean Moore
Centripetal Networks
USA
smoorephd@gmail.com

Errin W. Fulp
Wake Forest University
USA
fulp@wfu.edu

Hiroshi Wada
Unitrends
Australia
hiroshi.wada@nicta.com.au

Kenji Leibnitz
National Institute of Information and Communications Technology
Japan
leibnitz@nict.go.jp

Providing Current and Future Cellular Services to High Speed Trains

Martin Klaus Müller, Martin Taranetz, and Markus Rupp

ABSTRACT

The demand for a broadband wireless connection is nowadays no longer limited to stationary situations, but also required while traveling. Therefore, there exist combined efforts to also provide wireless access on high speed trains (HSTs), in order to add to the attractiveness of this means of transportation. Installing an additional relay on the train, to facilitate communication, is an approach that has already been extensively discussed in the literature. The possibility of a direct communication between the base station and the passenger has been neglected until now, despite it having numerous advantages. Therefore, a comparison between these two opposing approaches is presented in this article, accompanied by a detailed discussion of the related aspects. Additionally, we present simulation results for the two approaches when applying different schemes to supply a wireless connection. We also discuss the presented results from the perspectives of mobile operators and train operators.

INTRODUCTION

In the current market of train services, being able to provide mobile broadband access to costumers has become a main inducement for choosing this means of transportation. Due to the ubiquitous use of the Internet and the rapid adoption of novel devices such as smartphones and tablet computers, most passengers have become accustomed to experiencing high data rates and having the service following them no matter where they go. With the number of commuters expected to increase, high user mobility is also one of the most emphasized scenarios in the initiative for the fifth generation (5G) of wireless communications. LTE-Advanced (LTE-A), the contemporary standard for wireless communication, is not optimized for the challenges of high speed train (HST) scenarios. Hence, many train operators, mostly in collaboration with mobile operators, have increased efforts to satisfy the ever increasing requirements. There also exist international collaborations on a broad level to push for higher data rates and shorter latencies, for example the Shift2Rail initiative by the European Union.[1] Additionally, LTE for Railway (LTE-R) has been proposed in [1] as an evolution of Global System for Mobile Communications — Rail(way) (GSM-R), addressing required performance parameters and necessary adaptations on the architecture of the system. Nevertheless, LTE-R has not yet been standardized. Furthermore, the main concern of LTE-R is not passenger communication, which is the scope of this article. Therefore, we mostly focus on LTE-A.

Wireless communications in HST scenarios is confronted with unique conditions that have a considerable impact on network planning. In particular, the scenarios are characterized by user equipments (UEs) being densely packed inside the train and moving at high speed, as well as the specific propagation effects in a diversity of different environments.

Most publications on this topic assume a relay-based approach, assuming additional hardware installed on the train that communicates with the base station (BS) as well as with user equipments (UEs) without communicating directly ([2, 3]). However, the direct communication between UEs and BSs has its own advantages, but has not been given enough attention in the literature. This contribution provides an extensive comparison between the relay approach and the less studied direct link approach, and also discusses various other aspects that are specific to HST scenarios.

The structure of the article is as follows. We explain general issues that arise in HST scenarios, and compare the relay and direct-link approaches. We then deal with further aspects of HST, showing dependencies between different parameters and discussing the system aspects from an operator's point of view. Conclusions are drawn in the final section.

SPECIAL TRAIN ISSUES

Wireless communication in HST scenarios exhibits several key differences compared with traditional considerations of a coverage oriented network. What is more, there is not the same amount of extensive experience with such scenarios as with classical networks. Thus, many effects have not been fully understood, and it is not clear which are the most significant in HST scenarios. The state of the art is discussed in the following, along with the aspects that are currently identified as most important.

The authors are with TU Wien, Austria.

[1] http://www.shift2rail.org

0163-6804/15/$25.00 © 2015 IEEE

Following [4], newly built tracks for high speed trains should support at least 250 km/h, and many operational systems exceed this value. These speeds lead to a high Doppler shift, which causes several transceiver impairments such as channel estimation errors and inter carrier interference (ICI) in orthogonal frequency division multiplexing (OFDM) systems.

Moreover, the channel characteristics along the tracks vary greatly. The surroundings show several distinct categories, which encompass certain different features that influence signal propagation. Among those are tunnels, trenches, cuttings, stations, viaduct-like structures, or bridges. An example of such a diverse environment is shown in Fig. 1. An extensive description of such categories is described in [5]. In addition to the onerous propagation conditions, a large part of the signal may be shielded by the metal chassis of the carriage. If the signal is first received by a relay, this can lead to a very strong multi-path component in addition to the line of sight (LOS) path.

Another peculiarity of HST scenarios is the UE distribution and movement. In contrast to classical assumptions of random UE placement, all UEs are concentrated within the train. Their location and movement is approximately deterministic as it is predefined by the course of the tracks and the speed of the train. This implicates a certain a-priori knowledge that can be exploited to compensate for some of the above mentioned issues.

The high UE concentration results in pulse-like traffic in the cells along the tracks. This leads to BSs being either completely idle or confronted with a high load (in case the hardware is dedicated to serving the train) or to large performance fluctuations in a cell (in case the surrounding environment is also covered by the BS). From the high speed of the train it follows that small cells are traversed in a very short time, thus leading to a frequent necessity for handovers. When all UEs aboard a train simultaneously require a handover to the next cell, the control channel easily gets congested and suspends UEs from associating with the neighboring cell. As indicated above, the deterministic UE behavior may provide a considerable advantage for implementing elaborated handover schemes.

It is commonly agreed that a satisfactory quality of service can only be achieved by employing dedicated hardware in an HST scenario, especially in the form of remote units (RUs). Due to their smaller size and cost compared to classical BSs, it is possible to install them close to the tracks and in great numbers. Several RUs can be connected to one BS via radio over fiber (RoF). This allows for very flexible routing options, as explained in the subsequent section. An example of an RU installation along the tracks is presented in Fig. 1. In the remainder of this work we will assume such a scenario for our considerations.

Due to the diversity of environments, the modeling of the fast fading channel becomes a complex task. Many considerations regarding wireless communication in HST scenarios are based on simulations applying the Winner channel model [6]. Even though the model is highly adjustable, it was not originally intended to represent the characteristics of a HST scenario. Furthermore, it does not reflect the dynamic changes in the channel characteristics from one category to another. Thus, such simulations only yield first-order statements that may considerably deviate from reality and may not appropriately reflect the specifics of the environment. 3GPP has recently introduced a 3D channel model in [7]. However, specific aspects for train communications are not yet included and need to be determined.

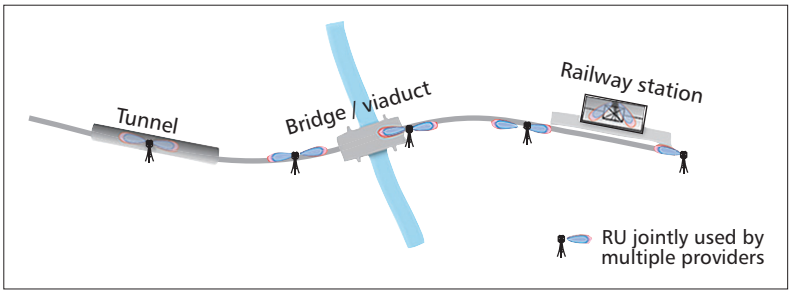


Figure 1. Various environments along railroad track. The track is supplied by RUs that are used by multiple network operators.

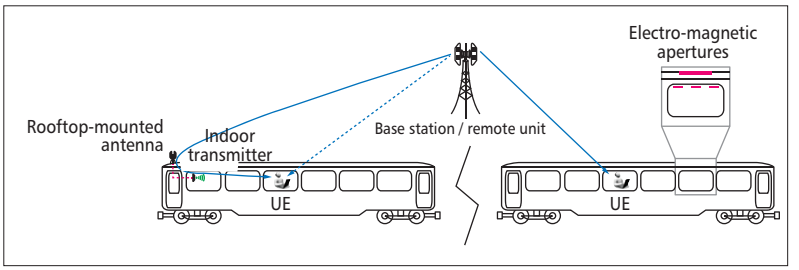


Figure 2. Signal reception with relay- and direct-link approach. Zoomed view indicates possibilities for applying electro-magnetic apertures.

Relay vs. Direct Link

In this section we consider the downlink direction exclusively. Similar considerations are valid for the uplink direction. Generally speaking, there exist two opposing approaches to provide wireless communications to passengers of an HST. In the first case, the UE directly associates with the BSs along the tracks, while in the second case this link is established via a relay, as shown in Fig. 2. Subsequently, a comparison between these two approaches is drawn and the advantages and drawbacks of both approaches are discussed.

Relay Approach

In the relay scenario, one or several antennas are mounted on the outside of the train. These are connected to one or more relays which are then distributing the signal inside the train. This approach has the major advantage that the signal is not attenuated by the windows of the carriage. Moreover, with this setup the relay can be configured such that it appears as a single UE to the BS, thus significantly reducing the number of handovers. Therefore, all traffic is aggregated by the relays and then distributed to the UEs.

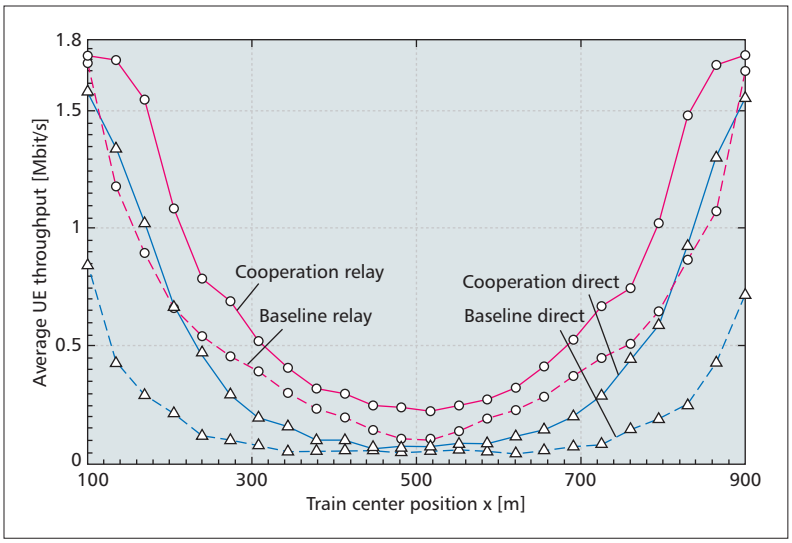


Figure 3. Train average throughput [Mbit/s] versus train center position [m]. Curves refer to results for baseline- and cooperation scheme among RUs for relay- and direct-link setup.

While the quality of the experienced link might be considerably improved, the employment of relays does not come without cost. First, relays need to be licensed for the specific band they are operating on. This is of minor consequence when no borders are crossed. In smaller countries, which is the case in most of Europe, one train connection can easily span three or more countries. For each country, the relays need to be registered individually or else they need to be switched off, which leaves the system in a direct-link state. Another consequence is that the employment of carriages becomes restricted to the countries/routes for which the relays on board are licensed and thus becomes less flexible.

A second issues appears with the choice of the frequencies for the RU-to-relay and the relay-to-UE connection. If the same frequency is used on both link sections, thus only bypassing the penetration loss of the carriage, the UE might still receive a considerable amount of the desired signal by a direct link through the window. The relative receive power of these two links depends on the individual position of the UE and the actual penetration loss. Due to the latency caused by the relay, the signal might be perceived via several multi-path components that cannot be equalized.

The aforementioned problem is completely avoided when two different frequencies are used on both sections of the link. For example, a dedicated frequency of a mobile operator is used from RU to the relay and a second frequency, e.g. in the ISM band, is used to provide WiFi inside. This setup is frequently referred to in the literature [2]. Nonetheless, it only provides a data connection for the passengers. Since only the relay is visible as a single UE for the BS, the passengers are not accessible for mobility management.

The performance of this setup may also considerably depend on the number of antennas and relays per train/carriage. It must be scrutinized whether the relays should work individually (e.g. one per carriage) or if the received signals should be combined. This is also affected by the possibility of connecting all relays to all carriages. Since a cable connection will not be feasible in practice, near-field communication standards at higher frequencies than the traditional 6 GHz band (e.g. in the upper mmWave band) may be considered for this task.

Direct-Link Approach

The direct-link approach assumes a direct connection between RU and UE. In comparison to the above scenario, the signal does experience a severe penetration loss into the carriage in this case. As the chassis of the carriages is usually made of metal, the signal enters the train mainly through the windows. However, the penetration loss may greatly vary among window types, as they are mostly metal coated themselves. Attenuation values range from 20 dB to 40 dB for metal coated windows of a German ICE-train [8], but other types of train exhibit different values, e.g. as observed in [3] with a combined range of 10 dB to 40 dB.

Note that these values reflect the situation for current carriages in use. Since the interest among train operators is increasingly to provide the best quality of experience to their customers, the design of future trains is likely to be adapted to the demands of the wireless link. Among various options is the possibility of introducing windows with small penetration loss (omitting the metal coating). Another option is to include apertures in the chassis or the window itself, by incorporating materials that are more permeable for electromagnetic waves, e.g. carbon fiber materials. Examples of such apertures are indicated in Fig. 2.

In order to compare the relay and direct-link setups, LTE-A system level simulations were carried out. The system model comprises a scenario with four equidistantly spaced sites that are placed directly next to the train tracks (thus exploiting available railway infrastructure). Each site employs two RUs pointing in opposite directions along the tracks. RUs facing away from the train are not considered interferers. Two RU collaboration schemes are compared:

- Each site is connected to an individual BS (baseline scheme).
- The two dominant RUs are connected to the same BS (cooperation scheme).

A fully occupied German ICE-train with 460 passengers is regarded, with 10 percent of the passengers having an active wireless connection. For the direct-link setup a penetration loss of 30 dB is applied referring to the mean value of the aforementioned penetration losses. *No* penetration loss is considered for the relay setup. The simulation parameters are summarized in Table 1. Simulations were performed with the Vienna LTE-A Downlink System Level Simulator.[2]

Simulation results are shown in Fig. 3 in terms of average UE throughput [Mbit/s] over the center position of the train [m]. Since the throughput at the relay represents the cumulative throughput of all active users, the results for the relay setup were normalized to 46 (UEs) for a fair comparison. Dashed lines represent the baseline scheme, solid lines the cooperation scheme. Comparing the performance of relay (circular markes) and direct-link setup (triangular markers), employing relays improves the performance roughly by a factor of two. This gain is surprisingly low, considering that

[2] http://www.nt.tuwien.ac.at/ltesimulator (current version v1.8 r1375).

the UEs in the direct setup experience a penetration loss of 30 dBs, and the relays are operating under ideal conditions (e.g. no processing delays and no overhead). Taking into account further aspects as discussed earlier makes the direct-link approach a notable alternative. From the curves it is further found that the performance can also be improved by advanced collaboration schemes. The throughput peaks in the vicinity of the base stations can be exploited by a sophisticated scheduler scheme. Regarding the traffic type, UEs with delay sensitive data could be assigned resources such that their requirements are fulfilled. UEs with best-effort traffic models could mostly be served when the train is closer to an RU and a higher total data rate is available.

While it is possible in the relay approach to let the whole train appear as the equivalent of one or two UEs (given the proper frequency setup), this option is not available for the direct-link approach. Each active passenger appears as an individual UE. Therefore, many handovers must be executed when the train moves from one cell to another. Considering the high speeds and that even "inactive" passengers, not actively transmitting data, must be handed over to the next cell, this can lead to a considerable amount of traffic on the control channel and might in extreme cases lead to blocking of UEs.

As mentioned above, the deterministic location of the train and the semi-deterministic location of the UEs (confined inside the train, but not known in particular) enables new concepts of handover and cell extension, including moving cell, smart handover schemes, and sliding handover.

The moving cell concept is shown in Fig. 4. For both approaches, relay and direct-link, it increases the cell length from around 1 km to a maximum of up to 50 km. Commercial systems with the necessary capabilities for that task are already available.[3] The working principle is that many remote units (RUs) are associated with a central control entity via RoF. When the train leaves the range of one RU, the central unit will reroute the data-stream toward the next RU along the tracks. In that manner, the cell "travels" with the train (in an abstract sense, this can be interpreted as a semi-static beamforming scheme), enabling a transparent and smooth transition, without the need for handovers over longer distances. Considering a speed of 250 km/h, the transition of a cell of 1 km length takes approximately 15 seconds and a handover becomes necessary. With the moving cell concept, the period of reoccurring handovers is enhanced to 12 minutes. The concept is further detailed in [9].

Smart handover schemes have been reported, e.g. in [10]. Such schemes improve handover performance, but assume cells simply touching each other on the outer cell borders. Nonetheless, they do not omit the issue that many handovers must be handled simultaneously when transitioning from one cell to the next. To further alleviate this issue, a possible solution would be to spread out the handover region wider than just between the two RUs at the cell borders. We call this a *sliding handover*. It is conceptually illustrated in Fig. 4. In the figure, it is observed that the cell overlap spans three RUs. When the train enters the handover zone, the control entity signals to a fraction of the UEs to perform a handover, while the others remain in the current cell. To find the optimal point for initializing the handover, the knowledge of the exact train position and the predefined trajectory (given by the tracks) can be exploited. In order to generate such an overlap region, the RU in the middle of the handover region is shared by both cell control centers. It can also transmit on different frequency bands and exploit coordinated multipoint (CoMP) schemes in order to avoid excessive interference in the handover zone. Fraction-wise handover avoids overloading the control channel. To further smooth the handover situation, the handover zone can be extended arbitrarily. This effectively leads to a

Parameter	Value
System bandwidth	20 MHz
Carrier frequency	2.14 GHz
Inter-RU distance	1000 m
eNodeB transmit power P_T	40 W
Antennas per RU	2
MIMO mode	CLSM
Path loss model	TS 36.942 'rural' [11]
Channel model	ITU-R Vehicular A [12]
Train speed	250 km/h
Receiver type	Zero forcing
Noise power spectral density	–174 dBm/Hz
Receiver noise figure	9 dB
Train length	200.84 m
Active UEs	46
UE distribution within train	Random, according to a uniform distribution
Antennas per UE	1
Traffic model	full buffer
Scheduler	Proportional fair
Channel knowledge	Perfect
Feedback	AMC: CQI MIMO: PMI and RI
RU backhaul connection	Radio over fiber, no delay

Table 1. Simulation parameters.

Many handovers have to be executed when the train moves from one cell to another. Considering the high speeds and that even "inactive" passengers, not actively transmitting data, have to be handed over to the next cell, this can lead to a considerable amount of traffic on the control channel and might in extreme cases lead to blocking of UEs.

[3] http://www.kathrein.de/indoor

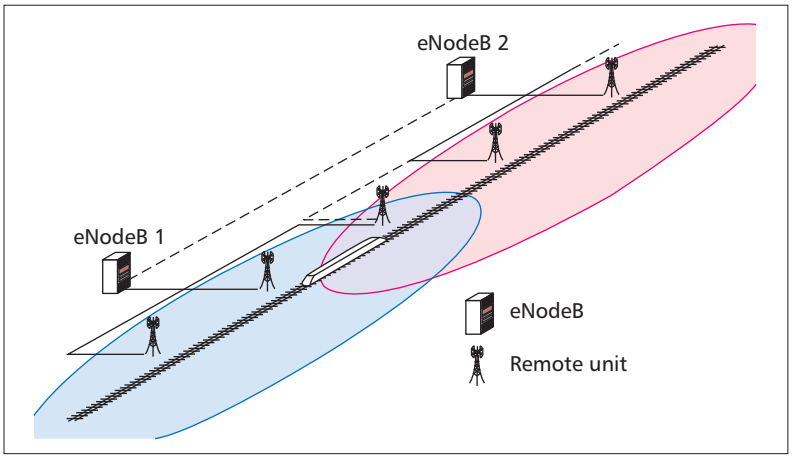


Figure 4. RU deployment along railroad tracks. Figure indicates moving cell- and sliding handover concept.

reduction of the total cell length (because of the longer overlap), which is, however, only a small fraction compared to the total length.

Further Aspects

The aspects of wireless communications in HST scenarios that have been discussed in the previous sections considerably affect many design choices for an actual system. The choice between the relay and direct-link approaches is aggravated by the fact that many of the presented aspects are interdependent, thus influencing each other. In this section these interdependencies are discussed from the perspectives of train operators and mobile-operators.

Technical Aspects

The interior of a train is characterized by many objects such as seats and baggage racks as well as passengers that reflect or absorb electro-magnetic waves. For the direct-link and the relay approaches, this has to be taken into account. The actual position of the UE might have an even greater impact than the characteristics of the outdoor link. In the relay approach, only the characteristics of the interior of the carriage need to be considered for the second part of the link. For the direct-link approach, a possible loss due to penetration through compartment walls needs to be considered. Such a path can even be dominant over one passing directly through the window closest to the UE, since penetration through windows might show a strong dependence on the angle of incidence. This has to be considered in the modeling of the direct-link channel. Possible shadowing of a train passing by in the opposite direction constitutes a further issue that might be included as a possible environment scenario, which occurs with low probability.

The selection of the carrier frequency also considerably impacts the overall system performance. The most relevant options are:

- Licensed bands (i.e. LTE-Advanced (LTE-A) frequencies).
- Unlicensed bands (industrial, scientific and medical (ISM) bands).
- Free bands in mmWaves.

On the one hand, licensed bands allow access to mobility management. On the other, distinct bands might experience a considerably different interference environment. In rural scenarios the interference from neighboring BSs can become a major issue, while for ISM bands, due to the limited equivalent isotropically radiated power (EIRP), and due to the lack of close interferers, the interference will be smaller. The limitation of the radiated power could be compensated with the deployment of very cost efficient RUs along the tracks, being spaced at very short distances. The application of mmWave is currently a hot topic for dense static deployment scenarios; however, their applicability for high mobility scenarios has yet to be scrutinized.[4] A possible application could be to radiate mmWaves from the tracks (i.e. from surface level) to the train base (assuming antennas are installed there).

The choice of a relay approach or a direct-link approach also strongly impacts the optimal placement and orientation of the RUs. While transmitter directions being aligned along the tracks work well in combination with relays, a positioning off the tracks with an antenna orientation perpendicular to the track-orientation can be beneficial for the direct-link approach, as it may overcome issues with a shallow angle of incidence, causing high penetration loss through the windows.

Operator View

Besides all the technical aspects, mobile operators will also consider economic factors such as CAPEX and OPEX, political issues, necessary agreements, and complexity of implementation. The two key players in a HST scenario are the train operators and the mobile operators, whose perspectives might considerably differ from each other.

For a mobile operator, the direct-link approach comes with the benefit of not having to rely on additional hardware being installed on the train. Thus, their system for supplying wireless access becomes more independent. The type of train only determines the expected penetration characteristics. Additionally, the mobile operator has all options for mobility management as opposed to a relay scenario with WiFi on the second part of the link. A train operator might also favor the direct approach, as they are not obligated to install and maintain additional hardware, but more importantly for not having to deal with any legal issues when trains are crossing borders. For both sides this approach comes with the benefit of reduced necessity to synchronize with each other.

Along with this first decision comes the question of the average performance that should be supplied to the passengers. Most importantly this affects the optimal placement and distance of the RUs. The total amount of RUs and connected hardware along the tracks (e.g. control entities, acquired sites) determines the CAPEX and OPEX of the whole system. Thus, RU spacing needs to be optimized to provide the required performance while minimizing the cost. Modifying the carriages might change the cost picture completely, thereby offloading some of the cost to the train operator. For example, installing windows with a low penetration loss or carriages with electro-magnetic apertures and employing the direct-link approach will make it possible to

[4] An example for an initiative, discussing the usage of mmWave in train scenarios, is IEEE 802.15 IG HRRC.

significantly increase the spacing of the RUs, thus substantially reducing CAPEX and OPEX.

An interesting question for mobile operators is whether the deployed hardware should also be used to supply the vicinity of the train tracks. From a cost perspective, this can be profitable but, simultaneously, it increases the complexity of network planning.

For train operators, it is beneficial to keep control over mobile communications when the task for supplying wireless access for their trains is not completely handed off to a mobile operator. On the one hand, the goal of mobile operators is mostly to achieve good coverage with minimal effort. Thus, they will focus on lucrative portions of highly frequented routes. On the other hand, train operators have the goal of providing coverage for the whole rail network, with a minimum performance guaranteed everywhere. Thus, the competitiveness of the train operator is consistently increased, as the additional service is provided comprehensively. Currently, there is growing interest among train operators to employ the hardware for mobile communications as well as for conveying control data for the trains. Since this aspect will be critical for the security of the system, it will be the operators' desire to ensure the *dependability* of the communication system, including aspects such as stringent latency constraints. From this perspective, many aforementioned arguments need to be reevaluated. For example, train operators typically do not buy their own licensed frequency bands. On the other hand, freely accessible bands such as the ISM bands come with the disadvantage of unpredictable interference, which is not feasible for security-relevant systems.

Conclusion

We discussed a variety of different aspects of HST scenarios in this article, with the main focus on the comparison of the direct-link and relay approaches. It was observed that the direct-link approach suffers from high penetration loss and a high number of simultaneous handovers, which can be overcome by utilizing materials with lower penetration losses (for windows and/or carriage walls), and by using sophisticated mobility management schemes such as a moving cell or a sliding handover. Moreover, with sophisticated scheduling schemes, delay-sensitive and best-effort traffic can be differentiated, depending on the train's position. If the technical issues with the direct-link approach can be overcome, it makes possible a less complex system setup and also avoids legal issues with additional hardware installed on the trains. Further considerations of train and mobile operators increase the complexity of the decision for specific system aspects, with cost issues and other factors coming into play, thus making such scenarios subject to multi-objective optimization.

Acknowledgements

This work has been funded by the Christian Doppler Laboratory for Wireless Technologies for Sustainable Mobility, and the KATHREIN-Werke KG. The financial support of the Federal Ministry of Economy, Family and Youth and the National Foundation for Research, Technology and Development is gratefully acknowledged.

References

[1] G. Tingting and S. Bin, "A High-Speed Railway Mobile Communication System Based on LTE," *2010 Int'l. Conf. Electronics and Information Engineering (ICEIE)*, vol. 1, Aug. 2010, pp. V1-414–V1-417.

[2] Y. Zhou *et al.*, "Broadband Wireless Communications on High Speed Trains," *20th Annual Wireless and Optical Commun. Conf. (WOCC)*, Apr. 2011.

[3] J.-Y. Zhang *et al.*, "A Multimode Multi-Band and Multi-System-Based Access Architecture for High-Speed Railways," *IEEE 72nd Vehic. Tech. Conf. Fall (VTC 2010-Fall)*, 2010, Sept. 2010, pp. 1–5.

[4] "Council Directive 96/48/EC of 23 July 1996 on the Interoperability of the Trans-European high-speed rail system," 23 July 1996, Council of the European Union.

[5] B. Ai *et al.*, "Challenges Toward Wireless Communications for High-Speed Railway," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, Oct 2014, pp. 2143–58.

[6] L. Hentil *et al.*, "MATLAB Implementation of the WINNER Phase II Channel Model ver1.1," https://www.ist-winner.org/phase 2 model.html, Dec. 2007.

[7] 3rd Generation Partnership Project (3GPP), "Study on 3D Channel Model for LTE," 3rd Generation Partnership Project (3GPP), TR 36.873, Sept. 2014.

[8] M. Uhlirz, "Adapting GSM for Use in High-Speed Railway Networks," Ph.D. dissertation, Institute für Nachrichtentechnik und Hochfrequenztechnik, Technische Universität Wien, 1995.

[9] C.D. Gavrilovich, Jr., "Broadband Communication on the Highways of Tomorrow," *IEEE Commun. Mag.*, vol. 39, no. 4, Apr. 2001, pp. 146–54.

[10] Z. Liu and P. Fan, "An Effective Handover Scheme Based on Antenna Selection in Ground-Train Distributed Antenna Systems," *IEEE Trans. Vehic. Tech.*, vol. 63, no. 7, Sept. 2014, pp. 3342–50.

[11] 3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) System Scenarios (Technical Specification TS 36.942), 3rd Generation Partnership Project (3GPP), Dec. 2010.

[12] ITU, Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000 (Recommendation ITU-R M.1225), International Telecommunications Union, 1997.

Biographies

Martin Klaus Müller [S] (mmueller@nt.tuwien.ac) received his Dipl.-Ing in telecommunications with highest distinctions from the TU Wien, Vienna, Austria, in 2013. He is currently pursuing the Ph.D. degree in telecommunications engineering with the Institute of Telecommunications, TU Wien. His research interests are supplying mobile access in train or highway environments and wireless communications in indoor scenarios.

Martin Taranetz [M] (mtaranet@nt.tuwien.ac.at) received the M.Sc. in telecommunications with highest distinctions from the TU Wien, Vienna, Austria, in 2011. He also received his Dr.-techn. degree (Ph.D. equivalent) in telecommunications engineering with highest honors from the TU Wien in 2015. In his dissertation, he focused on system level modeling and evaluation of heterogeneous cellular networks.

Markus Rupp [F] (mrupp@nt.tuwien.ac.at) received his Dipl.-Ing. degree in 1988 at the University of Saarbrücken, and his Dr.-Ing. degree in 1993 at the TU Darmstadt, Germany. From November 1993 to July 1995 he was a postdoc at the University of Santa Barbara, and from October 1995 to August 2001 he was a member of technical staff in the Wireless Technology Research Department of Bell-Labs at Crawford Hill, NJ. Since October 2001 he has been a full professor for digital signal processing in mobile communications at the Institute of Telecommunications of the TU Wien. He served as Dean from 2005–2007 and as Head of Institute from 2014–2015. He was associate editor of IEEE TSP from 2002–2005, and is currently associate editor of EURASIP's JASP and JES. He was elected an AdCom member of EURASIP from 2004 to 2012, and served as President of EURASIP from 2009 to 2010. He has authored or co-authored more than 500 scientific papers.

If the technical issues with the direct-link approach can be overcome, it makes possible a less complex system setup and also avoids legal issues with additional hardware installed on the trains. Further considerations of train and mobile operators increase the complexity of the decision for specific system aspects.

Automatic Train Control over LTE: Design and Performance Evaluation

Juyeop Kim, Sang Won Choi, Yong-Soo Song, Yong-Ki Yoon, and Yong-Kyu Kim

ABSTRACT

Due to technical advances in train control and wireless communications, unmanned train operation has gained in popularity of late. On the other hand, any errors involved in managing the QoS of train control traffic will cause negative consequences such as possible loss of life. Operators therefore naturally wish to scrutinize the specifications so that the wireless communications system is capable of guaranteeing the QoS of the train control traffic. In this article, we propose a feasible QoS management scheme for train control traffic based on the methodology used in a conventional LTE system. Based on the proposed scheme, we evaluate the feasibility of the LTE system using a testbed built in a commercial railway region. The key issues to support the train control services by the LTE system are the design of a QoS policy based on analyzing the characteristics of the train control traffic and the appropriate adjustment of the cell parameters during the cell planning and optimization procedures in order to resolve any network issues that may cause problems with data pause.

INTRODUCTION

Train control systems (TCSs) are responsible for all kinds of instructions for controlling train services at the wayside and train side. TCSs have been studied and developed to automatically guarantee safety according to the availability of various control technologies. By the end of 1930, locomotive engineers were operating trains by watching the signal lights on the sides of tracks and then making decisions manually in a low-speed environment. As the operational speed of trains and the number of operating trains increased, there was an unmanageable risk of collisions caused by human error. To address this, additional safety measures such as automatic train stop using beacons or balises were deployed. This ensured that locomotive engineers could not exceed the maximum allowed speed. After 1980, guaranteeing safety in high-speed environments became of paramount importance, and many other supplementary devices were introduced to control trains automatically. Automatic train control decides the proper speed based on:

- Information provided by ground control through a track circuit or a loop cable
- The current status of the train
- The trackside environment and weather conditions.

More recent research on TCSs has been aimed at extending automatic train control in order to achieve the goal of unmanned operations [1, 2]. To this end, the most significant change was to apply wireless and information technologies to TCSs. Traditional railway communications systems, such as track circuits and transponders, have critical problems in terms of maintenance. To solve the problems, the communication-based TCS was proposed for metropolitan railways. In this TCS, information transfer within the track region was achieved through wireless LANs. For high-speed railways, the European Train Control System (ETCS), which uses the Global System for Mobile (GSM) system for wireless communications between conventional trackside devices, was commercialized in Europe in 2004. Recently, the Korea Radio-Based Train Control System (KRTCS) project established in 2010 was completed for the use in all kinds of railway environments including metropolitan and high-speed railways.

According to this technical trend, TCSs have become closely aligned with a wireless railway communications system. The most widely known system is GSM-Railway (GSM-R), which is currently used in conjunction with ETCS in the European commercial field [3]. Moreover, railway services have gradually become more technologically advanced, and the demand for data has continued to increase. In this circumstance, many started to expect Long Term Evolution (LTE) to provide all kinds of railway services including voice communication, push-to-talk, multimedia-based supervision, and maintenance data transfer as well as train control. Consequently, considerable research efforts have been devoted to LTE as the next generation of a railway communications system [4–9]. Specifically, the feasibility of LTE as railway communications in a system aspect was validated in [4–6], and various algorithms and protocols were proposed in [7–9] to bring performance improvement of railway communications.

Since the fundamental concept of spectrum usage in LTE is quite different from that in GSM,

The authors are with Korea Railroad Research Institute.

0163-6804/15/$25.00 © 2015 IEEE

it is necessary to consider additional technical issues for providing a train control service based on LTE. GSM, mainly for narrowband communications, partitions the spectrum into several channels and allocates a channel to each service type. In contrast, LTE uses the entire wideband spectrum for all kinds of traffic, and data packets of various service types can be mixed in the spectrum. One of the critical issues in LTE involves guaranteeing quality of service (QoS) for each service. Especially, train control traffic includes vital information, and violation of the QoS for traffic will cause a problem for train service, which may threaten human safety. Therefore, LTE must be particularly concerned with guaranteeing the QoS of train control services. Conventional research including [10, 11] has considered QoS issues of voice and data services in LTE, but there is little research dealing with those of a TCS in the aspect of network management.

The goal of this article is to show the feasibility of LTE to serve TCSs in a practical environment. We aim to provide the parameter design of an LTE-based railway communications system for guaranteeing the QoS of train control services. Based on the architecture for QoS management in LTE, we provide a procedure for designing QoS parameters for TCSs. We then provide a verification procedure to validate whether the train control services work in practical scenarios based on a testbed built in a commercial railway region, and present performance evaluation results obtained from the testbed.

Train Control System Model

In this article, we assume that KRTCS is used as a TCS. Figure 1 shows the structure of the wayside and onboard systems in KRTCS. It is composed mainly of the following three components: automatic train supervision (ATS), automatic train operation (ATO), and automatic train protection (ATP). ATS is used to supervise the overall status of the train services at the wayside and perform remote control. ATO is used for the overall operation control in trains. ATO controls the train speed and train stops at regular positions at a station, issues commands to open and close the doors, and controls supplementary devices in stations. In addition, wayside and onboard ATPs perform real-time train positioning and decide moving authority so that a safe distance between consecutive trains can be maintained.

To accomplish its own mission regarding train control services, each of the components in KRTCS communicates with its corresponding component through wired and wireless links over IP. Specifically, ATS communicates with ATO to obtain a train status report and issue commands for various operations. Wayside and onboard ATPs communicate with each other to share the train status and perform various ATP operations regarding train movement. It is noted that the data communications between wayside and train side must pass through the air interface, which is covered by an LTE system in this article.

ATP traffic should be managed carefully because ATP plays a significant role in the safe movement of trains. The main mission of ATP is to move a train safely in a forward direction.

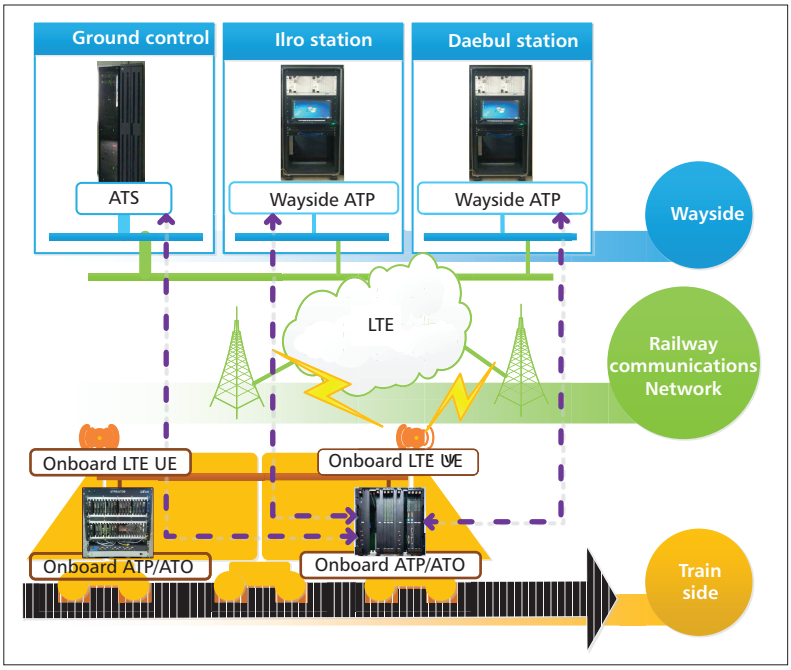


Figure 1. Overall structure of KRTCS.

Specifically, based on the train positioning report from the onboard ATP, the wayside ATP or ATS issues a moving authority, indicating a moving range in which the train is allowed to proceed, to the onboard ATP. The train should speed down or trigger the brake if it is about to reach the edge of the moving range. In addition, a wayside ATP can cover a limited area of a specific region, and handover is therefore performed between wayside ATPs. This procedure causes a train to be served seamlessly when the train crosses the coverage boundary of neighboring wayside ATPs. To accomplish their missions, onboard and wayside ATPs generate traffic and transfer to each other. Otherwise, problems will occur in performing their operation.

Analysis of KRTCS Traffic

To define QoS parameters suitable for KRTCS, it is necessary to analyze the characteristics of KRTCS traffic. Traffic from ATS and ATO are usually bursty at specific moments, such as on station entry. However, these types of traffic are generated rather infrequently and occur when the train is in the region of a station, in which the signal received from an LTE base station is in good condition. On the other hand, wayside and onboard ATPs generate their traffic continuously and periodically while the train is moving over the railway. Thus, ATP traffic can be generated anywhere in the region along the track and at any moment that the received signal environment is not be suitable for wireless communications due to geographical variations. Due to the fact that ATP traffic is the bottleneck in terms of guaranteeing QoS, it is important to analyze the characteristics of ATP traffic for grasping the QoS requirement of KRTCS.

In general, the amount of ATP traffic generated by a single train is maximally 50 kb/s for the downlink and 20 kb/s for the uplink. Because railway operators should make use of dualization

QCI	Resource type	Priority	Packet delay budget	Packet loss rate	Railway services
1	Guaranteed bit bate (GBR)	2	100 ms	10^{-2}	RailVoice dedicated, ATPControl, ATOControl dedicated
2		4	150 ms	10^{-3}	
3		3	50 ms	10^{-3}	
4		5	300 ms	10^{-6}	
5	Non-GBR	1	100 ms	10^{-6}	ATPControl, ATOControl default
6		6	300 ms	10^{-6}	
7		7	100 ms	10^{-3}	
8		8	300 ms	10^{-6}	
9		9			Internet default, RailVoice default

Table 1. QoS class identifier table.

on wireless sections to improve link availability, it is necessary to take into consideration the fact that twice the ATP traffic indeed occurs in the radio access layer. In addition, when handover between wayside ATPs takes place, onboard and wayside ATPs generate duplicate traffic. Therefore, the total amount of ATP traffic is up to 200 kb/s for the downlink and 80 kb/s for the uplink.

Similar to a voice service, data generated by wayside and onboard ATPs has a real-time property, and must be transferred to the other side within a specific time. This implies that delay is an important QoS parameter for ATP traffic just as it is for voice traffic. According to Annex C in [2], the maximum allowable transfer delays of train-to-wayside data and wayside-to-train data is 0.5–2 s. Based on the requirement, it would be reasonable to set the QoS parameter of the packet delay budget to 500 ms.

In the aspect of the theoretical cell capacity, it is sufficient for an LTE system to guarantee the above QoS requirements. However, it still matters that the LTE system guarantees the requirements strictly at any time. Because of its vital functionalities, ATP does not allow the violation of QoS at any point. This indicates that a service drop event should not occur throughout vast regions of service, including in tunnels and under bridges. Since it is desirable that packet loss be rare for ATP traffic, a higher priority must be allocated to the transmission of ATP traffic. From this aspect, QoS management for ATP should be differentiated from that for a voice service. Specifically, a voice service allows a certain amount of data pause in the aspect of service continuity, whereas ATP does not allow a short data pause.

QoS Guaranteed Framework Design for KRTCS

Preliminaries on the QoS Class Identifier

Since KRTCS traffic has unique characteristics and differs considerably from voice, video, or background data traffic, it is necessary to define a new set of QoS parameters dedicated to the KRTCS traffic. In general, the LTE standard provides a guideline for QoS parameters for efficient QoS management. The LTE standard defines typical QoS parameters and provides sets of predefined values of the QoS parameters for frequently used services. These predefinitions relax the complexity of implementation, and allow the functional entities of an LTE system to be optimized to guarantee the QoS based on the QoS parameters. In other words, it is hard to apply new and customized QoS parameters or modify the value of a QoS parameter in practice. For QoS management of KRTCS in a practical sense, we utilize the present status of the feasible QoS parameters in the LTE system.

Typical QoS parameters dealt with in LTE are resource type, priority, packet delay budget, and packet loss rate. Resource type indicates whether the minimal bit rate is guaranteed or not. The bearers with guaranteed bit rate have the minimal bandwidth such that the system is guaranteed in any network condition. Priority indicates the priority level for a bearer, and is applied during bearer establishment or modification. Packet delay budget is the upper bound on the end-to-end delay in the LTE system, and this affects scheduling operation in a base station. Packet loss rate means the maximal rate of packet loss at the link layer level.

In fact, the most important factor in QoS is the QoS class identifier (QCI), which represents the set of QoS parameter values, and is commonly used throughout the LTE system. Detailed QoS parameters are given in Table 1, which is defined in [12]. In Table 1, it can be deducted that QCIs 1 and 5 are suitable for KRTCS from their strict QoS parameters. Specifically, QCI 5 aims to transfer traffic that is sensitive to data loss. Thus, the packet loss rate of QCI 5 is set to be extremely low, and the priority of QCI 5 is set to be the highest. Due to those characteristics, QCI 5 is usually applied to bearers for application-level signaling messages in a commercial LTE system, such as call setup messages for voice services. However, QCI 5 does not guarantee a minimum bit rate per bearer, which is needed for providing a stable data rate to KRTCS traffic. On the other hand, QCI 1 is suitable for voice data. QCI 1 allows a certain amount of data loss, and requires a guarantee of a minimum bit rate per bearer. QCI 1 is usually applied to voice data bearers in a commercial LTE system. This QCI, however, allows packet loss in abnormal situations such as network congestion, which is not desirable for KRTCS services.

A Feasible QoS Management Scheme

The present status of the QCI table in LTE reveals that no QCI is perfectly suited to KRTCS traffic. An alternative way to support KRTCS

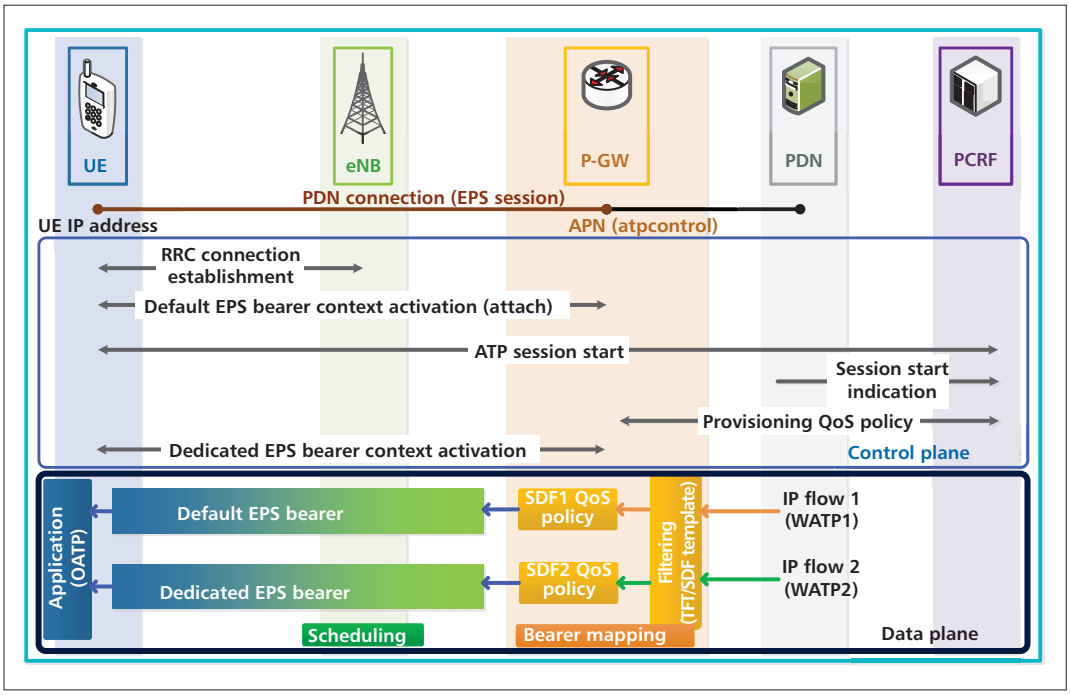


Figure 2. Architecture for guaranteeing QoS in an LTE system.

The LTE standard defines typical QoS parameters and provides sets of pre-defined values of the QoS parameters for frequently used services. These pre-definitions relax complexity of implementation, and allow the functional entities of an LTE system to be optimized to guarantee the QoS based on the QoS parameters.

traffic without modifying the QCI table is to use a mixture of QCIs 1 and 5. This can be achieved by making a KRTCS session have two data bearers of QCIs 1 and 5. The data bearer of QCI 1 carries most of the KRTCS traffic in a normal scenario. The data bearer of QCI 5 carries some significant messages, which require extremely low loss rate, or takes the role of the data bearer of QCI 1 when congestion or frequent data loss occurs.

Essentially, this method is based on a time sharing technique [13] in the information-theoretic sense. It is widely used to achieve a broad range of rate pairs given two achievable fixed rate pairs. The technical philosophy can fulfill the QoS requirements of KRTCS traffic in terms of delay, packet loss, and minimum data rate. In addition, this method uses network resources efficiently, because the data bearer of QCI 5, which has severe QoS parameters and is expected to use significant network resources, will be minimally utilized for transferring KRTCS traffic.

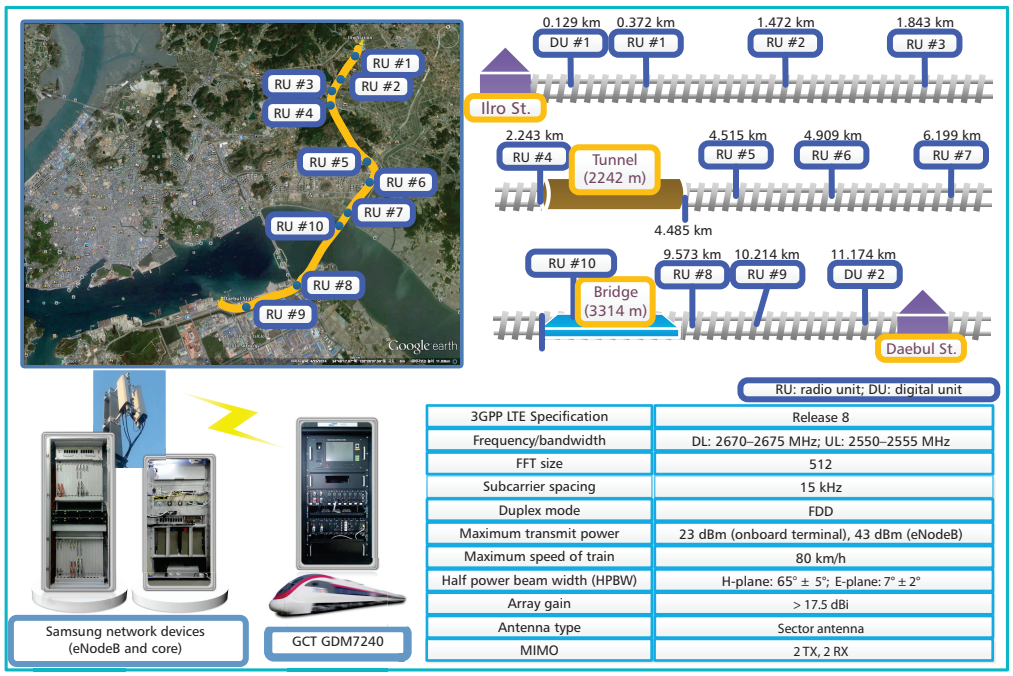
QoS Framework Design

Figure 2 provides the architectural view for guaranteeing QoS in an LTE system [14, 15]. As shown in [14], it is composed of an application server, a policy and charging rules function (PCRF), a packet data network gateway (P-GW), an eNodeB, and a user equipment (UE). The application server corresponds to ATS or wayside ATP. The PCRF contains sets of QoS parameters for each service and condition, and decides the QoS policy by forwarding the QoS parameters to the P-GW. The P-GW has the role of a gateway to the application servers, which are identified by an access point name (APN) in the LTE system. The P-GW leads to the management of data bearers according to the given QoS parameters. In addition, the eNodeB performs actual operations with respect to data packet transfer, such as scheduling, radio resource management, and mobility management, according to the QoS parameters of the data bearer [15].

The part of the control plane in Fig. 2 shows the procedure for starting a KRTCS session and configuring an Evolved Packet System (EPS) bearer, which is a data bearer between a UE and a P-GW. We allocated two independent APNs, *ATOControl* and *ATPControl*, to the ATS and wayside ATPs, respectively. During the train's power-on and attach procedure, a default EPS bearer of QCI 5 was configured for each APN. This default EPS bearer usually carries configuration and management messages of a KRTCS session. Also, this bearer can be fully utilized to transfer ATS or ATP traffic in case of network congestion.

Based on the status of the KRTCS session, PCRF plays an autonomous role in the dynamic decision of the QoS policy [12]. When the train service starts, the ATS and the wayside ATP generate sessions with ATO and ATP on the train side, respectively, and notify the PCRF. The PCRF then gives the QoS policy to the P-GW so that the P-GW configures a dedicated EPS bearer of QCI 1 for each APN. The ATS/ATO pair and the wayside ATP/onboard ATP pair then start exchanging their data with each other through the dedicated EPS bearers. The dedicated EPS bearer can guarantee a minimum data rate of KRTCS traffic for each train even in case of network congestion. In addition, network congestion may already occur before starting a KRTCS session, and there may be insufficient network resources to configure a new EPS bearer. In this case, the P-GW can attempt a preemption, in which existing EPS bearers of the lowest priority are released, and new EPS bearers are alternatively configured for

To carry out a field test in a practical railway environment, we built an LTE testbed in conjunction with a KRTCS testbed on a commercial railway. The testbed for KRTCS was built on the Daebul line in Mokpo, which is in southwestern South Korea.



3GPP LTE Specification	Release 8
Frequency/bandwidth	DL: 2670–2675 MHz; UL: 2550–2555 MHz
FFT size	512
Subcarrier spacing	15 kHz
Duplex mode	FDD
Maximum transmit power	23 dBm (onboard terminal), 43 dBm (eNodeB)
Maximum speed of train	80 km/h
Half power beam width (HPBW)	H-plane: 65° ± 5°; E-plane: 7° ± 2°
Array gain	> 17.5 dBi
Antenna type	Sector antenna
MIMO	2 TX, 2 RX

Figure 3. Cell deployment in the Daebul testbed.

the new KRTCS session. In addition, we set two more APNs, *VoIP* and *internet*, to provide voice and video services, respectively. For voice service, a default EPS bearer of QCI 9 is used for signaling, and a dedicated EPS bearer of QCI 1 is used for carrying voice data. The dedicated EPS bearer is configured at the moment the UE starts a voice session. It is estimated that allocating QCI 1 to a voice bearer would not affect the KRTCS data transfer significantly, because the amount of voice traffic is negligible in railway environments. On the other hand, since the amount of video traffic can be sufficiently large to cause congestion, we let the default EPS bearer of QCI 9 be configured for *internet*. This configuration enables the rate of video services per APN to be controlled.

Performance of KRTCS over LTE

Testbed Deployment

To carry out a field test in a practical railway environment, we built an LTE testbed in conjunction with a KRTCS testbed on a commercial railway. The testbed for KRTCS was built on the Daebul line in Mokpo, which is in southwestern South Korea. As shown in Fig. 3, the test site was 12 km long, and included a long (2.2 km) tunnel and a bridge. At this test site, five temporary stations were constructed to verify the operations in stations. With respect to a test train, we modified a commercial light train to be under control of onboard ATP and ATO. ATS was placed at the Ilro station, and wayside ATPs were deployed at the Ilro and Daebul stations, located at the end of the test site, so that inter-ATP handover could occur between the two stations. In addition, various supplementary devices, such as train closed circuit TVs (CCTVs) and screen doors, were deployed to assess the details of the unmanned operations.

For the LTE system, we used Samsung LTE Release 8 network devices with frequency band of 2.6 GHz. For the eNodeB deployment, we used 10 radio units and two digital units to cover the whole test site and connected them through backhaul. To extend the coverage to the tunnel and bridge regions, we positioned radio units at the ends of the tunnel and the bridge facing toward the middle of the region. The core network devices, including a P-GW and a PCRF, were placed at the Ilro station. For an onboard LTE UE, we used a GCT GDM7240 as an LTE modem chip. There were two onboard LTE UEs in the train for backup operation, and each onboard LTE UE was connected to a set of onboard ATP and ATO. Further details about system parameters in terms of the LTE testbed are described in Fig. 3.

To meet the QoS requirements of KRTCS across the whole railway service region, cell deployment was done systematically. Specifically, we endeavored to reflect the characteristics of KRTCS during cell deployment and optimization. Since the KRTCS service was affected by continuous data pause, we adjusted the cell parameters to minimize the occurrence and duration of data pause. For example, we reduced radio resource control (RRC) timer T310, which corresponds to an out-of-synchronization timer, to trigger cell selection quickly in that case. Also, RRC timers T300 and T311 are reduced such that the onboard terminal could quickly give up the on-going random access or connection re-establishment procedure followed by attemptting cell selection again. Unlike the commercial field, above setting is valid in a railway environment. It is because the received signal is strong in the

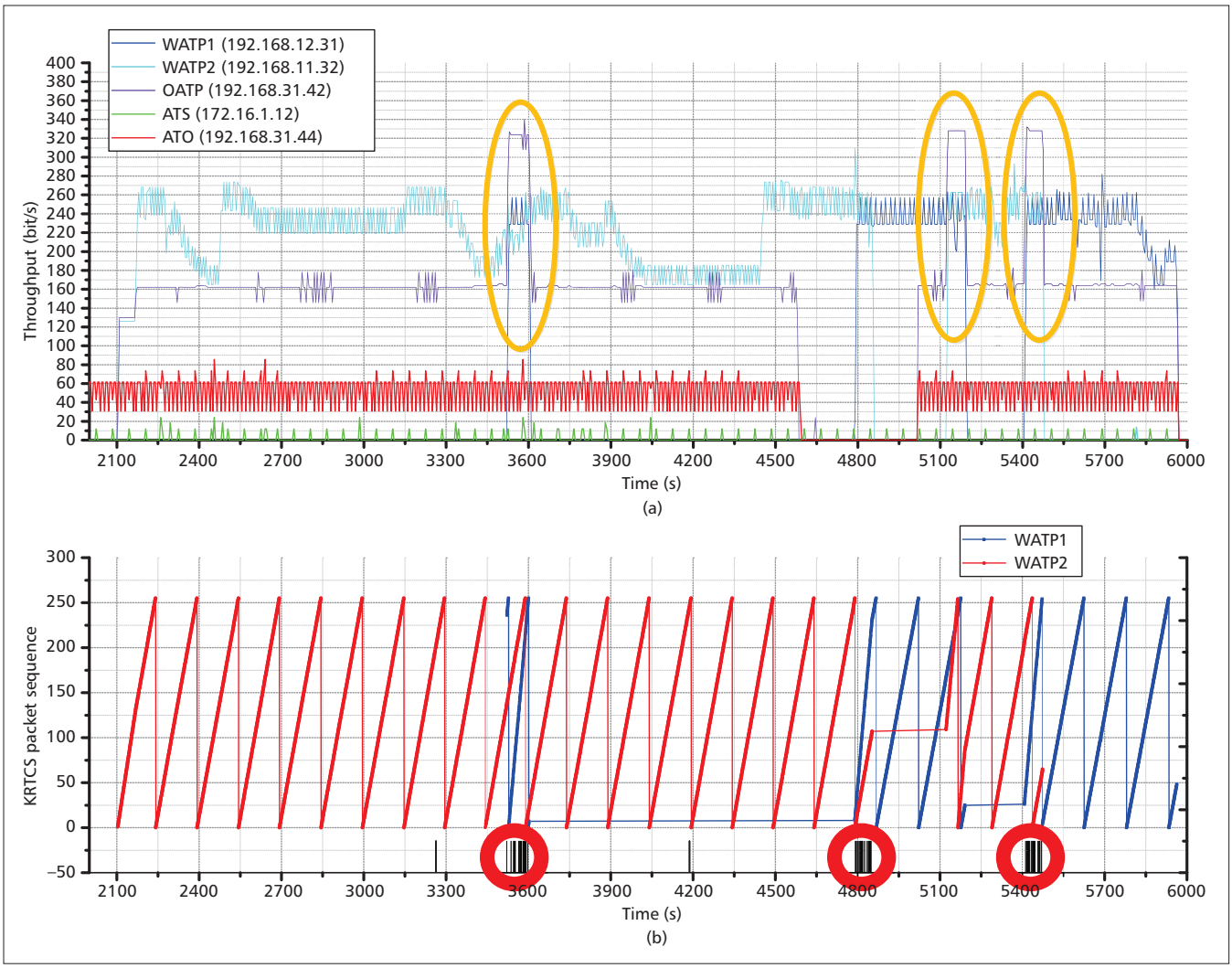


Figure 4. Observation of KRTCS traffic: a) observing KRTCS traffic generation; b) packet sequence trace.

track region and it takes short time to finish the cell selection or random access procedure for most cases.

Performance Evaluation

We evaluated 51 KRTCS test cases reflecting various unmanned operation scenarios that could occur on a commercial railway. The representative test cases include followings:

- Check if a train reports its position to wayside and gets a moving authority from wayside in time.
- Verify if a train is at the right position within a range of ±500*mm* with a probability of 100 percent, and within range of ±250*mm* with a probability of 90 percent when the train stops at a station.
- Validate whether inter-ATP handover procedure has been successfully completed without any loss of control signaling when a train crosses the boundary region.
- Verify whether the distance between two successive trains at a speed of 40 km/h is maintained or not.

During the test, we allowed the train to generate various background traffic such as video or voice data for railway operation and to transfer via the LTE system. This was for checking whether the QoS requirements of KRTCS could be satisfied preferentially while other services were ongoing.

The verdicts were passed for all the test cases with no service drop event. Figure 4 shows the snapshot of KRTCS traffic observed by Wireshark during the test. WATP1 and WATP2 indicate the wayside ATP in Daebul and the wayside ATP in Ilro, respectively. The graph in the upper part of Fig. 4 shows that the majority of the KRTCS traffic was ATP messages related to moving authority and train status reporting. In addition, the average throughput caused by the KRTCS traffic is much smaller than the capacity of an LTE cell, but the KRTCS traffic was continuously generated throughout the test. It implies that the LTE system should care for guaranteeing QoS of KRTCS traffic at any time regardless of where the train is and how the other traffic from CCTVs and voice services are generated. The graph in the lower part of Fig. 4 depicts the trace of packet sequence in time. This result ensures that the packet sequence increases continuously most of the time, and less than 1 packet was lost per second at some interval. This reveals that the LTE system serves

It is remarkable that the LTE system in the test-bed guarantees the QoS even in the region of problematic environments such as tunnel and bridge sections. Hence, we conclude that the LTE system is not only qualified to serve KRTCS but also can be a good candidate for serving other kinds of TCSs.

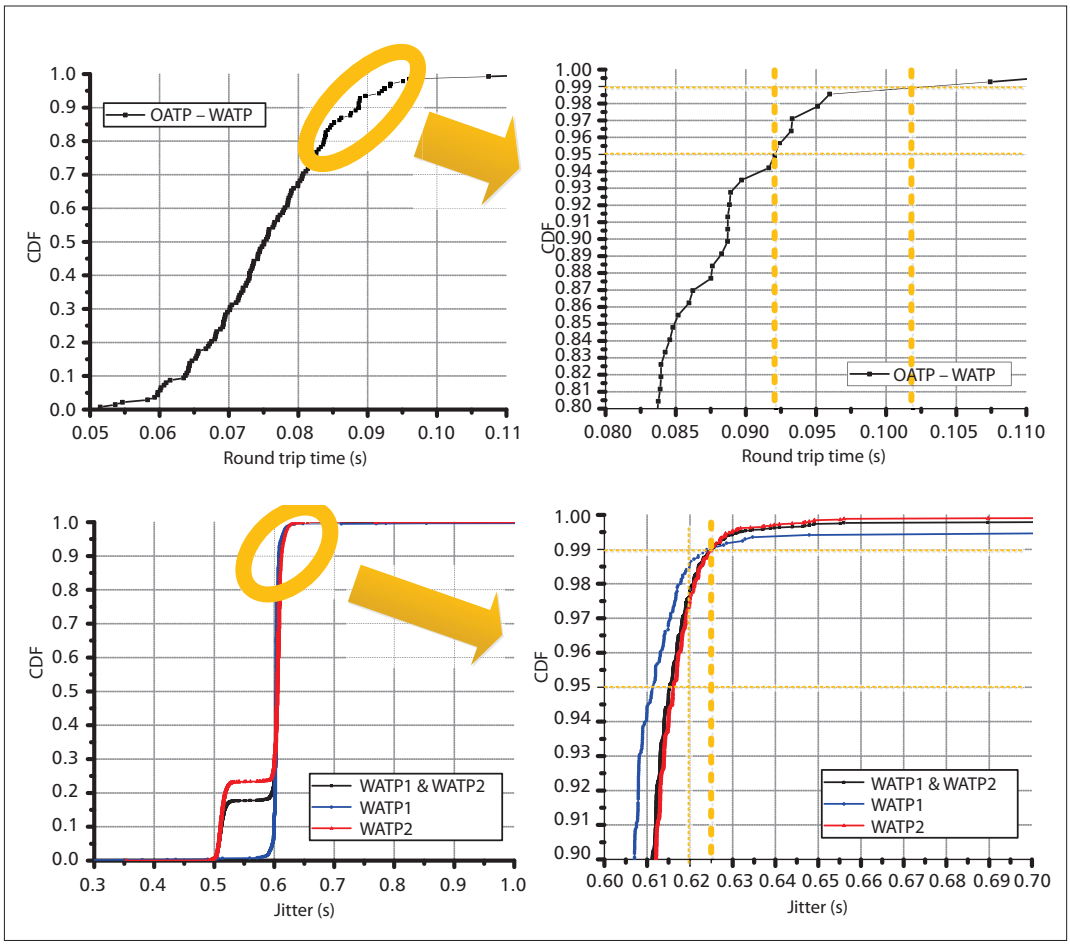


Figure 5. The performance results in terms of delay and jitter. Top: end-to-end delay; bottom: inter-arrival time.

KRTCS traffic without causing significant packet loss and service drop during the test.

Figure 5 shows more details about the performance of data transfer between wayside ATP and onboard ATP in terms of transfer delay. In view of end-to-end delay, we checked round-trip time (RTT). We measured transmission time of a request message from an onboard ATP and reception time of the corresponding response message at the onboard ATP. The result shows that RTT was about 75 ms on average and was less than 100–110 ms with 99 percent probability. This means that the end-to-end delay between the onboard ATP and the wayside ATP was typically 37.5 ms and mostly less than 55 ms. This result implies that the LTE system with our proposed scheme can fulfill the requirement of packet delay budget derived earlier. In addition, in view of jitter, we measured inter-arrival times of ATP messages, which are transmitted with a period of 600 ms. The result shows that the inter-arrival time was 610 ms on average and was less than about 620 ms with 99 percent probability. It is expected that the periodic messages generated by one side of ATP can be stably transferred to the other side of ATP through the LTE system so that periodicity of the ATP messages can be preserved at the reception side.

The performance results reveal that our proposed scheme can make the LTE system fulfill the QoS requirements of KRTCS traffic. It shows the feasibility that the traffic from ATS, ATO, and ATP can constantly proceed in the aspects of transfer delay and experience rare loss in the LTE system. Also, it shows the feasibility of the LTE system guaranteeing the QoS regardless of the train's operation and position. It is remarkable that the LTE system in the testbed guarantees QoS even in the region of problematic environments such as tunnel and bridge sections. Hence, we conclude that the LTE system is not only qualified to serve KRTCS, but also can be a good candidate for serving other kinds of TCSs.

Concluding Remarks

Many railway operators have recently stated a preference for unmanned TCSs for efficient management, but at the same time they are concerned about safety issues. In particular, errors in managing the QoS of the train control traffic could directly bring about a loss of human life. Consequently, operators wish to assess TCSs carefully and ensure that the wireless communications system is capable of guaranteeing the QoS of train control traffic. One of the key points in this article is to design a QoS policy based on analysis of the characteristics of the train control traffic. According to the traffic analysis, it is required to guarantee a minimum bit rate and low latency for each

KRTCS session. This critical issue was settled by allocating the two different functional EPS bearers with time sharing. Our proposed scheme is validated by the performance evaluation in terms of delay and jitter. The other key point is to adjust the cell parameters appropriately during cell planning and optimization procedures in order to resolve any network issues that may cause data pause. The performance in terms of packet loss was shown, while our testbed did not allow significant data pause and caused no service drop. Considering these key points will assure operators of the safety of unmanned train control over LTE.

Acknowledgment

This research was supported by a grant from the R&D Program of the Korea Railroad Research Institute, Republic of Korea.

References

[1] R. D. Pascoe and T. N. Eichorn, "What Is Communication-Based Train Control?," *IEEE Vehic. Tech. Mag.*, vol. 4, no. 4, Dec. 2009, pp 16–21.

[2] IEEE 1474.1-2004, "Standard for Communication-Based Train Control(CBTC) Performance and Functional Requirements," 2005.

[3] UIC CODE 951 v15.3.0, "EIRENE System Requirements Specification," 2012.

[4] R. Ivarez and J. Romn, "ETCS L2 and CBTC over LTE — Convergence of the Radio Layer in Advanced Train Control Systems," *IRSE Australasia*, Oct. 2013, pp. 1–12.

[5] A. Sniady and J. Soler, "LTE for Railways: Impact on Performance of ETCS Railway Signaling," *IEEE Vehic. Tech. Mag.*, vol. 9, no. 2, June 2014, pp. 69–75.

[6] J. Calle-Sanchez *et al.*, "Long Term Evolution in High Speed Railway Environments: Feasibility and Challenges," *Bell Labs Tech. J.*, vol. 18, no. 2, Aug. 2013, pp. 237–53.

[7] M. Cheng and X. Fang, "Location Information-Assisted Opportunistic Beamforming in LTE System for High-Speed Railway," *EURASIP J. Wireless Commun. and Net.*, July 2012, pp. 1–7.

[8] H. Gao *et al.*, "A QoS-Guaranteed Resource Scheduling Algorithm in High-Speed Mobile Convergence Network," *Proc. IEEE WCNC Wksp. '13*, Apr. 2013, pp. 45–50.

[9] J. Wang, H. Zhu, and N. J. Gomes, "Distributed Antenna Systems for Mobile Communications in High Speed Trains," *IEEE JSAC*, vol. 30, no. 4, May 2012.

[10] M. Alasti *et al.*, "Quality of Service in WiMAX and LTE Networks," *IEEE Commun. Mag.*, vol. 48, no. 5, May 2010, pp. 104–11.

[11] F. Capozzi *et al.*, "Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey," *IEEE Commun. Surveys and Tutorials*, vol. 15, no. 2, June 2013, pp. 678–700.

[12] 3GPP TS 23.203 v9.14.0, "Technical Specification Group Services and System Aspects; Policy and Charging Control Architecture," 2014.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 2012.

[15] 3GPP TS 23.401 v9.16.0, "Technical Specification Group Services and System Aspects; General Packet Radio Service (GPRS) Enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Access," 2014.

[15] 3GPP TS 36.300 v9.10.0, "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2," 2012.

Biographies

Juyeop Kim (jykim00@krri.re.kr) is a senior researcher in the ICT Convergence Team at Korea Railroad Research Institute (KRRI). He received his M.S. and Ph.D. in electrical engineering and computer science from Korea Advanced Institute of Science and Technology (KAIST) in 2010. His current research interests are railway communications systems, group communications, and mission-critical communications.

Sang Won Choi (swchoi@krri.re.kr) received his M.S. and Ph.D. in electrical engineering and computer science from KAIST in 2004 and 2010, respectively. He is currently a senior researcher in the ICT Convergence Research Team of KRRI. His research interests include mission-critical communications, mobile communication, communication signal processing, and multi-user information theory. He was the recipient of a Silver Prize at the Samsung Humantech Paper Contest in 2010.

Yong-Soo Song (adair@krri.re.kr) received his Master's degree in electrical engineering from Yonsei University in 2004. He has been with KRRI since 2004. He is working toward his Ph.D in electrical engineering from Yonsei University. His current research interests are in cell planning and handover in LTE railways.

Yong-Ki Yoon (ykyoon@krri.re.kr) received his M.Sc. degree in electrical engineering from Chungbuk National University, Republic of Korea, in 1996. He is currently a principal researcher in the metropolitan railroad system research center at KRRI. His current research interests are in communication based train control system, train position tracking method.

Yong-Kyu Kim (ygkim1@krri.re.kr) received his M.S. in electronic engineering from Dankook University, Korea, in 1987, and his D.E.A. and Ph.D. in automatic and digital signal processing from Institute National Polytechnique de Lorraine, France, in 1993 and 1997, respectively. He is currently an executive researcher inthe ICT Convergence Team at KRRI. His research interests are in automatic train control, communication-based train control, and driverless train operation.

According to the traffic analysis, it is required to guarantee a minimum bit rate and low latency for each KRTCS session. This critical issue was settled by allocating the two different functional EPS bearers with time sharing. Our proposed scheme is validated by the performance evaluation in terms of delay and jitter.

Cyber Security Analysis of the European Train Control System

Igor Lopez and Marina Aguado

ABSTRACT

One of the key research issues in further strengthening the role of railways in the transportation system is to achieve the highest possible level of cyber security against significant threats to the signaling and telecom systems. The European Rail Traffic Management System (ERTMS) was designed in the 1990s with the security measures and strategies available at the time. However, ERTMS' underlying communication technology, GSM-R, needs replacing, and replacement strategies are already underway. Due to their direct effects on safety, ERTMS security mechanisms also need updating to face current security threats. This article provides a security analysis of ERTMS' safety layer, Euroradio, in terms of current security threats. After identifying its vulnerabilities, we present four main recommendations: a more robust cryptographic mechanism, a new key distribution scheme, a new key storage and system integrity module, and a set of countermeasures for avoiding radio jamming attacks.

INTRODUCTION

Cyber security has been defined as a casualty of the transition from closed to open systems. In the industrial domain, migration from closed legacy systems to public data networks and advanced communication technologies has clearly introduced an enhancement in performance and resilience. However, such strategy also introduces a new challenge in that the communication networks require protection from current cyber security threats.

Extra challenging are information technology (IT) scenarios that deal with critical and vital services, such as vehicular and guided transportation scenarios. Services such as vehicular platoons, unattended train operation (UTO), virtually coupled train sets, and the signaling system itself clearly depend on the security and reliability of the underlying communication architecture.

The mass transport scenario is even more vulnerable to cyber attacks due to the predefined mobility pattern and the number of persons involved. The identification of this vulnerability, together with the need to foster and further strengthen the role of rail in the transportation system, have led to the following research question: How can we provide signaling and telecom systems with the highest possible level of cyber security in order to protect them against significant threats?

To help answer this question, in this article we focus on the most critical railway service, the signaling system, and more specifically the European Rail Traffic Management System (ERTMS), the standard for signaling and management systems in Europe. We provide an analysis of the potential vulnerabilities in current deployments, and offer security recommendations that can be considered for improvements in the short-term evolution. This analysis is timely since the system was designed back in the 1990s. Although the design took account of the proper security measures and strategies available at that time, threats and exposure have evolved since then.

The ERTMS is currently undergoing a transformation process. The main reason for this is that the underlying communication technology, Global System for Mobile Communications — Railway (GSM-R), a second generation mobile communication technology, has become obsolete. The current approach is therefore moving away from dedicated circuit-switched technology and toward advanced mobile technologies based on IP and shared medium access [1]. Likewise, the communication technologies, security policies, and strategies designed for the closed dedicated circuit-switched GSM-R technology in the 1990s — all specified in the Euroradio safety layer — need to be reviewed. Considering both current and future risk scenarios, this transformation provides us with an opportunity to perform a cyber security analysis.

Our contribution focuses on a cyber security analysis of the current ERTMS system, taking into consideration new risks that have emerged over the past two decades. This article is structured as follows: We first provide a taxonomy of current cyber attacks in the general IT domain. Next, we describe in detail the safety layer used in ERTMS as well as its different keys and its key distribution system. We then evaluate the protection mechanisms of the current ERTMS and the possible risks it faces in relation to the cyber attacks explained earlier. We present our different technologies that can potentially be used to update the current Euroradio safety

The authors are with the University of the Basque Country UPV/EHU.

0163-6804/15/$25.00 © 2015 IEEE

layer, as well as a set of accompanying recommendations.

Cyber Security Threats: A Taxonomy

This section provides a taxonomy of cyber security attacks in the general IT domain. These cyber attacks can be classified according to their interaction with the target, their goal, and the methodology used during the attack.

Passive Attacks

These types of attacks do not require any interaction with the target or the network under attack. Usually, they are difficult to detect, and their aim is to collect information for future, more complex, attacks.

Eavesdropping: This is the most common type of passive attack and is performed by capturing packets traveling along the network. This attack can be performed in both wireless networks and wired networks that are not correctly segmented.

Active Attacks

Unlike passive attacks, these attacks interact directly with the target or the network in order to cause intentional malfunctioning. These attacks are therefore more easily detected but also more dangerous.

Denial of service (DoS) attacks: As their name indicates, the goal of these attacks is to put the target out of service, usually by making the target work beyond its capabilities. Depending on how the attack is performed, DoS attacks can be classified into different groups. They can be categorized into physical and logical attacks, but they can also be defined according to the origin of the attack (i.e., whether the attack has one or multiple sources).

- Physical attacks (jamming): Jamming attacks do not require any logic or knowledge of the target and/or its network. The attack occurs when physical conditions that are able to interrupt communication are introduced into the network. For example, high-power electromagnetic emissions can abruptly reduce the signal-to-noise ratio (SNR) in the target's receptor.
- Logical attacks: The execution of these attacks requires exhaustive knowledge of the target's system, or of the network topology where it is located. Usually, these attacks — also called replay attacks — are based on the attacker copying valid packets of the service provided by the target and inserting these extra copies into the network. This action causes a data overflow in the target.
- Distributed attacks: When a DoS attack has more than a single source it is called distributed DoS (DDoS). DDoS attacks can be identical to regular DoS attacks, but by carrying out the attack from different sources, the probability of success increases considerably. Moreover, as there are multiple sources, it is more difficult for the target to recover from the attack.

Identity theft or spoofing attacks: These attacks permit the injection of packets into an unauthorized network by adopting the identity of an entity that is authorized by the network. When this identity theft is performed in both communication directions, it is called a "man-in-the-middle" attack; that is, attacker C alters communication between entities A and B by communicating with A as if it were B, and at the same time communicating with B as if it were A.

Equipment infection: The exploitation of known or unknown system vulnerabilities using viruses has become a common and effective means of cyber war. Cases such as StuxNet have demonstrated that incorrectly isolated industrial equipment can become a target. It is difficult to protect industrial equipment against these attacks for two main reasons. First, it is difficult to update such equipment in cases where the elements are geographically dispersed or are based on embedded systems. Second, protective software, such as antivirus software, has a negative effect on real-time performance.

Previously introduced security threats are common to any IT domain, including the critical and challenging vehicular-to-infrastructure domain. This domain includes car-to-infrastructure or train-to-ground communications. Previous research work [2] identifies the most severe security risks to the car-to-infrastructure communication architecture based on dedicated short-range communications (DSRC) technology and outlines possible countermeasures to the critical and major threats. However, train-to-infrastructure communication differs from car-to-infrastructure communication in several issues, such as number of entities involved, wireless technology used, and predefined behavior. Consequently, there is a compelling need to perform specific security risk analysis and countermeasures in this domain. We endeavor to present here this analysis on the most critical railway train-to-ground service, the train control system. The next section introduces the railway signaling system under study, ERTMS, and the security measures already present in its specification.

Train to infrastructure communication differs from car to infrastructure communication in several issues, such as number of entities involved, wireless technology used and predefined behavior. Consequently, there is a compelling need to perform specific security risk analysis and countermeasures in this domain.

The European Railway Signaling System, ERTMS, and the Euroradio Safety Layer

In the European railway signaling system, ERTMS, the Euroradio safety layer provides a secure data exchange by protecting delivery data in terms of authentication and integrity. This security is achieved by calculating a cipher block chaining-message authentication code (CBC-MAC). Additional protection is provided by the Automatic Train Protection (ATP) application working above Euroradio, which is responsible for protection against message delay, wrongly sequenced messages, message deletion, and message replay.

System Description

The CBC-MAC algorithm is based on the Triple Data Encryption Standard (3DES) and is defined in subset-037 v.3.1.0 of the ERTMS specification. The key material used in Euroradio is exchanged through a key management system (KMS) protected by different key materials. The ERTMS uses four different keys, which can be categorized in three levels. Table 1 summarizes

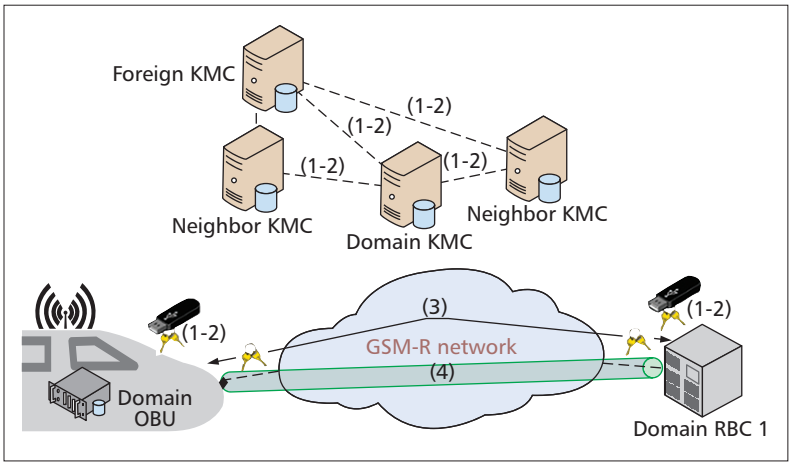


Figure 1. Current key distribution system in ERTMS: 1) KTRANS and K-KMC keys distribution; 2) KMAC keys distribution; 3) KSMAC derivation from KMAC; 4) safe communication using KSMAC.

this key material, including the entities that make use of each key and their functions.

Session keys, KSMACs, are generated from the authentication key material, KMAC, for each session. This procedure is described in detail in subset-037 and can be summarized in the following two steps:

- Exchange two random numbers in plain text between two entities.
- Perform the 3DES-CBC-MAC algorithm three times, using these two random numbers as seed material and three blocks of 64 bits taken from KMAC as keys.

Key Exchange

The key material in ERTMS is generated by the KMC, with the exception of KSMACs, which are negotiated for each session between ERTMS entities. Each ERTMS entity must have a valid KMAC shared with other ERTMS entities for establishing safe communication. In order to ensure the secure distribution of KMAC keys from the KMC to ERTMS entities and to other KMCs, transport keys (K-KMC and KTRANS) are used to provide confidentiality, authentication, and integrity. Half of the transport key is used for confidentiality by performing 3DES ciphering, and the other is used for authentication and integrity by calculating a CBC-MAC code.

The key distribution methodology, illustrated in Fig. 1, is based on messages defined in Subset-114 v.1.0.0 of ERTMS specification. These messages are exchanged through an offline mode, using physical storage devices such as USB sticks or CDROMs.

ERTMS Protection Mechanisms and Shortcomings against Cyber Attacks

This section considers the cyber security threats to current ERTMS security mechanisms based on the taxonomy of cyber attacks on the IT domain presented earlier. It is structured in two parts: the first part reports on those attacks initially considered in the ERTMS design phase, while the second part is concerned with the security threats that were not considered during the ERTMS design phase.

Analysis of ERTMS Security Mechanisms

The ERTMS faces security risks such as replay attacks and identity theft attacks at different levels. Such risks are counteracted in two main ways. On one hand, the introduction of timestamps in the ERTMS messages prevents replay attacks. On the other hand, the authentication and integrity of messages relies on cryptographic algorithms implemented by the Euroradio safety layer. Below, we analyze these mechanisms and algorithms in order to identify their strengths and weaknesses.

As mentioned above, to counteract replay attacks, the ERTMS introduces timestamps for sequencing the messages. However, this mechanism is not introduced during the session establishment process, making it vulnerable to such attacks. Additionally, since the sequence numbering using timestamps is made at the application layer level, the end-side must decrypt the MAC provided by Euroradio before checking the validity of the sequence number. This fact can be exploited to perform a flooding attack, resending several valid — but out-of-sequence — messages, which degrades system performance. Preventing such attacks depends on the confidentiality provided by the GSM-R encryption algorithm. However, this algorithm has already been cracked. Moreover, precomputed key tables, called rainbow tables and available on the Internet, allow an attacker to listen and even spoof entities in the same cell of the network. Thus, we can affirm that ERTMS is vulnerable to eavesdropping and replay flooding attacks due to GSM-R weakness and the current ERTMS protocol structure.

From the message integrity and authentication point of view, the potential vulnerability of ERTMS comes from two main factors: the vulnerability of the key material distribution and the weakness of the cryptographic algorithms used.

The key exchange methodology explained earlier deals with the security of the process. However, this offline process requires personnel to manually deliver the messages from the KMC to the ERTMS entities. Because this process is complex, there is a risk of simplifying it by using the same KMAC for large train fleets. As pointed out in a tender document [3] released by the Danish Railway Company, this fact has security and safety implications, because when many parties share a secret it is no longer a secret. Furthermore, the physical delivery of the key material introduces the possibility of attacks based on social engineering.

From the point of view of the cryptographic algorithm, critical vulnerabilities of DES, when it is used to compute CBC-MAC codes, have already been pointed out [4]. As Smith *et al.* demonstrated, DES is vulnerable to key-collision attacks based on the birthday paradox, which are more efficient than brute force attacks. Additionally, the use of 3DES does not introduce a much higher level of robustness compared to

DES, since it is still vulnerable to man-in-the-middle attacks [4]. Actually, the real robustness of 3DES is not higher than $O(2^{84})$ if 2^{28} cipher texts are available for the attack. This vulnerability is risky in terms of the authentication provided by the KSMAC. However, higher risk comes from two factors:

- The fact that multiple trains make use of the same KMAC for a long time
- The possibility of using weak random number generators during the KSMAC derivation

In fact, if the attack is performed against the session establishment with the goal of finding the KMAC, the whole system could be compromised: an attacker could take the identity of one or many trains during subsequent session establishments. As the KSMAC derivation from the KMAC is a public process and the random numbers travel in plain text, the effectiveness of the attack increases considerably.

Key name	Key size	Functions	Entities involved
Level 3: K-KMC	384 bits	Encryption, authentication and integrity	KMC-KMC[a]
KTRANS	384 bits	Encryption, authentication and integrity	KMC-RBC[b] KMC-OBU[c]
Level 2: KMAC	192 bits	Authentication and integrity	OBU-RBC
Level 1: KSMAC	192 bits	Authentication and integrity	OBU-RBC

[a] KMC: key management center; [b] RBC: radio block center; [c] OBU: onboard unit.

Table 1. Summary of key material used in ERTMS.

Security Threats Not Considered by ERTMS

The risks of suffering jamming attacks have traditionally been counteracted by isolating the railway networks. Using isolated circuit-switched networks and reserved frequency bands, the communication between ERTMS elements has been inaccessible to outsiders. However, the feasibility of these attacks now requires reconsideration on account of two factors: the popularization of commercial jammers working in the same band used by GSM-R, and the impending migration of ERTMS toward IP-based cellular technologies [1]. This security risk is common to other vehicular communication technologies, such as DSRC [5].

Table 2 summarizes the potential cyber attacks against ERTMS, their feasibility, and mitigation techniques of ERTMS for overcoming them. It also ranks security risks using estimated values for likelihood of occurrence and impact of each attack on the network. This risk analysis has been done using the Telecommunications and Internet Protocol Harmonization Over Networks (TIPHON) methodology published by the European Telecommunications Standards Institute (ETSI). When compared to the analysis presented in [2] for DSRC, we can conclude that the threat scenario is similar, and also that spoofing threats are the most critical ones. Unlike [2], we differentiate the feasibility of jamming and flooding attacks; therefore, the risk for the first one is critical, whereas the risk for the second one is major.

Potential Security Recommendations for ERTMS

With the aim of increasing the security of ERTMS and ensuring it is able to counteract attacks not previously considered, we propose four main security recommendations.

Cryptographic Algorithms

Any potential update of the cryptographic algorithms used in Euroradio will require choosing a substitute for 3DES, which involves two requirements. The first requirement is that the algorithm should be selected not only according to the current state of the art, but also with an eye to the future. For this reason, we should select the most suitable substitutes based on national and international security recommendations, pursuing validity after 2030. The second requirement is that the algorithm should fulfill the current functionalities demanded by the SFM: integrity and authentication. Thus, data confidentiality need not be evaluated because, due to the public definition of ERTMS messages, their encryption does not provide significant security improvement.

In our analysis of potential alternative crypto algorithms we consider three aspects: the cipher type, the integrity and authentication codes, and the key size and national recommendations.

Target	Attack type	Means	Knowledge needed	Detection capability	Mitigation technique specified in ERTMS	Occurrence likelihood	Impact	Risk level
OBU-RBC communications	Passive	Eavesdropping	Medium	Low	Dependent on GSM-R robustness	Possible	Low	Minor
	Active	Jamming	Low	High	Not considered	Likely	Medium: force degraded mode	Critical
		Spoofing	High	Medium	Authentication and integrity	Possible	High: wrong driving	Critical
		Flooding replay attacks	High	High	Responsibility of ATP application	Possible	Medium: force degraded mode	Major

Table 2. Review of potential attacks in ERTMS.

Standard	Validity date	Symmetric key size	Authentication and integrity algorithm
NIST [7]	>>2030	192	CMAC or SHA-3
ENISA [8]	>2030	128	CMAC or SHA-3
NSA [9]: Secret Top Secret	—	128 192	SHA-256 SHA-384
RFC 3766	2053	128 192	—

Table 3. National and international recommendations for cryptographic algorithms and key size.

1) Symmetric ciphers vs. asymmetric ciphers. In symmetric cryptography, since both parties involved in the communication must know a secret key, the weakness of the system depends on a correct and secure key exchange process. On the contrary, in asymmetric cryptography different keys are used for encryption and decryption, one of them public and the other private. Generally, this scheme is considered to be more secure and more suitable for communication between entities that do not know each other in advance. This is the case in broadcast communications, common in DSRC, where the targets of the message are any vehicle in the road. However, this is not the case in the railway environment, where the entities participating in the communication exchange are strictly controlled. Moreover, asymmetric schemes require more processing than symmetric ones. Thus, taking into account the real-time constraints, the optimal solution in the railway context is the use of symmetric cryptography for authentication and integrity. However, for the key exchange and negotiation processes, the use of asymmetric schemes is the most appropriate.

2) Authentication and integrity codes. Two of the most commonly used codes for providing authentication and integrity with a shared key are the CBC-MAC and hash-based MACs (HMACs). The first scheme uses a CBC process with a random initialization vector (IV), whereas the second is based on hash functions. The most popular algorithms that use both techniques are the CBC-MAC with AES and Secure Hash Algorithm (SHA-1) algorithms. It is worth noting that CBC-MAC-AES performs noticeably faster, especially when small packets are used —which is the case for ERTMS — due to the overhead generated by HMAC-SHA-1 [6]. Additionally, due to the collision vulnerability discovered in SHA-1, it has been recommended that it be substituted by the new SHA-3 algorithm. However, although this new algorithm is more robust, it generates even larger output than SHA-1 and therefore higher overhead in the communication.

3) Key size and national recommendations. Taking into account the computing improvement expected in the near future, national and international security agencies have published recommendations for choosing the optimal authentication and integrity algorithm and the most suitable key size. We summarize these recommendations in Table 3.

The key size determines the robustness of the secure communication. All recommendations made by security agencies state that after 2020, symmetric algorithms should not use a key shorter than 128 bits. Although an even longer key size may seem the best option, it is worth noting that an increment in key size implies an increment in processing overhead. As an example, a 192-bit key size represents an increment of 20 percent in the overhead compared to a 128-bit key size.

With regard to the crypto algorithm, the European Network and Information Security Agency (ENISA) [8] as well as the American National Institute of Standards and Technology (NIST) recommend the use of AES in cipher-based MAC (CMAC) mode. CMAC mode is defined in the document RFC 4493 and represents an improvement on the CBC-MAC mode since it solves CBC-MAC problems known to occur when using different size messages [7]. In contrast to the other agencies, the National Security Agency (NSA) recommends the use of SHA, but as discussed before, this algorithm is not suitable for railway signaling.

All in all, the most suitable option for updating the cryptographic algorithms of Euroradio appears to be AES in CMAC mode with 128-bit key size. From an efficiency point of view, a comparison of performance evaluations of DES, 3DES, and AES found in the literature [10, 11] indicates that AES works faster in software due to its design. Moreover, AES provides more security due to the larger block size and longer keys used, and it has no known cryptographic weakness.

Current MAC code in Euroradio is 64 bits long. It uses as input 64 data bit blocks and 64-bit-long keys. This MAC length would increase to 128 bits if — according to the national recommendations in Table 3 — the data block size and keys are increased to 128. Although this may seem a significant increase in message overhead, it is unavoidable for security reasons. Moreover, with the expected migration of access technology from GSM-R to IP-based broadband wireless technologies, the available bandwidth increases considerably, and this overhead increment from 64 bits to 128 bits can be considered negligible.

Key Exchange Systems

As mentioned above, the implementation of the offline key exchange system, defined in Subset-114, turns out to be rather complex in practical terms. In order to solve this problem, an adaptation of the specification has been presented by the Danish Railway Company [3]. This proposal is represented in Fig. 2 and introduces the key distribution center (KDC) element, which is responsible for distributing the keys generated by the KMC over the ERTMS entities. For secure distribution of this key material, the proposal suggests the use of transport layer security (TLS) combined with public key infrastructure (PKI) for validating the entities' identity.

To a large extent, this proposal meets the requirements of Subset-114, as it retains the KMC entity and KMAC and KSMAC keys. In terms of the key distribution process, this propos-

al substitutes the KTRANS and K-KMC keys with a public/private key scheme. Additionally, the use of TLS, which is based on International Telecommunication Union — Telecommunication Standards Sector (ITU-T) Recommendation X.509, means that the proposal fulfills the requirements defined by the European standard EN 50159-2 for safety-related communication in open transmission networks in the railway domain.

The scheme in Fig. 2 suggests that it is worth evaluating the real need to keep the KMS and the KSMAC derivation once we already have a public/private key scheme shared between ERTMS entities. In other words, could we simplify this scheme by introducing common security architectures used in commercial services such as e-commerce?

In Fig. 3, we simplify the scheme by introducing a distributed key negotiation scheme based on hybrid cryptography. As illustrated in Fig. 3, the KMC takes the role of a certification authority (CA) in PKIs, and its main duty is the generation of valid certificates that will guarantee the identity of the ERTMS entities. In this scheme, we propose the use of a hierarchical trust scheme with a common root CA for a common area (e.g., a single root CA for the whole of Europe) and multiple CAs in each country with a parent/child relationship. This design eases the federation of new KMCs or CAs as needed. The positive train control system (PTC) [12] and DSRC [5] standards propose a similar trust scheme based on CAs.

Both the public/private key pair and the accompanying certificate generated by the CA are delivered from KMCs to ERTMS entities using physical storage devices as in all previous schemes. However, once these keys and certificates have been correctly installed, and as long as the certificate has not expired, the ERTMS entity is able to renew both the key pair and the certificate using secure online procedures. However, it is worth noting that, due to the robustness of asymmetric cryptography, this cryptographic material could be valid for years before it needs updating.

The main difference between this scheme and the previous one is that the session key derivation has been updated. Using the public/private scheme supported by TLS, a session symmetric key is negotiated and exchanged, removing the risks found in the derivation process. As shown in Fig. 3, the number of messages necessary for establishing a secure connection is lower than that of other proposals. This implies two advantages: the whole process is faster, and the key material is less exposed to intruders.

Key Storage and System Integrity

In industrial security it can be very difficult to maintain the system integrity of the components connected to the network. If physical access to these elements is not correctly secured, no matter how secure the key distribution process is, the key will be accessible to an attacker, and the system will be vulnerable. Thus, it is important to store the keys internally and limit physical access to the system e.g. by removing physical connection ports accessible to potential attackers. To ensure the safe storage of asymmetric keys and certificates, a Trusted Platform Module (TPM) has recently been introduced in many hybrid systems. The TPM can be used to generate a TLS session's symmetric key within the hardware of the module, isolating the software-based attacks from the key negotiation procedure. TPM is also able to check the software integrity of the system and will report the introduction of any unauthorized software in the element, reducing the risk of malware or software corruption. The use of this technology has already been proposed for DSRC in order to prevent malware [2].

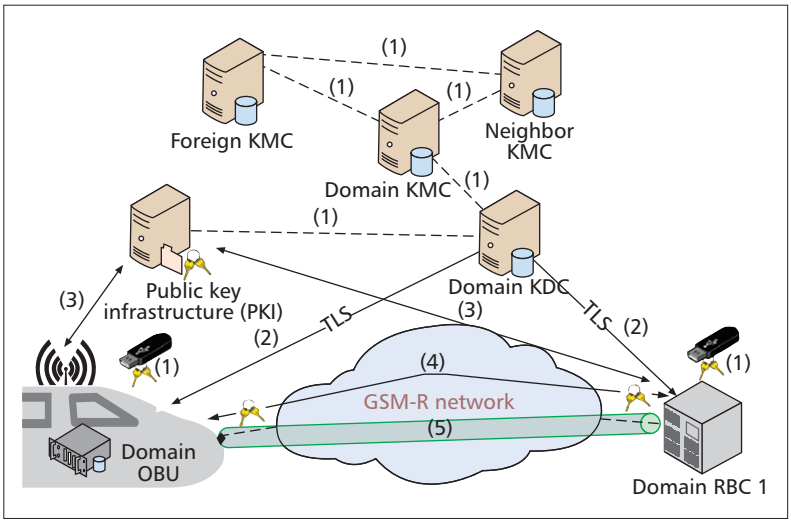


Figure 2. Danish proposal for an online key management system [3]: 1) KTRANS and K-KMC are replaced by a private/public key pair; 2) KMAC keys distribution using TLS with the private/public key scheme; 3) the identity of the parties is checked using a public key infrastructure scheme; 4) KSMAC derivation from KMAC; 5) safe communication using KSMAC.

Jamming Countermeasures

When addressing possible jamming attacks, it is important to detect them with as little delay as possible, and to design countermeasures to overcome them in real time. A real-time detection system for jamming attacks has already been presented for vehicular platoons [13], an intelligent transport system (ITS), which is most similar to the railway system. For overcoming a jamming attack, proposals have already been made for spread spectrum techniques for communications-based train control [14] or multipath techniques with heterogeneous bearers [15]. As shown in Table 2, this type of attack requires a low level of knowledge and can affect the availability of the railway service. Thus, future revisions of ERTMS should include techniques to counteract jamming.

Conclusions and Recommendations

In this article, we have provided a security evaluation of the European train control management system, ERTMS. This system was designed two decades ago in the 1990s. Since then, new risks have emerged; also, an ERTMS evolution

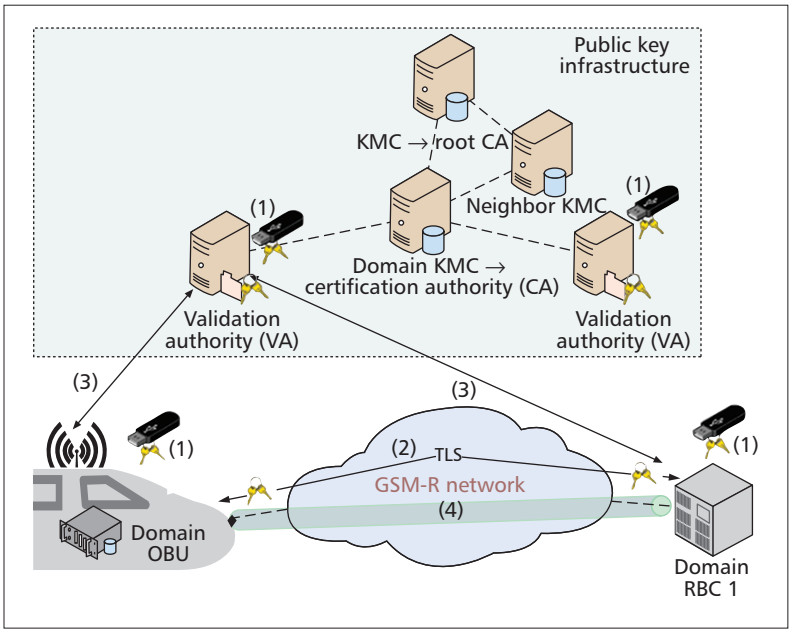


Figure 3. Proposal for an online key management system based on e-commerce systems: 1) KMAC keys are replaced by a private/public key pair; 2) the KSMAC symmetric key is negotiated during a TLS session establishment protected by private/public keys; 3) the identities of validated ERTMS entities are checked using a public key infrastructure scheme; 4) safe communication using KSMAC.

toward IP is currently on its way. Thus, this evaluation is timely and necessary. As a contribution, we have provided a review of potential attacks on ERTMS, the mitigation techniques specified in ERTMS, and their risk level according to the ETSI threat analysis methodology. We have also presented a set of viable and necessary security recommendations in four areas: a more robust cryptographic mechanism, a new key distribution scheme, a new key storage and system integrity module, and a set of countermeasures for avoiding radio jamming attacks. We recommend the use of an AES-CMAC algorithm with a 128-bit key for message authentication and integrity, and a key distribution scheme based on TLS and PKI infrastructure with federated CAs. We also recommend the inclusion of jamming detection and countermeasure techniques in ERTMS, as well as a TPM module for key storage and system integrity. Our proposal meets the requirements of the European standard EN 50159 for safety-related communication in open transmission networks in the railway domain and is in line with the European ENISA's recommendations for safe industrial communications.

Acknowledgment

The work described in this article was produced within the Training and Research Unit UFI11/16 funded by the UPV/EHU. This work was supported by the UPV/EHU Zabalduz program, an initiative hold within the EUSKAMPUS project, the International Campus, to promote the training of a research task force integrated within our regional industry and in cooperation with CAF (Construcciones y Auxiliar de Ferrocarriles).

This work was also supported by the Spanish Ministry of Economy and Competitiveness through the SAREMSIG TEC2013-47012-C2-1-R project (Contribution to a Safe Railway Operation: Evaluating the Effect of Electromagnetic Disturbances on Railway Control Signalling Systems) and funded under the call Programa Estatal de Investigación, Desarrollo e Innovación and oriented toward Retos de la Sociedad 2013.

References

[1] F. Pujol and J. S. Marcus, "Evolution of GSM-R. ERA/2014/04/ERTMS/OP. Final Report," Apr. 2015.
[2] C. Laurendeau, M. Barbeau, "Threats to Security in DSRC/WAVE," *Ad-Hoc, Mobile, and Wireless Networks*, 2006, pp. 266–79.
[3] Banedanmark Signalling Programme. "Online Key Management System Concept," Appendix 3.1 Att 11, SP-12-041120, Denmark, 2011.
[4] E. Biham, "How to Forge DES-Encrypted Messages in 228 Steps," Technion Comp. Sci. Dept. tech. rep. CS0884, Technion, Aug. 1996.
[5] J. B. Kenney, "Dedicated Short-Range Communications (DSRC) Standards in the United States," *Proc. IEEE*, vol. 99, no. 7, July 2011, pp. 1162–82.
[6] J. Deepakumara, H. M. Heys, and R. Venkatesan, "Performance Comparison of Message Authentication Code (MAC) Algorithms for Internet Protocol Security (IPSEC)," *Proc. Newfoundland Elec. and Comp. Eng. Conf.*, Nov. 2003.
[7] E. Barker and A. Roginsky, "Transitions: Recommendation for Transitioning the Use of Cryptographic Algorithms and Key Lengths," NIST Special Publication, 800-131A, Jan. 2011.
[8] European Network and Information Security Agency (ENISA)," The Algorithms, Key Size and Parameters Report," Nov. 2014.
[9] NSA, "Fact Sheet Suite B Cryptography," Sept. 2014.
[10] H. Alanazi *et al.*, "New Comparative Study between DES, 3DES and AES Within Nine Factors," *J. Computing*, vol. 2, no. 3, Mar. 2010, pp. 152–57.
[11] D. S. A. Elminaam, H. M. Abdual-Kader, and M. M. Hadhoud, "Evaluating The Performance of Symmetric Encryption Algorithms," *Int'l. J. Network Security*, vol. 10, no. 3, May. 2010, pp. 216–22.
[12] M. Hartong, R. Goel, and D. Wijesekera, "Key Management Requirements for PTC Operations," *IEEE Vehic. Tech. Mag.*, vol. 2, no. 2, June 2007, pp. 4–11.
[13] N. Lyamin *et al.*, "Real-Time Detection of Denial-of-Service Attacks in IEEE 802.11p Vehicular Networks," *IEEE Commun. Letters*, vol. 18, no. 1, Jan. 2014, pp. 110–13.
[14] Z. Peng, G. Li, H. Wang and Q. Wu, "A Wireless Transmission Mechanism of the Wireless CBTC System and Performance Analysis," *IEEE Int'l. Conf. Audio, Language and Image Processing*, July 2008, pp. 443–48.
[15] M. Heddebaut *et al.*, "Towards a Resilient Railway Communication Network against Electromagnetic Attacks," TRA-Transport Research Arena, Apr. 2014.

Biographies

Igor Lopez (igor.lopez@ehu.eus) has been a Ph.D candidate at the University of the Basque Country since 2013. He received his M.S. degree in information, systems and technology from the University of Paris-Sud and his Eng. degree from the University of the Basque Country in 2012. His current research focuses on railway communication systems, 4G mobile networks, communication reliability mechanisms, and cyber security for industrial networks.

Marina Aguado (marina.aguado@ehu.eus), who has a Ph.D. in telecommunications engineering, currently works as an associate professor at the University of the Basque Country and as a senior researcher in the research group I2T. Her expertise is focused on train control systems, more specifically on European Railway Train Control Systems (ERTMS) and communication technologies for transport systems. She has eight years of experience in the railway transportation industry.

Ultra-Wide Bandwidth Systems for the Surveillance of Railway Crossing Areas

Marco Govoni, Francesco Guidi, Enrico M. Vitucci, Vittorio Degli Esposti, Giovanni Tartarini, and Davide Dardari

ABSTRACT

Level crossings are critical elements of railway networks where a large number of accidents take place every year. With the recent enforcement of new and higher safety standards for railway transportation systems, dedicated and reliable technologies for level crossing surveillance must be introduced in order to comply with the safety requirements. In this survey the worldwide problem of level crossing surveillance is addressed, with particular attention to the recent European safety regulations. In this context, the capability of detecting, localizing, and discriminating the vehicle/obstacle that might be entrapped in a level crossing area is considered of paramount importance to save lives, and at the same time avoid costly false alarms. In this article the main solutions available today are illustrated and their pros and cons discussed. In particular, the recent ultra-wide bandwidth technology, combined with proper signal processing and backhauling over the already deployed optical fiber backbone, is shown to represent a promising solution for safety improvement in level crossings.

INTRODUCTION

Level crossings (LCs), where almost 50 percent of all train accident events caused by third parties take place, are very difficult for the railway sector to control. In the U.S. there are approximately 270 deaths each year at public and private grade crossings, and nearly every 180 minutes someone is hit by a train [1]. The Federal Railroad Administration (FRA), through the efforts of its Highway-Rail Crossing and Trespasser Prevention Division, is committed to reduce that number. Federal funding to install automatic warning devices and other improvements for public highway-rail crossings is managed by the Federal Highway Administration and commonly referred to as the Section 130 program.

In 2010 the European Railway Agency (ERA) disclosed the European benchmark in LC safety reporting 619 significant LC accidents resulting in 359 fatalities and 327 serious injuries. In the EU, LC accidents represented 27 percent of all significant railway accidents and 28 percent of all fatalities on railways, suicides excluded [2]. There are currently approximately 1.2 million LCs in the EU, and on average there are five LCs per 10 line-km. Half of them are active LCs with some sort of user-side warning, while the reminders are passive LCs typically equipped only with the St. Andrew's cross traffic sign. A similar active/passive percentage (43%/57%) also applies to the reported 250/523 highway-rail grade crossings in the United States [1]. LCs with automatic user-side warning (typically flashing lights and sound) are the most common type of active crossings (38 percent) in Europe, closely followed by LCs with automatic user-side protection and warning (barriers with lights) (34 percent).

The economic impact of fatalities and serious injuries in LC accidents was estimated to be 350 million Euros in 2010 [2]. The 2004/49 CE directive is oriented to promote the development and improvement of safety on the EU Community's railways by harmonizing the regulatory structure in the member states. The concepts of *common safety targets* and *common safety methods* have been introduced here to ensure that a high level of safety is maintained and possibly enhanced, and one of the most critical points is the protection of *LCs*, which is defined in the common safety indicator (CSI) of this directive. CSIs are based on common definitions and calculation methods. The data set is structured following significant accidents, deaths and serious injuries, economic impact of accidents, technical aspects (level crossings by type and automatic train protection systems), and management of safety [3]. In order to maximize LC safety levels while preserving a reliable and fast network, companies are then required to develop technical solutions complying with the EU safety requirements.

In this context, the availability of large backhaul communication networks based on fiber-optic technology, which have been deployed in several countries along railway lines in recent years, paves the way for a new generation of LC surveillance systems, such as those relying on ultrawide-band (UWB) signals.

In this article we illustrate the application of UWB technology to LC safety and, successively, we propose a solution exploiting the combina-

Marco Govoni is with Rete Ferroviaria Italiana (RFI) and theUniversity of Bologna.

Francesco Guidi, Enrico M. Vitucci, Vittorio Degli Esposti, Giovanni Tartarini, and Davide Dardari are with the University of Bologna.

 0163-6804/15/$25.00

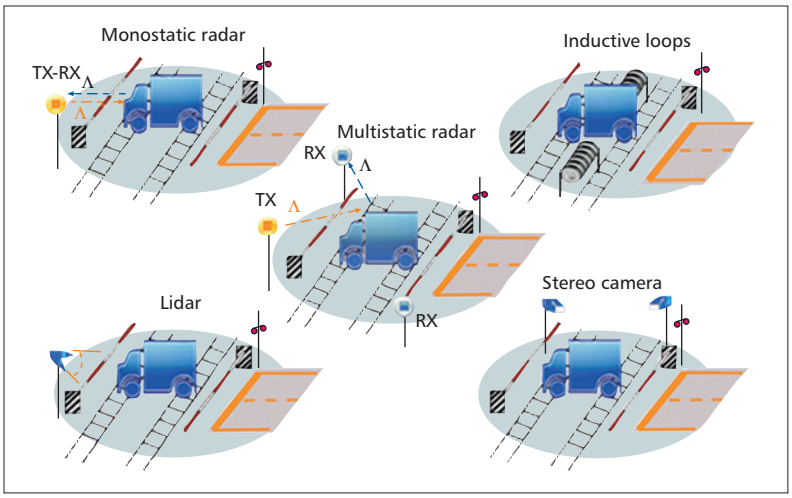


Figure 1. Different technologies available to detect the presence of an entrapped object inside the level crossing area.

tion of UWB and fiber-optic links to centralize all signal processing tasks to a remote central unit.

The article is organized as follows. The most recent technologies for LC surveillance are surveyed by illustrating their main characteristics and limitations with reference to the current safety requirements. Our proposed solution is then presented, where a case study is also reported to show the performance achievable using such a system. Finally, potential future developments are described and conclusions are drawn.

Railway Crossing Safety Systems

Safety Requirements

Safety requirements and regulations are specific to each country, and in some cases to single rail infrastructure operators. Therefore, in the following we use the European scenario as a reference.

Passive safety requirements for railway vehicles are defined in the Commission Decision UE 291/2011 (par. 4.2.2.5), and refer to all subsystems, comprising surveillance system for LC areas, that can operate independently from the railway infrastructure. This definition fits the LC scenario where, despite the automatic user-side warning as well as barrier closing, a road vehicle could be entrapped inside a LC, causing an extremely dangerous situation that could lead to a collision with the oncoming train. Two reference collision cases are classified for LCs: the impact of the train with a *large obstacle* or a *small obstacle*. The entrapped object of the first case is described in EN15227/2008 (Table 2, Sec. 5) as a heavy truck or a tank. In the second case the EU decision does not provide specific information about the smallest size of the obstacle. To fix a requirement, it is reasonable to consider the minimum size of a vehicle that must be detected inside the LC and generate an alarm. This can be approximated with a parallelepiped volume placed on the ground with dimension equal to $2 \times 1.1 \times 1.3\text{m}^3$. This dimension is slightly smaller than that of the smallest minicar available on the market. Under a conservative setting, the critical dimension for the performance assessment of different surveillance systems can be chosen equal to one cubic meter. Therefore, one of the key performance parameter is the capability of the system to discriminate the volume of the obstacle (when present), as only obstacles larger than one cubic meter must generate an alarm with consequent stop of the train.

According to current EU regulations, LC surveillance systems must also satisfy functional requirements in terms of robustness to weather conditions, cost, and ease of installation on existing infrastructures, making their design challenging. In particular, the *tolerable hazard rate* is defined as a target measure of both systematic and unpredictable failure integrity. For instance, LC surveillance systems must guarantee a *false alarm rate* (i.e. the detection of obstacles even if they are not present in the area) less than 1.9×10^{-4}, which is equivalent to one false alarm per year with traffic of 20 trains per day. On the other hand, the *misdetection rate* (i.e. the missed detection of an obstacle when it is effectively present in the area) must be below 10^{-8} [3]. In the following we describe current possible solutions that aim to preserve safety in LC areas. Specifically, only systems that can be integrated with the railway infrastructure without human interaction will be considered here. As a consequence, other solutions such as those that rely on car speed reduction through bumpers, on traffic signal improvement, or on vision-based surveillance, are out of the scope of this work.

Surveillance Systems State of the Art

In recent years several systems have been proposed for LC surveillance, each supported by a different technology:

- Microwave/millimeter-wave radar.
- Inductive loops detector.
- Laser imaging detection and ranging (LIDAR).
- Stereo camera detection.

The radar concept was born with microwave technology with the main intent to detect the presence of an intruder inside a monitored area. A key radar indicator is the radar cross section (RCS), which represents the projected area of a metal sphere that would scatter the same power in the same direction as the target does. For most radar systems, their ability to determine the dimensions of an object is based on an RCS estimate by analyzing the reflected signal (backscatter). Several operating frequency bands as well as radar architecture configurations have been exploited. For example, in [4] two UWB mono-static radars cover a half-portion of the monitored area, respectively (Fig. 1a). Each UWB radar detects obstacles eventually present in its covered area portion by analyzing the backscattered signal to obtain only a rough approximation of the obstacle's size. Unfortunately, the coverage area separation among sensors determines a low localization resolution and makes the system performance particularly sensitive to single sensor outage. In [5] the multiple-input multiple-output (MIMO) antenna array concept is developed for a frequency modulated continuous wave (FMCW) radar operating at

Technology	System architecture	Dimension estimation	Localization capability	Heavy rain	Dense fog	Cost range
Mono-static radar	Two sensors [4]	RCS estimation	None	Reliable	Reliable	Low
Ka-band radar	Independent multiple sensors [6]	RCS estimation	None	Reliable	Reliable	Medium
V-band radar	Antenna array [7]	RCS estimation	Approximated 2D localization	Reliable	Reliable	Medium
MIMO radar	Antenna array [5]	RCS estimation	Approximated 2D localization	Reliable	Reliable	Medium
FOS radar	Multiple sensors [13]	Good resolution	Good 3D localization	Reliable	Reliable	Medium
Iductive loops	Multiple buried turns [8]	Low resolution	None	Reliable	Reliable	Low
LIDAR	Single head [9]	High resolution	High 3D localization	Blind spot	Unreliable, image degradation	High
Stereo camera	Single head [10, 11]	High resolution	High 3D localization	Unreliable, image degradation	Unreliable, image degradation	High

Table 1. Comparison between the characteristics of different monitoring technologies.

25 GHz. Even though determining the dimensions of the obstacle still relies on RCS, the MIMO configuration allows for a 2D angular resolution in azimuth.

With the liberalization and the possibility to use the Ka-band and higher bands (e.g. V-band), different systems exploit these new frequencies that are characterized by a low level of interference. In particular, the authors in [6] adopt FMCW at 36.5 GHz using up to nine sensors, whereas the authors in [7] investigate a spread spectrum radar at 60 GHz using the correlation of pseudo-noise codes to detect the obstacle in distance and azimuth, respectively.

A common limitation of all these systems is that RCS is not a reliable indicator of an object's dimensions as it strongly depends on the object's reflection characteristics and shape, which are not known a priori. Moreover, since electromagnetic wave propagation cannot be confined in a specific area, false alarms caused by objects located outside the LC area are possible. This is actually the main limitation of radio-based systems that can be mitigated only if high-accuracy localization capabilities are implemented, as detailed later.

The solution based on inductive loops [8] was among the first to be proposed. The loops are excited with signals whose frequencies range from 10 KHz to 50 KHz, and when a vehicle stops on or passes over, their inductance is decreased. Depending on the resonance frequency generated by the wire loop, it is possible to identify specific metal portions of the vehicle. This system is very simple but extremely inaccurate in estimating an obstacle's dimensions. The massive presence of metal in the railway causes problems in threshold setting and, additionally, wire loops are subjected to traffic stresses and temperature effects.

Another solution is represented by LIDAR technology, which exploits ultraviolet, visible, or near infrared light to illuminate a target with a laser. Object detection and 3D image reconstruction are based on the time-of-flight (TOF) of the electromagnetic wave. In [9] environment scanning is performed through a single-head 3D laser range finder which is tilted to create a 3D image of the scene, as shown in Fig. 1c. As for the stereo camera solutions, several works have been published extracting 3D information from digital images by comparing the same scene taken from two advantageous locations.

A common problem of image detection methods is static background estimation, which causes the necessity to detect and track incoming, staying, or outgoing objects in the LC area, as investigated in [10], where 3D localization is performed by hierarchical belief propagation algorithms. Differently, the authors in [11] assume that the displacement of the image contents between two nearby instants (frames) is small and approximately constant within a neighborhood of the point under consideration. Thus the optical flow equation can be assumed to hold for all pixels within a window centered at that point.

The main characteristics of current technologies are summarized in Table 1, where it can be noted that present microwave and millimeter-wave solutions do not provide any or only rough information about the obstacle's volume and position. On the other side, the solutions based on image detection, while exhibiting very high resolution in obstacle shape detection, might suffer from image degradation when working in non-optimal weather conditions. In Table I the different solutions are also compared in terms of cost of the technological apparatus. We must point out that other related costs, such as site-specific costs deriving from the installation of

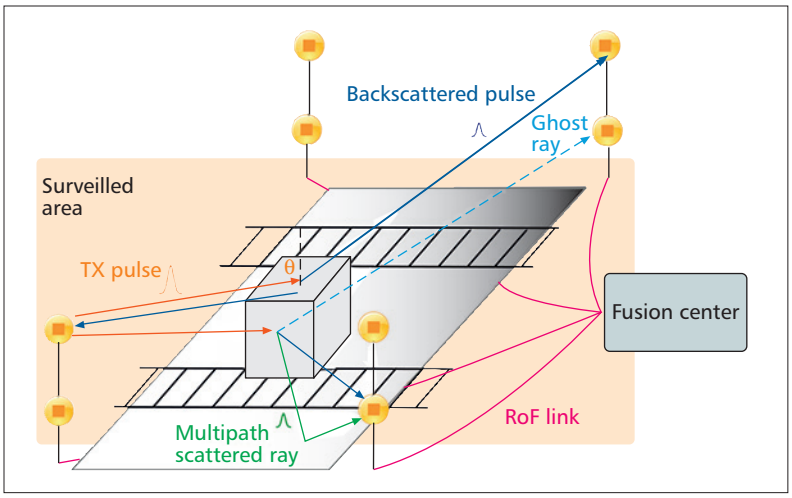


Figure 2. Architecture of the level crossing surveillance system with the FOS UWB radar.

the supporting infrastructure, might have a determinant impact and can be difficult to generalize.

Another important parameter is maintainability, which falls into the so called management of RAMS (reliability, availability, maintainability, and safety). One of the figures of merit defined in the EN-50126 European Standard for Railway Applications, also adopted by other international projects such as the California High Speed Train (CHSTP), is the mean down time (MDT), which is the average time when a system is not operational. All the considered systems exhibit a degree of quantifiable self-imposed down time for periodical calibration. However, the time related to repair, corrective, and preventive maintainance, and logistic or administrative delays, depends again on the individual instance considered, and therefore a general comparison in terms of MDT has not been included in Table 1.

UWB Radar/Imaging Systems

In this section we discuss how it is possible to design a LC surveillance system to enhance LC safety by exploiting UWB radar technology with a partial multi-static architecture.

UWB Radar

A promising wireless technique to detect, localize, and estimate the dimensions of obstacles is the UWB technology characterized, in its impulse radio implementation, by the transmission of sub-nanosecond duration pulses. The employment of UWB signals enables the resolution of multipath and extraordinary localization precision based on TOF estimation of the signal. The UWB radar is therefore considered to be an interesting option for surveillance tasks in terms of spatial resolution. UWB radar architectures can be *multi-static*, where one or more transmitters send the interrogation signal and more receivers, located in different positions, process the signals backscattered by the environment, as shown in Fig. 1b. When transmitters and receivers coincide and the receiver manages only its own backscattered signals, we obtain a *mono-static radar* architecture as illustrated in Fig. 1a. Both solutions are well investigated but their direct application to the problem of LC surveillance could be problematic. Indeed, regarding multi-static radars, the potentially large size of obstacles prevents the adoption of classical solutions that are implicitly based on unrealistic assumptions such as isotropic scattering, moving objects, and punctual obstacle size [12]. On the other hand, mono-static UWB imaging systems (UWB scanners) provide high-accuracy obstacle imaging but require a large amount of antennas mounted on a mechanical arm that circumnavigates the obstacle contour, which is obviously not feasible in the application under consideration. For these reasons, in the following we discuss a possible solution recently proposed in [13] to counteract the above mentioned issues.

The Partial Multi-Static UWB FOS

As previously discussed, classical UWB multi-static radar schemes work under quite unrealistic assumptions and might fail when applied to railway crossing areas. In fact, the finite size and the anisotropic scattering of the obstacle might prevent some nodes (e.g. those located in the opposite direction) from receiving the backscattered signal. Therefore, the absence of expected components (e.g. obstruction, high angle of incidence of the wave on the object) and the presence of unexpected components in the received signal (e.g. multipath) might generate unsolvable ambiguities when detecting and estimating the right position of the object, thus seriously compromising image formation, as will be shown.

To overcome such limitations, we describe a recently proposed solution, namely a fixed object scanner (FOS), capable of detecting, localizing, and estimating the obstacle's volume, even in static conditions. It makes use of a fixed set of UWB nodes to obtain the information about the volume of the obstacle and discriminate between large or small obstacles. Specifically, the surveillance system is composed of a set of transmitter (TX) and receiver (RX) nodes, located at different heights at the vertices of the monitored area, as shown in Fig. 2. The sounding of the environment via UWB interrogation signals and subsequent analysis of backscattered signals is split in different phases to which only a subset of nodes participate, leading to a partial multi-static radar configuration. In particular, the FOS algorithm performs five phases, four for the lateral sides and one for the top of the area. During each phase, only the TX-RX pairs located in the considered side are activated, thus miming the presence of several mono-static imaging scanners with fixed nodes. In this way, the resulting partial multi-static radar operates most likely in conditions where the incident angle of the electromagnetic wave impinging the obstacle is $< 90°$, with a consequent significant mitigation of the aforementioned ambiguities during the imaging process in Fig. 2.

With the purpose of facilitating the 3D imaging algorithm, the monitored area is subdivided into small 3D cubic pixels. The 3D imaging process of the obstacle can be summarized in the following steps: *clutter removal*, *pixel detection*, *imaging*, and *volume estimation*.

Clutter Removal: An important issue when detecting the presence of steady obstacles is the static environment response (static clutter) caused, for example, by the rail and poles. This component is removed by using an empty-room approach in which the reference signals, recorded in the absence of obstacles, are subtracted from the actual received signals. Note that when an obstacle is present, part of the static clutter could be hidden, leading to imperfect clutter suppression. To counteract this ghost effect, only the signal components corresponding to positive variations in the received energy are taken into account during the clutter removal process.

All measurements are successively collected by a fusion node responsible for making an overall decision on the event. To reduce the number of LCs to be monitored by a given fusion center, an interesting opportunity is to connect the sensor nodes and the fusion center through fiber-optic links, as will be discussed in the next section.

Pixel Detection and Imaging: Obstacle detection and image formation consist of checking whether the generic pixel is a candidate for containing part of the obstacle (if present). This is accomplished by performing, during each phase and for each pixel, a specific binary detection test where the corresponding likelihood ratio is compared with a threshold as described in [13]. This procedure is repeated for each pixel and phase. In the end, all binary test outputs are combined to form the 3D image. In particular, the presence of part of an obstacle in a 3D pixel is detected if at least one pixel is above the threshold during the scanning phases.

Volume Computation: The result of the 3D image formation is used as input for volume computation to estimate the size of the obstacle, if present, and generate an alarm to stop the train if the estimate value is greater then 1 cubic meter. One possible approach is to compute the average parallelepiped volume starting from the moment of the pixels along each dimension with respect to the barycentre as proposed in [13].

Note that the FOS algorithm can be considered a hybrid approach combining the UWB multi-static radar and the mono-static imaging scanner configurations. As a consequence, it allows for gaining some of the advantages of both configurations and mitigating their drawbacks. Indeed, it overcomes the limitations of optical based systems, [9, 10] and at the same time offers good obstacle detection and localization performance inside the LC as will be shown in a later section.

Remote Processing using the Optical Fiber Infrastructure

The implementation in loco of all signal processing tasks involving UWB signals could be costly and imply difficult maintenance procedures. Therefore, remote operations would be preferable. On the other hand, the transmission of raw UWB measurements coming from sensor nodes is extremely demanding in terms of the required bandwidth. Fortunately, most countries in Europe have deployed dedicated fiber-optic communication infrastructures that could be exploited to move the signal processing tasks from LCs to a central unit. For instance, the Italian railway operator Rete Ferroviaria Italiana (RFI) has developed an optical network that covers more than 10,000 km. The cable deployment started in the 1980s to develop a high capacity network and fulfill the growing technology demand on the railway process. Fifty percent of fiber links are available for future applications. To eliminate the need of extremely high-speed analog-to-digital converters at sensor nodes, the UWB radio-over-fiber (RoF) approach is of particular interest. This technique is indeed widely utilized in many applications for antenna remotization, since it allows the transparent transmission of the received signal to the central unit [14]. Each antenna is then equipped with a RoF link, which in turn is formed by a RoF transmitter, based on a distributed feed back (DFB) laser, a strand of G-652-compliant optical fiber, and a RoF receiver based on a PIN-photodiode followed by an RF amplifier. The modulation bandwidths of today's DFB lasers extend easily to some GHz, and this allows their direct modulation, avoiding the use of costly external electro-optical modulators. The wavelength of operation for the DFB has been chosen as $\lambda = 1310$ nm, so that operating in the second optical window, where the chromatic dispersion is very low, the distortion effects due to laser frequency chirp become negligible. The same effects hold for laser non-linearities, which in the case of DFB are more limited with respect to other kinds of lasers (e.g. Fabry Perot or vertical cavity), and also considering the low power levels of the input UWB signals, can be neglected as well. Consequently, the main detrimental effect in RoF transmission results from the increase in the noise figure of the system caused by the RoF link. For short link lengths the increase is mainly caused by the relative intensity noise of the DFB and by the shot noise of the PIN, while for longer link lengths the thermal noise of the receiver RF amplifier becomes dominant.

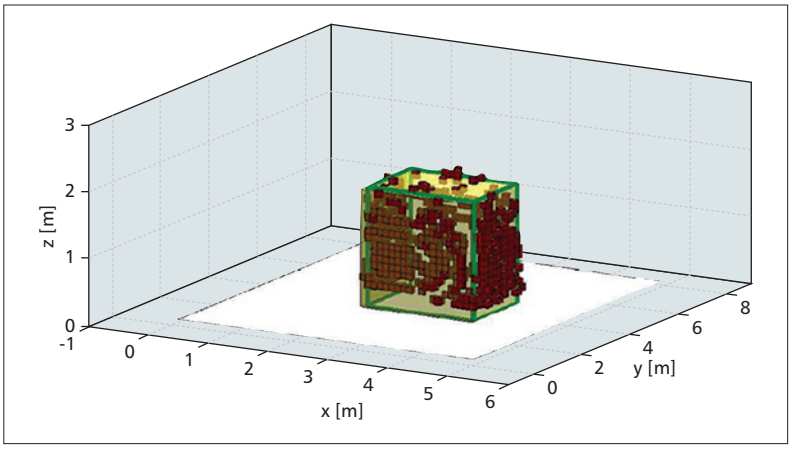


Figure 3. 3D image of a metal box of 5.72 m^3 in the middle of the surveillance area using the FOS approach.

Case Study

Now we present a case study where a classical UWB multi-static approach and the FOS algorithm are compared to assess their capability in discriminating the volume of the obstacle.

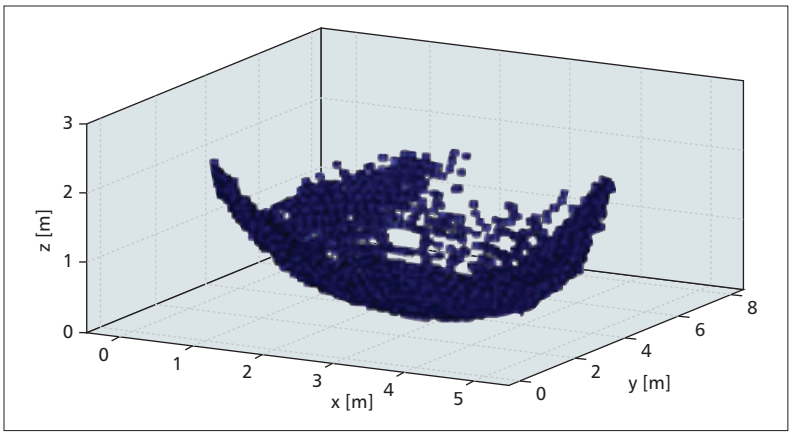


Figure 4. 3D image of a metal box of 5.83 m^3 in the middle of the surveillance area using the classic UWB multi-static radar approach.

The surveillance area is divided into 3D pixels of side $\Delta = 10$cm. The channel transfer function between each TX–RX pair has been simulated with the aid of the 3D ray tracing (RT) software described in [15]. In addition to specular reflection and edge/corner diffraction, modeled through geometrical optics (GO) and uniform theory of diffraction (UTD), the RT tool accounts for the effect of diffuse scattering, modeled through the effective roughness (ER) approach [15]. One of the main parameters of the ER model is the scattering parameter S, where S^2 represents the amount of power that is diffused in all directions at the expense of specular reflection, due to the presence of surface and volume irregularities. The obstacle is modeled as a metal box, whereas ground, barriers, tracks, and antenna poles are modeled as slabs.

The 3D imaging approach previously described has then been applied to the output of RT simulations, for obstacles having volume 5.83, 1, and 0.34 m^3 placed inside the surveillance area.

Figure 3 shows the 3D image output of the FOS algorithm for a metal box of 5.83 m^3, located in the middle of the area. A typical value for outdoor environments of $S = 0.3$ has been used in this case, thus meaning that about 10 percent of the reflected power is diffused in non-specular directions [15]. The actual geometry of the object, as modeled in RT simulations, is also included as a reference in the figure (green line). The yellow parallelepiped juxtaposed can be taken as representative of the actual volume of the obstacle.

For comparison, the same simulation setup has been used to derive the results in Fig. 4 where the classical UWB multi-static radar approach is considered [12]. Even though the presence of the obstacle is detected, a huge number of outlier pixels arise due to ambiguities, thus making impossible a realistic volume computation and/or localization of the obstacle. Compared with Fig. 3, the gain introduced by the FOS algorithm is evident.

Table 2 summarizes the volumes computed for some simulation setup, i.e. different volumes and positions of the metal box, and exploiting optical fiber connection. As can be noticed, the classic approach makes impossible a realistic volume computation and/or localization of the obstacle. Instead, the same volume computation exploiting FOS detection/imaging provides an approximate estimation of the actual volume of the obstacle. Moreover, the stability of the FOS approach is tested for increasing lengths of the optical connection between sensors and the central unit. As can be noted, the noise introduced by the RoF link irremediably corrupts UWB signals after 40 km. We can conclude that up to 40 km optical links are tolerable without creating a significant performance degradation in terms of obstacle imaging.

Conclusions and Future Prospects

Recent safety requirements adopted in Europe as well as in other countries for the surveillance of railway level crossing areas are very demanding, and therefore new and sophisticated technologies for automatic surveillance are necessary. Some existing solutions have been discussed, concluding that a technology change is unavoidable to fully meet the safety requirements. An important step in this direction is represented by the adoption of the UWB technology in conjunction with UWB radio-over-fiber communication exploiting the fiber-optic networks already deployed in most railway networks. In this article we have shown that this technology, if opportunistically paired with the FOS approach, can enhance the capability of the surveillance system to discriminate the dimensions of objects inside the monitored area, thus overcoming the limitations of other approaches. Extensive studies corroborated by experimental campaigns considering a wider set of requirements are necessary to confirm the validity of this and other solutions.

Box volume [m^3]	Position	Classical approach [m^3]	FOS [m^3] for different RoF links				
			0 km	10 km	20 km	30 km	50 km
5.83	Middle	>10	6.49	6.52	6.52	6.53	11.01
1.00	Middle	>10	1.14	1.14	1.14	1.16	3.15
1.00	Corner	≈10	1.23	1.28	1.23	1.12	13.45
0.34	Middle	>1	0.24	0.37	0.37	0.41	1.13

Table 2. Object volume estimation capabilities in different configurations.

Anyway, it could be pretentious to rely completely on a single technology; higher levels of safety could be achieved through a smart integration of different technologies (UWB, video, laser, etc.) and by designing advanced data fusion algorithms and communication schemes.

Acknowledgment

Thiswork is the result of the collaboration between Rete Ferroviaria Italiana and the University of Bologna within the project "Protezione Automatica Integrata dei Passaggi a Livello", and it is supported in part by the European project H2020 XCYCLE.

References

[1] D. Deborja and B. A. Hamilton, "Compilation of State Laws and Regulations Affecting Highway-Rail Grade Crossings," *6th Ed. Federal Railroad Administration (FRA)*, U.S. Department of Transportation, Apr. 2014.

[2] I. Forvente, "Existing International Rules and Regulations on Level Crossing," International Union of Railways (UIC), no. 1, 2012.

[3] European Parliament and Council of the European Union," Directive 2004/49/ EC," Brussels, 2004, pp. 1–24.

[4] S. Lohmeier, R. Rajaraman, and V. Ramasami, "Development of an Ultra-Wideband Radar System for Vehicle Detection at Railway Crossings," *Proc. IEEE Conf. Ultra Wideband Syst. and Technologies*, May 2002, pp. 207–211.

[5] A. Narayanan *et al.*, "Railway Level Crossing Obstruction Detection Using MIMO Radar," *Proc. European Radar Conference (EuRAD)*, Oct 2011, pp. 57–60.

[6] G. Ermak *et al.*, "Autodyne Sensors for Hump Yard and Rail Crossing Applications," *Proc. 13th Int. Radar Symposium (IRS)*, May 2012, pp. 209–12.

[7] M. Watanabe *et al.*, "An Obstacle Sensing Radar System for a Railway Crossing Application: A 60 GHz Millimeter Wave Spread Spectrum Radar," *Proc. IEEE MTT-S Int. Microwave Symp. Digest*, vol. 2, June 2002 vol. 2, pp. 791–94.

[8] R. Campbell, "Vital Inductive Loop Processor Detection," Aug. 2014.

[9] G. Kim *et al.*, "Design of Safety Equipment for Railroad Level Crossings Using Laser Range Finder," *Proc. 9th Int'l. Conf. Fuzzy Systems and Knowledge Discovery (FSKD)*, May 2012, pp. 2909–13.

[10] N. Fakhfakh *et al.*, "Background Subtraction and 3D Localization of Moving and Stationary Obstacles at Level Crossings," *Proc. 2nd Int'l. Conf. Image Process. Theory Tools and Applicat. (IPTA)*, July 2010, pp. 72–78.

[11] Z. Silar and M. Dobrovolny, "The Obstacle Detection on the Railway Crossing Based on Optical Flow and Clustering," *Proc. 36th Int'l. Conf. Telecommun. and Signal Process. (TSP)*, July 2013, pp. 755–59.

[12] M. Chiani *et al.*, "Target Detection Metrics and Tracking for UWB Radar Sensor Networks," *Proc. IEEE Int'l. Conf. Ultra-Wideband (ICUWB)*, Sept. 2009, pp. 469–74.

[13] M. Govoni *et al.*, "UWB Multistatic Radars for Obstacle Detection and Imaging in Railroad Crossing Areas," *Proc. 12th Wksp. Positioning Navigation and Commun. (WPNC)*, Dresden, Germany, 2015.

[14] M. Jazayerifar, B. Cabon, and J. Salehi, "Transmission of Multi-Band OFDM and Impulse Radio Ultra-Wideband Signals over Single Mode fiber," *J. Lightwave Technology*, vol. 26, no. 15, Aug 2008, pp. 2594–603.

[15] V. Degli Esposti *et al.*, "Measurement and Modelling of Scattering from Buildings," *IEEE Trans. Antennas Propag.*, vol. 55, no. 1, Jan 2007, pp. 143–53.

Biographies

Marco Govoni (m.govoni@rfi.it) is a Ph.D. candidate in telecommunications engineering at DEI, University of Bologna (UniBo), and with Direzione Tecnica Standard Tecnologici e Sperimentali, Rete Ferroviaria Italiana (RFI). His research interest include ultrawide-band radar networks, ray tracing, and channel propagation models, in addition to fiber optics. Since 2014 he has been participating in the Italian Project PAI-PL (Protezione Automatica Integrata Passaggi a Livello).

Francesco Guidi (f.guidi@unibo.it) received the Ph.D. degree both from Ecole Polytechnique ParisTech and the University of Bologna in electronics, telecommunications, and information technologies. He is currently a postdoctoral researcher at the University of Bologna. He received the best student paper award at the 2014 IEEE International Conference on Ultra-Wideband. His research interests include RFID technology, joint antenna and channel characterization, and signal processing.

Enrico M. Vitucci (enricomaria.vitucci@unibo.it) is a postdoctoral fellow at the Center for Industrial Research on ICT (CIRI ICT), University of Bologna. His research interests are in mobile radio propagation, ray tracing models, and MIMO channel modelling. He participated in the European Cooperation Actions COST 2100 and COST IC1004, in the European Networks of Excellence NEWCOM and NEWCOM++, and in the EU Integrated Project FP7-ICT-ALPHA. He has authored or co-authored more than 40 papers in international journals and conferences.

Vittorio Degli Esposti (v.degliesposti@unibo.it) is an associate professor in the Department of Electrical Engineering (DEI), University of Bologna. He is the author or co-author of more than 100 peer-reviewed technical papers in the fields of applied electromagnetics, radio propagation, and wireless systems. He was appointed Vice-Chair of EuCAP2010 and EuCAP 2011. He is an elected member of the Radio Propagation Working Group of the European Association on Antennas and Propagation (EuRAAP).

Giovanni Tartarini (giovanni.tartarini@unibo.it) is currently an associate professor of electromagnetic fields at DEI, University of Bologna (UniBo). His research interests include the area of microwave photonics. From 2008 to 2011 he was the UniBo responsible of the EU project Architectures for Flexible Photonic Home and Access Networks. Since 2013 he has been the responsible of a collaboration with the Italian Institute of Astrophysics for the development of the Radio over Fiber Receiver within the International Radio-Astronomy Project Square Kilometer Array.

Davide Dardari (davide.dardari@unibo.it) is an associated professor at the University of Bologna. His interests are in ultra-wide bandwidth systems, localization techniques, distributed signal processing, and wireless sensor networks. He is past Chair for the Radio Communications Committee of the IEEE Communication Society. He has served as guest editor for several journals, and was an editor for *IEEE Transactions on Wireless Communications* from 2006 to 2012.

Recent safety requirements adopted in Europe as well as in other countries for the surveillance of railway level crossing areas are very demanding, and therefore new and sophisticated technologies for automatic surveillance are necessary. Some existing solutions have been discussed concluding that a technology change is unavoidable to fully meet the safety requirements.

SOCIAL NETWORKS MEET NEXT GENERATION MOBILE MULTIMEDIA INTERNET





Seshadri Mohan *Nitin Agarwal* *Ashutosh Dutta* *Sudhir Dixit* *Ramjee Prasad*

With ever growing popularity and widespread adoption of mobile social applications, the traffic handled by mobile networks and the Internet has grown significantly. While researchers have been making advances in the study of social networks and independently in the area of next generation wireless networks, very little attention has been given to the interplay between the two, and their impact on each other and society. The challenge of the interplay between social networks and mobile networks is compounded by the fact that advances in smart handheld devices and those in wireless technologies have paved the way for increasing bandwidth catering to very high data rates. Sophisticated social applications such as Second Life, and those involving 3D and real-time data can take advantage of such advances. It is entirely likely that such advances in turn could lead to novel social applications not yet thought of. For example, new social applications could emerge in the area of social health or social games with new forms of massively multi-player, multimedia, 3D, and role playing games.

Next generation wireless (5G and beyond) will emerge that will likely place more emphasis on device intelligence, for example, by exploiting cognitive radio, context, and device-to-device communications. It is also likely to exploit software defined networking, cloud, software-defined radio access networks, and massive MIMO with vastly increased bandwidth that is orders of magnitude more than is available at present, and offer mobile user equipment the ability to establish ad hoc and peer-to-peer communications. These new capabilities will in turn serve as a catalyst to usher in more sophisticated social applications. Applications already exist that can create ad hoc mobile social networks. For example, FireChat can rapidly create an ad hoc mobile network over Bluetooth and WiFi among a large number of users, and their widespread usage globally could facilitate the creation of global ad hoc networks.

Much research has been carried out in the field of social networks involving modeling the growth of social networks and mining of massive information collected through social networks. The results provide insights into human behavior in their interactions with social networks and the formation of new research areas. The emergence of research in the area of big data and its applicability to social networking can hardly be overemphasized. Together, the evolution and deployment of cloud RANs, and cloud-based and software-defined networks could lead to new forms of big data, for example, involving user locations, usage patterns, user mobility, and other user-specific behavior. Also, the ability on the part of next generation networks to facilitate peer-to-peer and ad hoc networking paves the way for new forms of interactions with mobile social networks and applications. Hence, important research questions arise while attempting to understand the interplay between mobile social networks and the next generation mobile multimedia Internet.

This Feature Topic (FT), which serves as the sequel to the June 2012 *IEEE Communications Magazine* FT, Social Networks Meet Mobile Networks,[1] assembles six interesting articles that address several important issues within the highly complex interdisciplinary field addressing the theme of the FT. In these articles readers can find answers to a set of key questions that are bound to stimulate further interest and research.

In the first article, "Socially Enabled Wireless Networks: Resource Allocation via Bipartite Graph Matching" by Wang *et al.*, the authors answer the question: how do social interactions between users and users' devices influence resource allocation, including radio spectrum, and impact the design of next generation wireless networks? The authors approach the problem first by defining *social graphs* and *interest graphs*, where a social graph represents the social tie between users through either real life relationship such as family members, classmates, and colleagues, or social media such as Facebook and LinkedIn, and an interest graph represents the users' interests such as hobbies, watching similar categories of movies, or other similar interests. Subsequently, through a hierarchical bipartite graph partitioning approach, they define an upper layer bipartite graph to pair users with similar social ties and interests, and argue that such an approach would reduce the cost of content sharing. In a similar manner, a lower layer bipartite graph is defined in which social ties and

[1] S. Mohan, N. Agarwal, and A. Dutta, Guest Editors, "Social Networks Meet Mobile Networks," *IEEE Communications Magazine*, June 2012.

interests are resolved into radio resource requirements and facilitate the sharing of radio resources.

The second article, "Location-Based Social Video Sharing over Next Generation cellular Networks" by Roy *et al.*, addresses the question: would sharing of location information improve the usage of social applications in a community and improve the quality of experience (QoE) of users? They argue that due to the increasing need to share video content in the social context among users, there is an urgent need to devise novel solutions to minimize CAPEX, increase efficiency, and improve user QoE. They propose that by sharing the location information of the users between 4G LTE mobile network operators, not only would the routing of live video streaming be done more efficiently, but users' overall QoE would improve significantly by reducing jitter, supporting higher bit rate streams, and reducing latency in video playback start time.

In a manner similar to sharing location information as proposed by the previous article, by exploiting the knowledge of mobility patterns of users in a community, can efficient routing protocols be designed for sharing and distributing content? The third article, "NCCU Trace: Social-Network-Aware Mobility Trace" by Tsai *et al.*, addresses the challenge with an approach to collect mobility traces and model them. The article studies the impact of users' mobility on the routing protocols for delay-tolerant networks, as this would be largely determined by social networking behavior. Using a location- and behavior-aware Android application, the authors have collected the mobility traces of college students in a campus environment and designed a mobility model to capture such movement. The mobility traces were then imported into a simulator to verify the performance of the routing protocols and evaluate the performance of social-based routing methods. Simulation results show that this trace-based mobility model is much closer to the real movement situation and can evaluate the performance of social-based routing method.

Extending the previous questions further, how could a network become socially aware with knowledge of communities to facilitate efficient sharing of content such as video? The fourth article, "Social-Aware Mobile Peer-to-Peer Communications for Community Multimedia Streaming Services," by Xu *et al.*, addresses this question. The article discusses the challenges in defining and implementing a virtual community, that is, a socially aware mobile multimedia community (SMMC), for sharing and delivering multimedia content, and presents an overview of the work on the estimation methods for similarity in terms of demand, socialization, and mobility. The authors describe an approach for community construction and suggest an integrated network context-aware concurrent multi-path transmission solution for content delivery. They also compare performance between the proposed scheme and an alternative solution, the ant-inspired mini-community-based solution for video-on-demand services (AMCV).

With the ever increasing popularity of crowdsourcing as a powerful paradigm in harnessing the knowledge and work of a large number of people to accomplish a given task (e.g., the compilation of Wikipedia), how does one evaluate the reliability and trustworthiness of information or data generated, and how do crowdsourcing and mobile sensing impact each other? Participatory computing, especially crowdsourcing, has been tremendously influenced by smart mobile and handheld devices. Chen *et al.*, the authors of the fifth article, "When Crowdsourcing Meets Mobile Sensing: A Social Network Perspective," have leveraged crowdsourcing to enhance mobile sensing, thereby demonstrating a synergistic advancement of both the disciplines (i.e., mobile networks and human computation). Mobile sensing has emerged from several decades of research on wireless sensor networks and as a result of recent technological advances that have transformed multiple mobile devices into one multipurpose sensing device. Mobile sensing utilizes agent-participatory data to improve decision making, but such data could be extremely noisy. Consequently, identifying trustworthy data and reliable agents becomes an essential task in mobile sensing. The article leverages the concepts of wisdom of crowds from the social networks discipline to address this challenge. Authors have conducted evaluations of their model on real-world data and report promising findings. The authors' work has far-reaching implications not only in the mobile networks domain but also in human computation, citizen science research, crowdsourcing tasks, and many other participatory computing paradigms.

Since data sharing assumes utmost importance in crowdsourcing and mobile sensing applications, can systems be designed that allow sharing of the collected data independent of the communications infrastructure, its capabilities and location, and applications that require the data? In the last article of this FT, "Pervasive Data Sharing for Support of Mobile Citizen Sensing Applications," Moreira and Mendes provide insight into the effect of pervasiveness of large-scale sensing systems, and how this may lead to improvement of social and personal welfare by exploiting novel mobile citizen sensing applications. They also underscore the need for these devices to be able to share sensing data independent of the available communication infrastructure, their location, and applications that make use of this data. The authors have taken a pragmatic and constructive engineering approach to define a process for the design of a pervasive data sharing application. They give an introduction to the networking requirements of mobile citizen sensing and propose four design paradigms that represent basic building blocks for the solutions.

We strongly believe that the set of articles in this FT will stimulate further research and create a venue to bring together researchers and practitioners from different disciplines, especially computer and information sciences and next generation mobile/wireless multimedia Internet/networks as well as other related disciplines to share, exchange, learn, and develop preliminary results, new concepts, ideas, principles, and methodologies, aiming to advance mobile networks in the information and communication technologies involved in next generation mobile multimedia Internet. We hope the readers enjoy these articles as much as we did!

BIOGRAPHIES

SESHADRI MOHAN (sxmohan@ualr.edu) is currently a professor in the Systems Engineering Department at the University of Arkansas at Little Rock (UALR), where, from August 2004 to June 2013, he served as chair of the Department of Systems Engineering. Prior to his current position, he served as the chief technology officer with Telsima, Santa Clara, California; chief technology officer with Comverse, Wakefield, Massachusetts; a senior research scientist with Telcordia, Morristown, New Jersey; and a member of technical staff with Bell Laboratories, Holmdel, New Jersey. Besides his industry positions, he also held faculty positions at Clarkson and Wayne State Universities. He has authored/coauthored over 100 publications in the form of books, patents, and papers in refereed journals and conference proceedings. He co-authored the textbook *Source and Channel Coding: An Algorithmic Approach*. He has contributed to several books, including *Mobile Communications Handbook* and *The Communications Handbook*. He holds 14 patents in the area of wireless location management and authentication strategies as well as in the area of enhanced services for wireless. He is the recipient of the SAIC Publication Prize for Information and Communications Technology. He has served or is serving as a Technical Editor of *IEEE Personal Communications* (now *IEEE Wireless Communications*), *IEEE Communications Surveys and Tutorials*, and *IEEE Communications Magazine*. He has also served as a Guest Editor for several Special Issues and Feature Topics in *IEEE Network*, *IEEE Communications Magazine*, and *ACM MONET*. He

served as a Guest Editor of the March 2012 *IEEE Communications Magazine* Feature Topic "Convergence of Applications Services in Next Generation Networks" as well as the June 2012 Feature Topic "Social Networks Meet Wireless Networks." He is the recipient of the 2010 IEEE Region 5 Outstanding Engineering Educator Award. He coauthored the paper "A Multi-Path Routing Scheme for GMPLS-Controlled WDM Networks" presented at the 4th IEEE Advanced Networks and Telecommunications Systems, which received the Best Paper Award. He holds a Ph.D. degree in electrical and computer engineering from McMaster University, Canada, a Master's degree in electrical engineering from Indian Institute of Technology, Kanpur, India, and a Bachelor's degree in electronics and telecommunications from the University of Madras, India.

NITIN AGARWAL is the Jerry L. Maulden-Entergy Chair Professor of Information Science at UALR. His research interests lie in the areas of social computing (e.g., knowledge discovery in social media, modeling social dynamics including collective action, influence, trust, community evolution, and collective intelligence), data mining and machine learning (especially big "social" data analytics), and privacy. In this direction, he has developed an interdisciplinary research program on social computing at UALR with foundational as well as applicational contributions. Foundational contributions are made to computational social network analysis; social science theories such as collective action, collective behavior, and homophily; data mining; privacy; and virtual organizations. The applicational contributions include, but not limited to, event analysis, monitoring cyberthreats through social media, smart health and wellbeing, social media in learning environments, network and communication, and socially aware mobile networks. Aside from information science, the research program brings together researchers from various disciplines such as social science, economics, political science, communication and organization science, and computer and mobile networks and practitioners including defense analysts from NATO, U.S. Naval Research Lab, Dillards, Acxiom, @WalmartLabs, and other organizations. The research has resulted in publications in various prestigious forms, including 5 books, 19 journal articles, 13 book chapters/encyclopedia entries, and over 50 conference proceeding papers. The research studies have received the Best Information System Publication of 2012 Award recognized by the AIS Senior Scholar Consortium, a Best Paper Award, and several best paper nominations. He has guest edited several Special Issues for *Elsevier Journal of Systems and Software, Oxford's The Computer Journal,Springer's Lecture Notes in Social Networks, IEEE Communications Magazine, Elsevier Journal of Computational Science, and IEEE Internet Computing*. His research has been supported by grants from the U.S. National Science Foundation (NSF), U.S. Office of Naval Research (ONR), U.S. Air Force Research Lab (AFRL), and U.S. Army Research Office (ARO). He obtained his Ph.D. in computer science from Arizona State University with outstanding dissertation recognition in 2009. He has a Bachelor's of Technology in information technology from the Indian Institute of Information Technology, India. For more details see http://ualr.edu/nxagarwal/.

ASHUTOSH DUTTA [SM] is a Senior Member of the ACM. He obtained his B.S. in electrical engineering from NIT Rourkela, India, his M.S. in computer science from New Jersey Institute of Technology, and his M. Phil. and Ph.D. in electrical engineering from Columbia University, New York. He is currently lead member of technical staff at AT&T's Security and Mobility Organization within the Chief Security Office, where he leads the design and architecture of security for next generation mobility networks. His 25-year career includes CTO of Wireless at a cybersecurity company, NIKSUN, senior scientist at Telcordia Applied Research, director of the Central Research Facility at Columbia University, and computer engineer at TATA Motors. He has more than 80 conference and journal publications, three book chapters, and 28 issued patents; he has given tutorials in mobility management at various conferences. His research interests include wireless Internet, multimedia signaling, mobility management, 4G networks, IMS, VoIP, and session control protocols. He is co-author of the book *Mobility Protocols and Handover Optimization: Design, Evaluation and Application* (Wiley). He serves as Editor-in-Chief for the *Journal of Cybersecurity and Mobility* (River Publishers). He is a Senior Member of ACM, and has served as Chair of the IEEE Princeton/Central Jersey Section, Industry Relation Chair for Region 1 and MGA, Pre-University Coordinator for IEEE MGA, and Chair of the Ad Hoc Committee for Public Visibility of IEEE ComSoc. As Vice Chair of the Education Society Chapter of PCJS, he co-founded the IEEE STEM Conference in 2011 and helped to implement Engineering Projects in Community Service in the high schools within PCJS. He currently serves as Director of Marketing and Industry Relations for IEEE ComSoc. He was the recipient of the prestigious 2009 IEEE MGA Leadership award and the 2010 IEEE-USA Professional Leadership Award. He currently serves as Vice Chair of the Global ICT Standardization Forum for India's Service Oriented Networking WG.

SUDHIR DIXIT [F] has been a Distinguished Chief Technologist and CTO at Communications and Media Services (Americas), HP Enterprise Services since December 2013, and is based in Palo Alto, Californa. Prior to this he was director of HP Labs India since September 2009. From June to August 2009, he was a director at HP Labs. Prior to joining HP Labs, he held a joint appointment as CTO at the Centre for Internet Excellence and a research manager at the Centre for Wireless Communications, Oulu, Finland. From 1996 to 2008, he held various positions with leading companies, such as BlackBerry as a senior director, Nokia and Nokia Networks in the United States as senior research manager, Nokia Research Fellow, head of Nokia Research Center (Boston, Massachusetts), and head of Network Technology (USA). From 1987 to 1996, he was at NYNEX Science and Technology and GTE Laboratories (both now Verizon Communications). He has 21 U.S. patents granted, has published over 200 papers, and edited, co-edited, or authored six books: *Wi-Fi, WiMAX and LTE Multi-hop Mesh Networks* (Wiley, 2013), *Globalization of Mobile and Wireless Communications* (Springer, 2011), *Technologies for Home Networking* (Wiley, 2008), *Content Networking in the Mobile Internet* (Wiley, 2004), *IP over WDM* (Wiley, 2003), and *Wireless IP and Building the Mobile Internet* (Artech House, 2002). He is presently on the Editorial Boards of *IEEE Spectrum*, Cambridge University Press Wireless Series, and Springer's *Wireless Personal Communications Journal* and the *Central European Journal of Computer Science*. He is Chairman of the Vision Committee and Vice Chair for the Americas at the Wireless World Research Forum. From 2010 to 2012, he was an adjunct professor of computer science at the University of California, Davis, and has been a docent (adjunct professor) of broadband mobile communications for emerging economies at the University of Oulu, Finland. A Fellow of IET and IETE, he received a Ph.D. degree in electronic science and telecommunications from the University of Strathclyde, Glasgow, United Kingdom, and an M.B.A. from the Florida Institute of Technology, Mel, Florida. He received his M.E. (electronics) degree from Birla Institute of Technology and Science, Pilani, India, and his B.E. (electrical engineering) from Maulana Azad National Institute of Technology, Bhopal, India.

RAMJEE PRASAD [F] is currently director of the Center for TeleInFrastruktur (CTIF) at Aalborg University, Denmark, and Wireless Information Multimedia Communication Chair Professor. He is the Founding Chair of the Global ICT Standardisation Forum for India (www.gisfi.org) established in 2009. GISFI has the purpose of increasing the collaboration between European, Indian, Japanese, North American, and other worldwide standardization activities in ICT and related application areas. He was Founding Chair of the HERMES Partnership, a network of leading independent European research centers established in 1997, of which he is now Honorary Chair. He is a Fellow of IETE, India, IET, United Kingdom, WirelessWorld Research Forum, and a member of the Netherlands Electronics and Radio Society and the Danish Engineering Society. He is also a Knight ("Ridder") of the Order of Dannebrog (2010), a distinguished award by the Queen of Denmark. He has received several international awards, the latest being the 2014 IEEE AESS Outstanding Organizational Leadership Award for "organizational leadership in developing and globalizing the Center for TeleInFrastruktur Research Network." He is the founding Editor-in-Chief of the *Springer International Journal on Wireless Personal Communications*. He is a member of the Editorial Board of other renowned international journals including those of River Publishers. He is a member of the Steering Committees of many renowned annual international conferences, such as Wireless Personal Multimedia Communications Symposium, Wireless VITAE, and Global Wireless Summit. He has published more than 30 books, over 900 journal and conference publications, and more than 15 patents, and has mentored more than 90 Ph.D. students and over 200 Master's students. Several of his students are now telecommunication leaders worldwide.

CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE
WIRELESS COMMUNICATIONS, NETWORKING, AND POSITIONING WITH UNMANNED AERIAL VEHICLES

BACKGROUND

Enabled by the advances in computing, communication, and sensing as well as the miniaturization of devices, unmanned aerial vehicles (UAVs) such as balloons, quadcopters, and gliders, have been receiving significant attention in the research community. Indeed, UAVs have become an integral component in several critical applications such as border surveillance, disaster monitoring, traffic monitoring, remote sensing, and the transportation of goods, medicine, and first-aid. More recently, new possibilities for commercial applications and public service for UAVs have begun to emerge, with the potential to dramatically change the way in which we lead our daily lives. For instance, in 2013, Amazon announced a research and development initiative focused on its next-generation Prime Air delivery service. The goal of this service is to deliver packages into customers' hands in 30 minutes or less using small UAVs, each with a payload of several pounds. 2014 has been a pivotal year that has witnessed an unprecedented proliferation of personal drones, such as the Phantom and Inspire from DJI, AR Drone and Bebop Drone from Parrot, and IRIS Drone from 3D Robotics.

Among the many technical challenges accompanying the aforementioned applications, leveraging the use of UAVs for delivering broadband connectivity plays a central role in next generation communication systems. Facebook and Google announced in 2014 that they will use a network of drones which circle in the stratosphere over specific population centers to deliver broadband connectivity. Such solar-powered drones are capable of flying several years without refueling. UAVs have also been proposed as an effective solution for delivering broadband data rates in emergency situations through low-altitude platforms. For example, the ABSOLUTE, ANCHORS, and AVIGLE projects in Europe have been investigating the use of aerial base stations to establish opportunistic links and ad-hoc radio coverage during unexpected and temporary events. They can serve as a temporary, dynamic, and agile infrastructure for enabling broadband communications, and quickly localizing victims in case of disaster scenarios.

This proposed Feature Topic (FT) issue will gather articles from a wide range of perspectives in different industrial and research communities. The primary FT goals are to advance the understanding of the challenges faced in UAV communications, networking, and positioning over the next decade, and provide further awareness in the communications and networking communities on these challenges, thus fostering future research. Original research papers are to be solicited in topics including, but not limited to, the following themes on communications, networking, and positioning with UAVs.

- Existing and future communication architectures and technologies for small UAVs
- Delay-tolerant networking for cooperative UAV operations
- Design and evaluation of wireless UAV test beds, prototypes, and platforms
- Multi-hop and device-to-device communications with UAVs
- Interfaces and cross-platform communication for UAVs
- QoS mechanisms and performance evaluation for UAV networks
- Game-theoretic and control-theoretic mechanisms for UAV communications
- Use of civilian networks for small UAV communications
- Integrating 4G and 5G wireless technologies into UAV communications, such as millimeter wave communications, beamforming, moving networks, and machine type communications
- Use of UAVs for public safety and emergency communications, networking, and positioning
- Integration of software defined radio and cognitive radio techniques with UAVs
- Channel propagation measurements and modeling for UAV communication channels

SUBMISSIONS

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words (from introduction through conclusions). Figures and tables should be limited to a combined total of six. The number of references is recommended not to exceed 15. In some rare cases, more mathematical equations, figures, and tables may be allowed if well-justified. In general, however, mathematics should be avoided; instead, references to papers containing the relevant mathematics should be provided. Complete guidelines for preparation of the manuscripts are posted athttp://www.comsoc.org/commag/paper-submission-guidelines. Please send a pdf (preferred) or MSWORD formatted paper via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to Author Center. Follow the instructions there. Select "May 2016 / Wireless Communications, Networking and Positioning with UAVs" as the Feature Topic category for your submission.

SCHEDULE FOR SUBMISSIONS

- Submission Deadline: November 1, 2015
- Notification Due Date: January 15, 2016
- Final Version Due Date: March 1, 2016
- Feature Topic Publication Date: May 2016

GUEST EDITORS

Ismail Guvenc
Florida International Univ., USA
iguvenc@fiu.edu

Walid Saad
Virgina Tech, USA
walids@vt.edu

Mehdi Bennis
Univ. of Oulu, Finland
bennis@ee.oulu.fi

Christian Wietfeld
TU Dortmund Univ., Germany
christian.wietfeld@tu-dortmund.de

Ming Ding
NICTA, Australia
ming.ding@nicta.com.au

Lee Pike
Galois, Inc., USA
leepike@galois.com

Socially Enabled Wireless Networks: Resource Allocation via Bipartite Graph Matching

Li Wang, Huaqing Wu, Wei Wang, and Kwang-Cheng Chen

ABSTRACT

The influence of social interactions among mobile devices and network components in wireless networks has attracted substantial attention due to its potential impact on resource allocation of spectrum and power in particular. We present an organized social graphical view on resource allocation and then extend to multi-objective resource allocation of wireless networks. We subsequently consider taking advantage of multi-dimensional resources, including radio resource, user behavior, and content characteristics, such that we can successfully integrate caching capability, interest similarity, and content popularity and distribution into wireless network design. As an illustration, device-to-device communications is utilized to form pairs and clusters of mobile devices regarding optimal resource matching via a bipartite graph. This socially enabled methodology highlights new potential to design wireless networks and 5G mobile communications.

INTRODUCTION

Modern design methodology of wireless networks proceeds on the ground of the Open Systems Interconnection (OSI) layer structure. A typical approach starts from a physical layer transmission mechanism, multi-access on a medium, and then networking algorithms and protocols. Such an approach is very reasonable, particularly when the traffic is connection-based such as voice calls, and can be extended to variable rate video and best effort data.

As the Internet and cloud computing have created tremendous services and applications, social media and the Internet of Things (IoT) supply major traffic for mobile wireless networks, which actually suggests a new paradigm in the design of wireless networks. As indicated in [1], the interplay of technological networks and social networks has not been fully exploited. It is also noted that any entities of probabilistic or dynamic relationship can be generalized into a social network. When social media and human-centric applications drive the progress of networking technology, the methodology to leverage social network analysis to design wireless networks is of particular interest but remains open [2]. In this article, we orient graph matching in social network analysis to systematically enhance the design of wireless networks, with focus on radio resource allocation.

Applying graph theory to resource allocation in a network is not a strange idea. However, this approach is scattered in the literature and mostly done in a rather ad hoc manner. In this article, we present a systematic approach of generalizing social networks and also apply graph theory for appropriate resource allocation in wireless networks, particularly by leveraging bipartite graphs. Our overall system scenario is depicted in Fig. 1, not targeting radio resource allocation directly according to only the air interface, but taking social interactions, caching capability, and interest similarity for users into account simultaneously. Later in this article, we elaborate more on content sharing and resource sharing in socially enabled wireless networks.

RESOURCE ALLOCATION VIA BIPARTITE GRAPH

Wireless physical radio resource can be considered jointly in the frequency and time domains. The traditional physical layer transmission, or *single-user PHY*, allows only one transmission at a certain time over a certain frequency band. Even in orthogonal frequency-division multiplexing (OFDM), adopted by IEEE 802.11a/g, all 48 data subcarriers are dedicated in each transmission. With the introduction of orthogonal frequency-division multiple access (OFDMA) in fourth generation Long Term Evolution (4G-LTE), also known as multi-user OFDM, we have entered the era of employing *multi-user PHY*; thus, radio resource is utilized by radio blocks in terms of frequency-time resource elements. Here, we introduce a systematic and generally

Li Wang and Huaqing Wu are with Beijing Key Laboratory of Work Safety Intelligent Monitoring, School of Electronic Engineering, Beijing University of Posts and Telecommunications.

Wei Wang is with Zhejiang University.

Kwang-Cheng Chen is with National Taiwan University.

0163-6804/15/$25.00 © 2015 IEEE

applicable way to abstract radio resource optimization and thus allocation in modern complex wireless networks.

As shown in Fig. 2, suppose there are M users who buffer or cache popular contents for content sharing with partners via wireless communications. Each user can select its partner based on whether the chosen partner has the desired (target) content it needs, whether the chosen partner can be trusted or tightly connected, and whether the allocated spectrum resource and channel state for the corresponding links are stable enough to support the whole data transmission, as shown in Fig. 1. All these demands can be satisfied by selecting appropriate partners, which is exactly a bipartite graph problem. The partner pairing problem is illustrated in Fig. 2a. More generally, the users we discuss could also be the base stations (BSs) or the operators. In these cases, the cell selection or operator selection can be addressed by bipartite graph matching as well.

Then suppose there are N partitioned units of radio resource that can be PHY time-frequency resource blocks, labeled $R_1, \ldots, R_N$, to support users 1, 2, 3, … , M. The communication or networking bandwidth demands are therefore represented by $a_1, a_2, \ldots, a_M$ (or $b_1, b_2, \ldots, b_M$). For user satisfaction, user i's bandwidth demand a_i can be assigned to one of the available radio resource units, say R_1, R_2, or R_3, as illustrated in Figs. 2b and 2c, which is exactly another bipartite graph. The resource allocation is therefore a bipartite graph matching that is well studied in graph theory, and [4] is one early example of applying such a concept. Similarly, the users here could be BSs or operators.

Let us use a new set of notations to study a bipartite graph representing a system. Suppose $\{S_1, S_1, \ldots, S_n\}$ is a collection of $n \geq 2$ finite nonempty sets. This collection is said to have a system of distinct representatives if there are n distinct elements $x_1, x_2, \ldots, x_n$ such that $x_i \in S_i$, $\forall i$. In other words, x_i is the representative of the set S_i. A condition that collections of sets contain a system of distinct representatives is known as Hall's theorem in graph theory.

Hall's Theorem

A collection $\{S_1, S_2, \ldots, S_n\}$ of n nonempty finite sets has a system of distinct representatives if and only if for each integer k with $1 \leq k \leq n$, the union of any k of these sets contains at least k elements.

Hall's theorem suggests that the identification of cluster leaders in a bipartite graph can be equivalent to resource allocation based on each entity's requirements. Computationally efficient algorithms have been widely investigated in the mathematical literature, including weighted bipartite graphs [3]. This remarkable theorem provides a sufficient condition for the existence of a perfect matching. The violation of its condition (e.g., two pairs after a selection algorithm may aim to use the same resource, resulting in violating the achievement of perfect matching and subsequent performance degradation) leads to performance degradation. Fortunately, Hall's theorem holds for most cases in wireless networks, and is successfully applied in many networking algorithms in engineering literature by satisfying the statistical performance constraints.

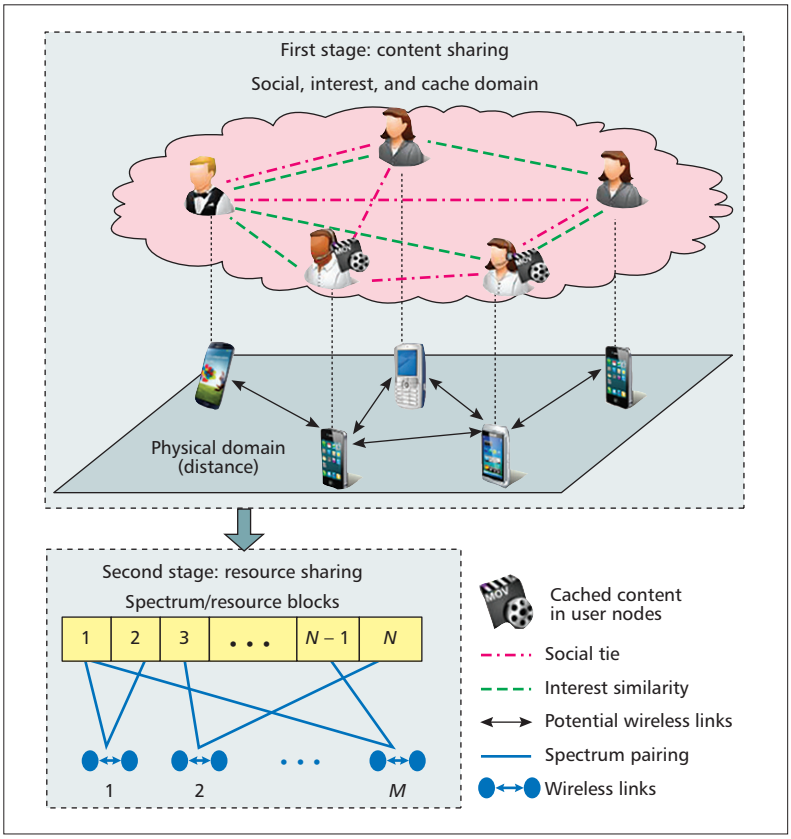


Figure 1. A hierarchical and multi-domain oriented graph matching framework. Both contents and radio resources could be considered as resources in general to design wireless networks.

Literature Survey

One of very early connections between bipartite graph and resource control in networks was reported in [4], which links the network control algorithms and the edge coloring algorithms for bipartite graphs. As a matter of fact, bipartite graph matching has been deeply studied and widely applied to discrete resource allocation in social economics for a long time [2, 5], while other typical applications include house assignment, hospital bed matching, and college admission/selection, among others.

In recent years, the bipartite graph has been utilized in several cases of radio resource allocation in wireless networks. To simultaneously consider both the bandwidth utilization and starvation problems, a bipartite graphical method was introduced to dynamic spectrum allocation in wireless mesh networks [6]. A remarkable idea, to construct a conflict graph to avoid interference, can be established during the process. Algorithms were developed in bipartite graph matching based on maximum cardinality or maximum weight [7]. For spectrum efficiency, maximum weighted graphical modeling is usually useful. Along the same thinking, optimal channel and relay assignment in a multi-pairing OFDM relay network can be translated into a maximum weighted bipartite matching problem [8]. It also

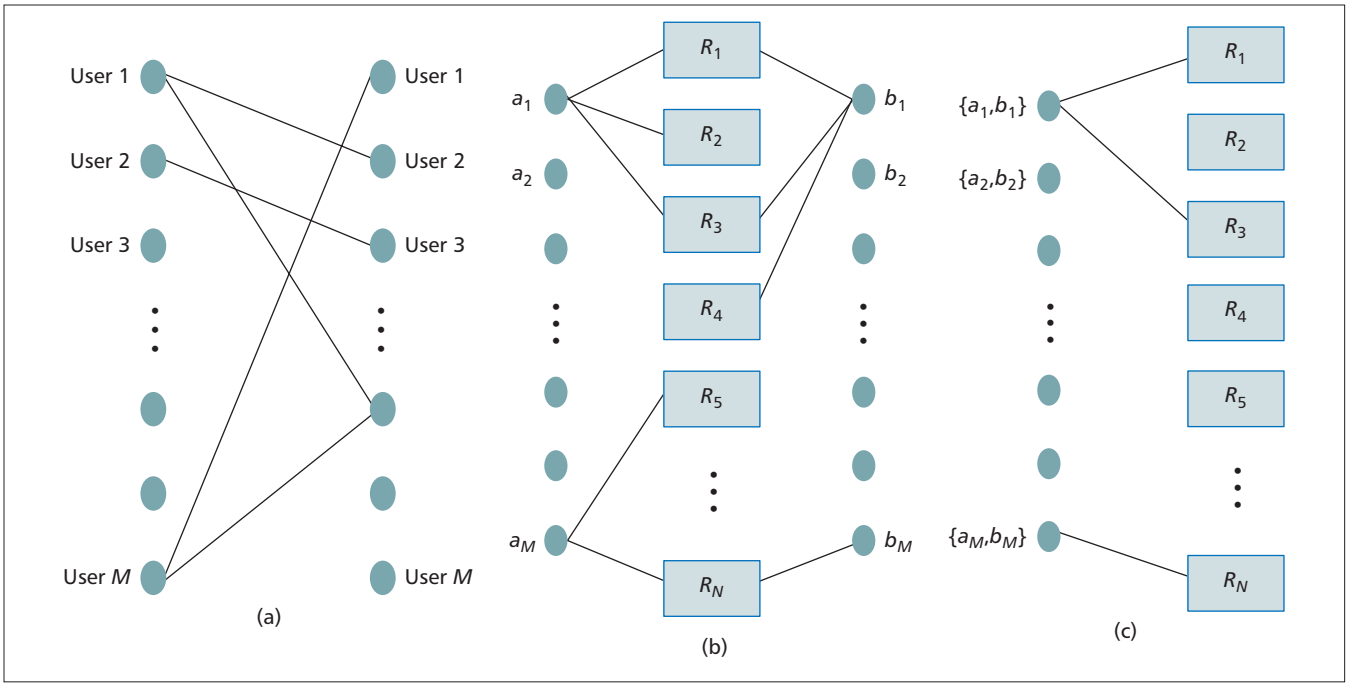


Figure 2. Multi-objective oriented bipartite graphs and evolving of concepts to facilitate: a) multi-dimensional partner selection; b) resource allocation via graph; c) regenerative bipartite graph.

implicitly suggests that multi-dimensional factors can be considered in the optimization of resource allocation.

Multi-Objective and Hierarchical Bipartite Graph Matching

Although techniques have been scattered in literature and summarized above, an effective methodology is crucial to achieve appropriate resource allocation in state-of-the-art complex wireless networks supporting diverse application scenarios, especially for cases taking social interaction into account. Starting from a straightforward extension using bipartite graph matching, we discuss multi-objective generalization of graph matching to fully utilize social networking. Furthermore, we propose a novel concept of hierarchical bipartite (HBP) to elaborate practical realization.

Multi-Objective Bipartite Graph Matching

An immediate extension of multi-objective bipartite graph matching is bandwidth aggregation, while modern mobile communications networks have to serve traffic ranging from social video and high-definition movies to IoT traffic of small packets that might require low-latency transportation. User traffic might require huge bandwidth in terms of multiple radio resource units, to satisfy either quality of service (QoS) or quality of experience (QoE). Since multi-objective resource allocation can be modeled as a new regenerative bipartite graph matching, as illustrated in Fig. 2, in which multiple repetitions are needed for ensuring allocation success.

According to graph theory, Hall's theorem guarantees the conditions of successful allocation. However, recent advances in networking suggest further dimensions of consideration in resource allocation, particularly in wireless networks. In particular, social characteristics for mobile users emerge to play a key role in wireless network design [9, 10]. Social-interaction-based resource allocation optimization becomes multi-objective, that is, maximizing the sum system capacity with different QoS constraints or minimizing the outage probability, which inspires generalizing the social graph theoretical resource allocation into multi-objective purposes as well. Furthermore, we can also target the secrecy rate of mobile users in the presence of eavesdroppers.

Luckily, it is rather straightforward to generalize bipartite graph matching to multi-dimensional matching. Suppose a user (or traffic) needs radio resource allocation by satisfying different design criteria, without loss of generality, a_1 and b_1. To achieve a_1, we create a bipartite graph, and similarly to achieve b_1, illustrated by Fig. 2. Different from earlier bipartite graph matching problems, a user or traffic can have multi-objective resource allocation, where the objectives may come from different criteria associated with each stream of traffic from a social networking perspective. We may create a multi-dimensional bipartite graph (Fig. 2b). If we collect all objectives together to form vectors of request, $\{a_m, b_m\}$ for user/traffic m,$1 \le m \le M$, the candidate resource blocks must satisfy both criteria. For example, R_1, R_2, and R_3 satisfy a_1, and R_1, R_3, and R_4 satisfy b_1. The request vector $\{a_1, b_1\}$ for user/traffic 1 only connects to R_1 and R_3 (i.e., the intersection of sets formed by resource blocks matching both criteria, Fig. 2c). In this way, we can obtain a regenerative bipartite graph for multi-objective resource allocation; then existing allocation algorithms based on Hall's theorem can be extended in a straightforward way.

Hierarchical Bipartite Graph Matching

In practice, various objectives determine different kinds of bipartite graph we should adopt. An unweighted bipartite graph can be adopted to achieve maximum cardinality, that is, maximizing the number of wireless links assigned spectrum resource for transmission. For weighted bipartite graphs, the weight of the edges can be defined differently for various objectives.

Homophily is widely known in social networks: users with similar interests or tight relationships are likely to have similar behavior, such as downloading the same popular contents. Users can select an appropriate partner that has the target content by considering multi-dimensional factors, such as the physical distance and social trust between different user nodes, and the cache capacity of user nodes. Then a user can obtain its need from its partner(s), which can significantly improve efficiency and offload from BSs. Afterward, wireless resource allocation can be carried out to assign proper resource blocks to maximize system performance.

Therefore, we propose the concept of HBP, which consists of upper layer bipartite for communication partner selection and lower layer bipartite for resource allocation. When it comes to the upper layer case, the benefit of caching for neighboring partners cannot be ignored. From the perspective of the upper layer, hit probability and access delay represent the weight when content sharing is taken into account. On the other hand, when it comes to the lower layer case, we can simply consider physical wireless resource allocation in wireless networks; thus, the weight can be the achievable data rate, secrecy rate, and outage probability of wireless links. Furthermore, the system performance can be measured by the weight of one particular edge or the sum weight by utilizing the maximum weighted bipartite matching.

In addition to the hierarchical bipartite graph described above, when users communicate with BSs directly rather than forming communication partners, the resource allocation can be accomplished by using one single bipartite considering multi-dimensional factors referring to Figs. 2b and 2c. In this case, we do not have to discuss the formation of a user partner, targeting the access time, hit probability, data rate, and so on, which represents a bipartite graph corresponding to Fig. 2a. However, we note that partner formation can be much more important and effective in some random cases or events, and can usually be ignored among tightly connected communities or societies.

HBP Framework for Socially Enabled Wireless Networks

Recall that the social tie (relationship), interest similarities, caching capabilities, and physical distance among users can be represented by graphs, which are composed of nodes and edges among them. To reflect the efficient, accurate, and practical graph regarding multi-domain factors while avoiding unnecessary recomputations and energy consumption as much as possible, we conduct the methodology mentioned above in the physical, social, interest, and cache domains as shown in Fig. 3, in terms of the partner formation (e.g., pairing or clustering) and physical resource allocation for communication links, which are upper layer bipartite and lower layer bipartite, respectively.

According to Fig. 2c, with slight modification, we can obtain the ranking list of candidates for the partner finding case relating to Fig. 2a, and physical resources to different users or multiple preferred factors of specific users corresponding to Fig. 2b. To avoid re-computing for cases with different objectives, we can use the overlapped ranking list between them as our candidates, as shown in Fig. 2c. It is also suitable for cases when we consider time-variant conditions to reduce complexity and avoid redundancy in computing.

Content-Sharing-Oriented Upper Layer Bipartite

Two devices within the maximum physical distance of each other can transmit data and form a communication pair (links or partners), which eventually establish a connected graph. Toward content sharing, in addition to the *physical distance*, it is worth considering the social and cache considerations jointly in the user pairing process. *Social ties* and *interest similarities* (common interests) ensure that the users who already have the desired contents would like to share their data with higher security. Note that social ties can be evaluated by social trust, as shown in Fig. 3, since trust and privacy are going to be more and more important for human society. *Cache capability* makes it possible for a user to cache more than the required data and improve the efficiency of content sharing. Considering these multiple effective factors that affect user pairing performance, we can target the objective of user pairing as data rate, cache hit probability, access delay, and many others.

Wireless-Resource-Sharing-Oriented Lower Layer Bipartite

After the user pairing considering upper layer factors, we consider the wireless resource allocation for these constructed user pairs according to the lower layer factors, including the spectrum properties and social relationship. Different spectrum bands accommodate various transmission opportunities, propagation properties, and so on, which supply different communication bandwidth for the data transmission of user pairs. With limited wireless resources, spectrum efficiency can be improved by spatial reuse, where multiple user pairs share their resource. For efficient resource sharing, the mutual social relationship and interactions between user pairs that use the same wireless resource should also be considered. With tight social ties (high social trust) among user pairs using the same resource, it is possible to achieve more efficient resource sharing, such as efficient interference coordination among these user pairs, by leveraging their transmit power. For wireless resource sharing, in addition to the achievable data rate and system capacity, security is also of great importance. Hence, we can also consider security objectives

An unweighted bipartite graph can be adopted to achieve maximum cardinality, that is, maximizing the number of wireless links assigned spectrum resource for transmission. For weighted bipartite graphs, the weight of the edges can be defined differently for various objectives.

(i.e., secrecy capacity or secrecy outage probability) for all the matched transmitter-receiver pairs.

Clustering Extension for Hierarchical Bipartite

In addition to pairing users, it is also necessary to divide the users into groups (clusters) regarding to social tie, interest similarity (common interest), caching ability, and computing capability as well, for sharing the same resources later. However, for the upper layer case (i.e., the 1st stage), the clustering case of taking multiple factors into consideration might be quite challenging since the utility of a user being added into or removed out of a cluster depends on all the other users' influence and willingness to share within the same cluster, and thus the edge weights of bipartite graph are generally not constants. On the other hand, for the lower layer case (i.e., the second stage), we not only need to pair the given spectrum resource with one single user but a number of clustered users instead. Different system reward metrics are considered (e.g., achievable network capacity, outage probability, energy efficiency, and secrecy-oriented metrics as well). That means, we also need take the co-channel interference among cluster users into account simultaneously, as shown in Fig 4. The clustering in Fig. 4 degenerates to a pairing problem when the number of members in cluster M drops to 1 (i.e., $K_m = 1$).

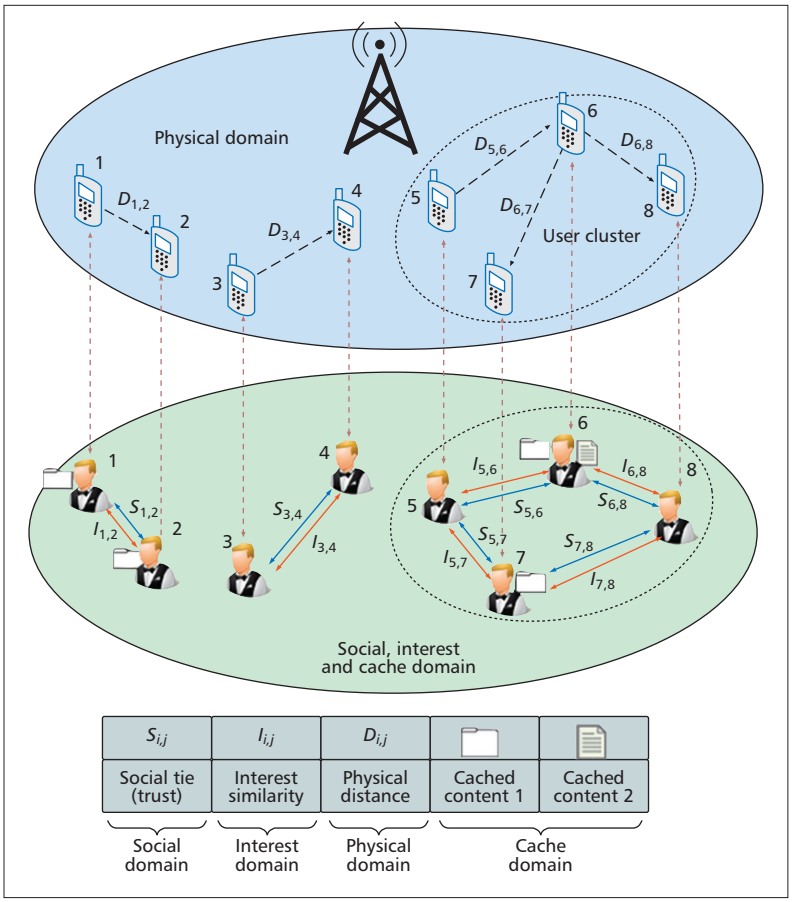


Figure 3. Multiple factors oriented users' pairing and clustering for the content sharing case: upper layer bipartite.

Weights for Resource Allocation via HBP

To establish the hierarchical bipartite of resource allocation, it is necessary to identify effective factors for content sharing and wireless resource sharing. With users' social interactions and effective physical communication, we would particularly like to include social trust, interest similarity, and cache capability in the HBP framework in terms of social ties and content caching.

Social-Tie-Driven Scenario

We start by defining graphs with different purposes to comprehend the relationship and consequent framework. A social graph is a connected graph that indicates the social ties among users who know each other in real life (e.g., family members, classmates, colleagues, and club members) or in social networking society (e.g., Facebook and LinkedIn). An interest graph indicates members' hobbies, interests, or individual needs. It can be generated by the new feeds that users follow, products that they purchase (e.g., on Amazon), movies or media that they prefer (e.g., ratings on YouTube), or survey information that they supply.

Social tie and interest similarity are simultaneously involved in the weights of resource allocation bipartite graphs, and will be of vital significance to better exploit contemporary social-network-based wireless communication.

In the upper layer bipartite of HBP, social ties bring trust in the real world into the weights for content sharing. It is probably more decent for privacy to share resources between users who are familiar with each other, which leads to a large weight in the bipartite graph. Users in the same society usually have similar interests, such as downloading the same popular dancing video for the members of one dance club. Taking social and interest graphs into account will potentially improve the security and success ratio and reduce the system cost for content sharing.

In the lower layer bipartite of HBP, social ties facilitate the cooperation of wireless resource sharing among user pairs that use the same resource, and thus improve system performance. User pairs with tight social ties have large weights for sharing the same resource.

Content-Caching-Driven Scenario

Traffic localization by distributed caching provides an efficient way to reduce the cost of content sharing, which is the core idea of content-centric networks (CCNs) [11]. As the storage sizes of wireless communication devices grow, the cache contents of a user includes not only his/her own pertinent data but also specifically required popular data cached for other users [12]. The benefits induced by caching the former data mainly depend on the similarity of individual interests or preference among devices for users, and that for the latter is affected by the cache capability of user devices.

Recall that in the interest graph concept defined above, users in the same society usually

have similar interests. For example, the members of one dance club would like to download the same popular dancing video. For better content sharing, which users are selected to download this video from the server and cache it locally depends on the users' storage capability. Some user devices have stronger cache capability, so these corresponding users can cache more popular contents for sharing with other users.

Different from the social-tie-driven case, the cache contents affect the upper layer bipartite only. Cache content related to users' interests, content multiformity, and replica locations over the network topology should be taken into account in the weights of the upper layer bipartite graph to achieve maximum content hit probability and minimum content access delay [13].

HBP Graph Matching Implementation in Wireless Networks

According to the concept of HBP, the design and implementation of wireless networks can be composed of two steps (stages), which utilize the information from user partners and allocate wireless/radio resources, respectively.

In the first stage, regarding the exploitation of multiple domains, we establish user partners considering not only the physical distance but also the social interaction and interest similarity of mobile users in HBP, as shown in Fig. 3. Specifically, $s(m, n) \in [0, 1]$ and $c(m, n) \in [0, 1]$ are the social trust index and interest similarity between mobile users m and n, respectively. Furthermore, $r(m, n) \in [0, 1]$ can also represent caching capability for user m, the ratio of the cached content to the whole demand of its partner n.

After pairing the user partners, in the second stage, radio resource blocks should be assigned to those communication links to accomplish the data transmission and service requirements. Thus, the achievable data rate of communication link j between mobile users m and n using resource block i (spectrum) can be expressed as

$$R(m,n) = \sum_j q_{i,j} s_1(m,n) \cdot \log_2(1+\xi_j), \quad (1)$$

where $q_{i,j}$ is the matching indicator variable between resource block i and communication link j. $q_{i,j} = 1$ indicates that communication link j uses the resource block i; otherwise, $q_{i,j} = 0$. $s_1(m, n)$ denotes the union of social-oriented multi-factors between two mobile users who can potentially form a communication link, $s_1(m, n) = a \cdot s(m, n) + b \cdot c(m, n) + c \cdot r(\mathrm{m}, n)$, where a, b, c are constant weight parameters according to different factors, and a Œ $[0, 1]$, $b \in [0, 1]$, $c \in [0, 1]$, $a + b + c = 1$. ξ_j is the signal-to-interference-plus noise ratio (SINR) of communication link j.

To bring the critical security concern into this scenario, we can replace the data rate in Eq. 1 in favor of the secrecy rate for joint optimization of secrecy- and efficiency-oriented resource allocation. However, we omit the similar optimization equation because of length limitation.

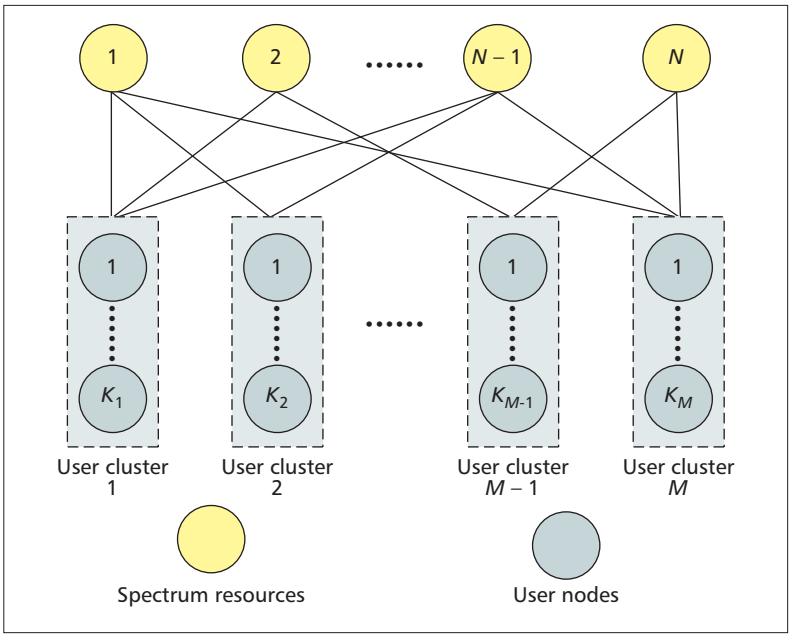


Figure 4. Resource sharing for user pairing and clustering: lower layer bipartite.

D2D Communications: An Illustration

In mobile social networks, device-to-device (D2D) communications have proven efficient for mobile services with high rate demands, such as video sharing and dissemination, since it can potentially establish direct links between proximity users without going through the BS.

We take D2D communications as an example to implement our generalized social enabled wireless network framework. Multi-dimensional generalization, discussed later, allows us to include energy efficiency, location, and radio range, and more relationships among devices in optimizing radio resource allocation. D2D underlay and overlay [14] are two typical modes in terms of different spectrum sharing. In D2D underlay, co-channel interference should be considered in addition to the D2D overlay cases. We note that the secrecy-oriented objectives are also applicable for D2D cases. Without loss of generality, we consider D2D underlay here as an illustration, which can be viewed as an extension of multi-pairing in wireless mesh networks. Therefore, after appropriate admission control and power control to form stable and reliable D2D pairs, maximum weighted bipartite matching was again employed to maximize overall network throughput [15]. Accordingly, Eq. 1 can be regarded as the achievable data rate of D2D link j formed by DUE m and n, sharing the resource of CUE i, and rewritten as

$$R(i,j) = s_2(i,j) \sum_j q_{i,j} s_1(m,n) \cdot \log_2(1+\xi_j^d), \quad (2)$$

where ξ_j^d is the SINR of D2D link j when it shares the same resource with CUE i. $q_{i,j}$ is the matching indicator variable between CUE i and D2D link j, $q_{i,j} = 1$ indicates that D2D link j

reuses the resource of CUE i; otherwise, $q_{i,j} = 0$. P_i^c and P_j^d are the transmit power of CUE i and the transmitter of D2D link j, respectively. Note that $s_2(i, j)$ represents the social trust between CUE i and D2D link j, which are sharing the same resource, that is, the probability that CUE i would be willing to share its resource with D2D link j.

Recall that we define stage 1 as the process of D2D partner selection regarding to high-layer preference factors, and stage 2 is defined as the procedure of CUE and D2D-link matching depending on wireless resource level. We shall discuss the achievable data rate for D2D links by taking different factors into consideration in stages 1 and 2, with different number of CUEs and D2D links, and their corresponding transmit power limitation.

For notational simplicity, we use PDO-1 and PDO-2 to denote the cases when considering factors in the physical domain only in stages 1 and 2, respectively. Similarly, MD-1 indicates that multi-domain factors are taken into account in stage 1, targeting physical distance, mutual social trust, interest similarity, and cache capability of DUEs. However, both MD-2-A and MD-2-E consider multi-domain factors, which are physical distance and social trust between CUEs and D2D links in stage 2. Furthermore, we use MD-2-A to indicate the case with accurate social trust between CUEs and D2D links, whereas social trust estimation error exists in MD-2-E.

For further analysis, we combine the cases discussed above, as shown in Fig. 5. For example, PDO-1 + MD-2-A implies that only physical domain factors are considered in stage 1, and multi-domain factors are taken into consideration in stage 2 with accurate social trust. Accordingly, the closest M pairs of DUEs are chosen to form D2D links in the PDO-1 + MD-2-A strategy. However, the MD-1 + MD-2-A strategy forms D2D links with multiple objective preferences, referring to Eq. 1. Furthermore, pairwise power optimization carries out the maximization of $R(i, j)$ after the D2D pairing, before optimizing the overall sum data rate of D2D links via optimal matching.

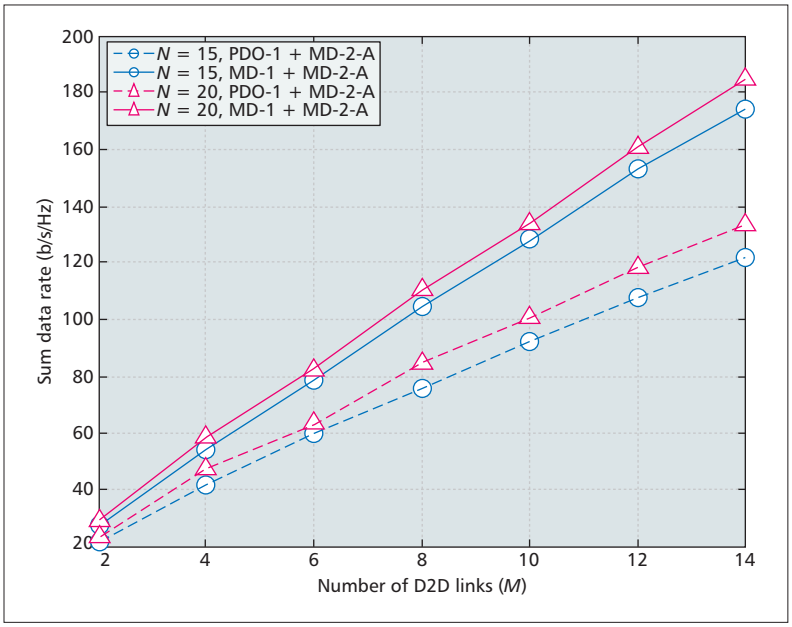


Figure 5. Sum system data rate with different number of CUEs and D2D links. We consider a single cell of radius R, where N conventional CUEs and M D2D links are uniformly distributed in the cell with the radius of R = 300 m. The noise power for each channel is assumed as $\sigma_N^2 = -96$ dBm. The distance between two DUEs to form a potential D2D link is less than the maximum D2D tolerant distance, $d_{\max}^d = 30$ m, and CUE i can only be regarded as a resource sharing candidate of D2D link j if the corresponding distance is no less than the minimum reuse distance depending on the maximum tolerant co-channel interference, $d_{\min}^c = 100$ m. Note that the maximum transmit power of CUEs and D2D transmitters are $P_{i,\min}^c = 24$ dBm and $P_{j,\min}^d = 19$ dBm, respectively.

Figure 5 shows that the sum rate of the system grows with the increasing number of CUEs and successful pairing D2D links, because more D2D links and CUEs make it possible for better resource allocation, thereby leading to larger sum data rate. We find that the MD-1 + MD-2-A strategy achieves a significantly higher rate compared to PDO-1 + MD-2-A, since the latter has higher probability of choosing D2D users with low $s_1(m, n)$, thereby making it difficult to guarantee the value of $R(i, j)$. To overcome this drawback, the MD-1 + MD-2-A strategy chooses D2D users by considering both physical and social-oriented (multiple) factors to form D2D links, which ensures the superiority of system performance to a large extent.

Different from Fig. 5, Fig. 6 demonstrates that larger $P_{i,\max}^c$ and $P_{j,\max}^d$ allow each pair of D2D links to reach a higher rate and subsequently a higher sum data rate. It can also be easily observed from Fig. 6 that MD-1 + MD-2-A and MD-1 + MD-2-E outperform MD-1 + PDO-2, because the matching results in MD-1 + PDO-2 may suffer from poor utilization of social information, further resulting in the less favorable system reward, whereas MD-1 + MD-2-A and MD-1 + MD-2-E give full consideration to the social trust between CUEs and D2D links. Thus, the matching result guarantees the maximization of the achievable sum data rate and relative rate for D2D links as well. In addition, the gap between those curves that are holding estimated social trust and accurate parameters is not huge, and can be ignored by considering multiple social oriented factors. In other words, the robustness of the multi-domain strategy outweighs that of the MD-1 + PDO-2 strategy.

Conclusions

In this article, we have presented a novel graph-theoretical framework for resource allocation by exploiting the inherent interplay between social networks and wireless communications. We have provided a systematic approach for multi-objective optimization and then applying bipartite matching to enhance the design of wireless networks. Specifically, we propose hierarchical bipartite-based pairing and clustering by jointly considering multi-dimensional factors in terms of two stages, selecting the partner first and optimizing the resource allocation after.

Numerical results have demonstrated the merits of our proposed concept. Further investigations remain open, such as improving system robustness, since the key factors considered for

resource allocation and content sharing can be time varying and prone to estimation errors in wireless networks.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundations of China under Grant No. 61372117, 61201150, and 61571056; the National High Technology Research and Development Program of China under Grant No. 2014AA01A701, and the Beijing Higher Education Young Elite Teacher Project under Grant No. YETP0442, the Ministry of Science and Technology under contract 104-2221-E-002-082, and a research grant from the Institute of Information Industry, Taiwan.

REFERENCES

[1] K. C. Chen, M. Chiang, and H. Vincent Poor, "From Technological Networks to Social Networks," *IEEE JSAC*, vol. 31, no. 9, 2013, pp. 548–72.

[2] T. Sönmez and M. U. Ünver, "Matching, Allocation, and Exchange of Discrete Resources," *Handbook of Social Economics*, vol. 1A, Ch. 17, J. Benhabib, M. O. Jackson, and A. Bisin, Eds., 2011, pp. 781-852.

[3] L. Wang and H. Wu, "Fast Pairing of Device-to-Device Link Underlay for Spectrum Sharing with Cellular Users," *IEEE Commun. Lett.*, vol. 18, no. 10, 2014., pp. 1803–06.

[4] F. K. Hwang, "Control Algorithms for Rearrangeable Clos Networks," *IEEE Trans. Commun.*, vol. 31, no. 8, 1983, pp. 952–54

[5] X. Niu *et al.*, "Predicting Image Popularity in an Incomplete Social Media Community by a Weighted Bi-Partite Graph," *Proc. IEEE Int'l. Conf. Multimedia and Expo*, 2012, pp. 735–40.

[6] J. Yang and Z. Fei, "Bipartite Graph Based Dynamic Spectrum Allocation for Wireless Mesh Networks," *Proc. IEEE Int'l. Conf. Distrib. Comp. Sys.*, 2008, pp. 96–101.

[7] B. Bai *et al.*, "Max-Matching Diversity in OFDMA Systems," *IEEE Trans. Commun.*, vol. 58, no. 4, Apr. 2010, pp. 1161–71.

[8] Y. Liu and M. Tao, "Optimal Channel and Relay Assignment in OFDM-Based Multi-Relay Multi-Pair Two-Way Communication Networks," *IEEE Trans. Commun.*, vol. 60, no. 2, Feb. 2012, pp. 317–21.

[9] L. Wang, H. Tang, and M. Cjemy, "Device-to-Device Link Admission Policy Based on Social Interaction Information," *IEEE Trans. Vehic. Tech.*, vol. 64, no. 9, Sept. 2015, pp. 4180–86.

[10] M. Doostmohammadian and U. A. Khan, "Graph-Theoretic Distributed Inference in Social Networks," *J. Sel. Topics Signal Process*. vol. 8, no. 4, Aug. 2014, pp. 613–23.

[11] J. Iqbal and P. Giaccone, "Interest-Based Cooperative Caching in Multi-Hop Wireless Networks," *Proc. IEEE GLOBECOM 2013 Wksp.*, 2013, pp. 617–22.

[12] H. J. Kang and C. G. Kang, "Mobile Device-to-Device (D2D) Content Delivery Networking: A Design and Optimization Framework," *J. Commun. Networks*, vol. 16, no. 5, Oct. 2014, pp. 568–77.

[13] X. Wang *et al.*, "Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 131–39.

[14] X. Lin, J. G. Andrews, and A. Ghosh, "Spectrum Sharing for Device-to-Device Communication in Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, Dec. 2014, pp. 6727–40.

[15] D. Feng *et al.*, "Device-to-Device Communications Underlaying Cellular Networks," *IEEE Trans. Commun.*, vol. 61, no. 8, Aug. 2013, pp. 3541–51.

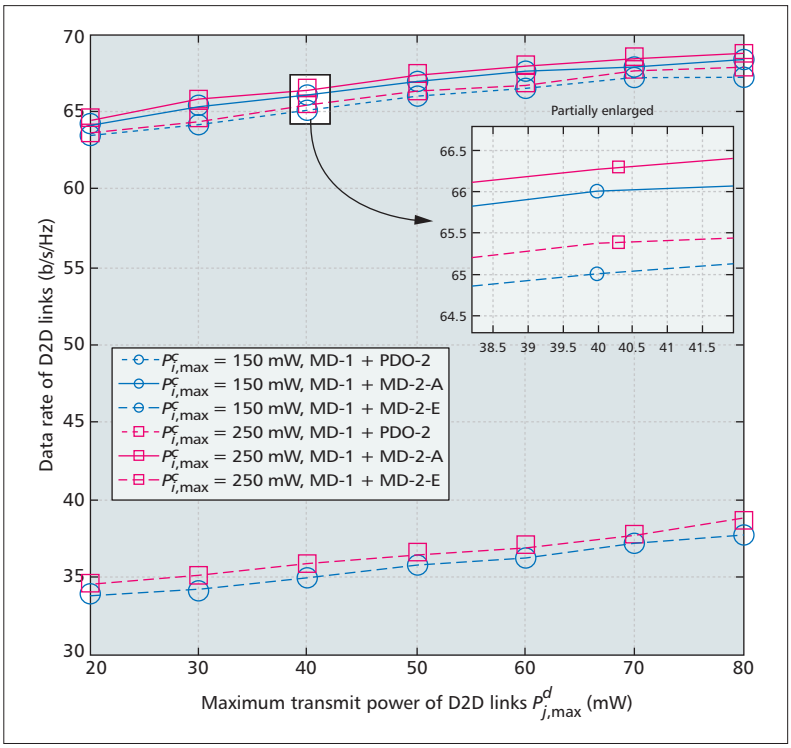


Figure 6. Sum data rate for D2D links with different maximum transmit power of CUEs and D2D transmitters.

BIOGRAPHIES

LI WANG [S'08, M'14] is an associate professor in the School of Electronic Engineering, Beijing University of Post and Telecommunications (BUPT), where she leads the Lab of High Performance Computing and Networks. She received her Ph.D. degree in 2009 from BUPT. She received the 2013 Beijing Young Elite Educator for Higher Education Award. From December 2013 to January 2015, she held a visiting research appointment at the School of Electrical and Computer Engineering at Georgia Tech, Atlanta. From August 2015 to November 2015, she was a guest researcher with the Department of Signals and Systems at Chalmers University of Technology, Gothenburg, Sweden. Her research interests include wireless networking, secure communications, device-to-device communication systems, and peer-to-peer networks. She has served or is serving on the Technical Program Committees of IEEE CCNC 2009, IEEE CCNC 2010, IEEE WCSP 2013, IEEE GLOBECOM 2014, IEEE WCNC 2015 , IEEE ICC 2015, IEEE ICNC 2015, IEEE ICCC 2015, IEEE GLOBECOM 2015, IEEE ICNC 2016, IEEE ICC 2016, and so on.

HUAQING WU [S'15]reveived her Bachelor's degree from BUPT) in 2014. Since 2014, she has been pursuing graduate studies at BUPT. Her research interests include wireless networking, secure communications, and device-to-deveice communication systems.

WEI WANG [S'08, M'10, SM'15] received his B.S. and Ph.D. degrees from BUPT in 2004 and 2009, respectively. Currently, he is an associate professor with the Department of Information Science and Electronic Engineering, Zhejiang University, China. From September 2007 to September 2008, he was a visiting student at the University of Michigan, Ann Arbor. From February 2013 to February 2015, he was a Hong Kong Scholar with the Hong Kong University of Science and Technology. His research interests mainly focus on cognitive radio networks, green communications, and radio resource allocation for wireless networks.

KWANG-CHENG CHEN [M'89, SM'94, F'07] is the Distinguished Professor, Graduate Institute of Communication Engineering, National Taiwan University. He has received a number of awards, including the 2011 IEEE ComSoc WTC Recognition Award, 2014 IEEE Jack Neubauer Memorial Award, and 2014 IEEE ComSoc AP Outstanding Paper Award, and has co-authored a few award-winning papers published in IEEE ComSoc conferences. His research interests include wireless communications, social networks and network science, and data analytics.

Location-Based Social Video Sharing over Next Generation Cellular Networks

Abhishek Roy, Pradipta De, and Navrati Saxena

ABSTRACT

The popularity of video sharing over social networks is straining service providers to maintain high quality of service. With the introduction of live video streaming over social networks, video sharing is truly becoming social and real time. However, existing video distribution architecture, split across content providers and mobile network operators, cannot leverage cross-provider information. We argue that MNOs can explore the knowledge of users' location to route video traffic intelligently. This can enable better quality of experience for live social video sharing. We show different scenarios in existing cellular networks, where user experience can suffer due to current video distribution mechanism. We propose that use of the location information of the receiver with respect to the sender can be exploited by next generation cellular networks for improving video streaming quality. Our simulation experiments validate that on all QoE metrics there is opportunity for significant improvement in live video streaming by using relative locations of the receivers and senders.

INTRODUCTION

Video streaming has long been touted as one of the killer applications for mobile devices. But limitations of technology, whether due to limited resources on mobile devices or inadequate infrastructure support, have hindered the rise of video streaming in social media space. As the network capacity has evolved, and mobile devices have become more powerful, the scenario is changing fast. We are witnessing the rise of several companies, such as Vine and Instagram, which have popularized short video sharing, and Meerkat, Twitter's Periscope, and Twitch.tv, which are focused on live streaming of user generated video content. In 2011, Erman *et al.* reported that 40 percent of cellular traffic is composed of video streaming [1]. In the Internet, video streaming constitutes 66 percent of the total traffic, according to Cisco Visual Networking Index: Forecast and Methodology, 2013–2018, 2014.[1] Social video streaming is certain to fuel the growth of video traffic further.

Social video sharing is a natural progression of text- and image-based social networks. Strover and Moner studied the generational shift in viewing media. They concluded that young people, especially those in colleges with high-speed Internet connectivity, are contributing toward sharing and creating new content [2]. They form the new generation of "ProdUsers," where the roles of consumers and users of content are intertwined, thereby leading to the new hybrid role [3]. The ProdUsers' desire to share content is met by the market with novel applications, the latest among them being live video sharing over social networks. This mode of live video sharing truly embodies social sharing as it connects groups in real time. We consider live streaming of user generated video to social groups as "social video sharing."

The traditional Internet architecture follows a layered design, where dedicated service providers cater specific services, like content management, content distribution, and infrastructure support. As shown in Fig. 1, there are three key entities in the content distribution: content providers, content distribution networks (CDNs), and mobile network operators (MNO). Often these providers work in isolation with limited cross-provider information sharing. A user-generated video content must travel upstream to the content provider, and then flow downstream to the consumers. At each layer, the aim of the provider is to adapt independently to deliver better quality of experience (QoE) to the user. However, lack of cooperation across providers leads to complex and sub-optimal adaptation techniques, as shown by [4].

The benefits of information sharing across providers is not unknown. Jiang *et al.* observed that Internet service providers are taking on the lucrative role of content providers. Merging of the roles can lead to better coordination in traffic engineering, which can be especially beneficial for high-volume latency-sensitive video traffic [5]. Google and Amazon are more recent examples of providers, which are taking on multiple roles to enhance their service quality. There have also been alliances across CDNs and MNOs to improve service delivery [6]. It is not far-fetched to assume that network architectures will

Abhishek Roy is with Samsung Electronics.

Pradipta De is with the State University of New York, Korea and Stonybrook University.

Navrati Saxena is with Sungkyunkwan University.

[1] http://tinyurl.com/mev32z8

0163-6804/15/$25.00 © 2015 IEEE

evolve to allow more information sharing across providers [7]. With this shift in network design toward greater collaboration across providers, it is expected that in the future, video distribution architecture will also evolve and share information across providers to deliver higher QoE to the users. This is our assumption for the next generation cellular network (NGCN).

This article explores avenues for enhancements in video distribution over cellular networks from the perspective of the MNOs. We specifically focus on live video streaming, where MNOs can leverage the users' location information within cells to improve QoE. In the next section we explore the challenges in current video distribution. Next, we present multiple scenarios that we believe open up opportunities to enhance quality of social video sharing over fourth generation (4G) Long Term Evolution (LTE) networks. We show the benefits of bypassing the content providers in the video distribution path, especially when content providers and MNOs can share information about the senders and receivers of the live video stream. Our observations are validated using simulation experiments.

Challenges in Social Video Sharing over Existing Cellular Networks

Figure 2 shows the video distribution process over 4G LTE cellular networks. The evolved NodeB (eNB) in 4G cellular networks is the evolution of the base station (or Node B of 3G) for all radio network functions, like scheduling, radio resource management, and security. Unlike 2G/3G networks, 4G eNBs do not have any separate controller and are interconnected with each other by X2 interfaces. eNBs are also connected with the System Architecture Evolution (SAE) gateway in the Evolved Packet Core (EPC) via the S1 interface. EPC is the all-IP evolution of the general packet radio service (GPRS) core network. It is responsible for all 4G core network operations, like connection (session) establishment, overall control, mobility management, quality of service (QoS), and policy enforcement. The SAE gateway is the terminal node of 4G cellular networks for packet transmission to and from external networks. Traditional video servers connect to the SAE gateway to deliver content to the users.

In existing 4G LTE systems, video traffic is routed to the video content server via LTE eNB and EPC networks. While LTE cell capacity is a few hundred users, a single EPC and a video server typically serve hundreds of eNBs, servicing thousands of users. Thus, even with Gigabit Ethernet link capacity, the network links between the video server and the core network, as well as the downstream links to the users, may saturate quickly. With high data traffic required for streaming, this can degrade streaming quality, especially when multiple users are streaming simultaneously.

The increasing popularity of mobile video sharing through social networking websites and mobile-originated, peer-to-peer video sharing are threatening to make the problem even worse. Unfortunately, legacy cellular 4G networks do not exploit the location of social video users while routing the packets. Video traffic is sent upstream to the content servers, and then sent downstream from the video servers to the receivers. Streaming to the content provider is necessary if the video needs to be stored, but for faster delivery to users, it can be enhanced by generating another stream which can follow a different network path that leverages receiver location with respect to the sender. We consider the scenario where the video content creator is streaming live to a receiver. Depending on the relative location of the receivers, the MNO can use the location information to route the video more intelligently than the content provider.

With the impending wireless evolution toward 5G, there is also a shift toward more device-centric architecture from traditional network-centric architecture design [8]. Evolved multimedia broadcast and multicast services (eMBMS) [9], multicast broadcast single frequency network (MBSFN) [9], and emerging device-to device (D2D) communications [10] can also be explored for better utilization of cellular network resources. While eMBMS and MBSFN are currently tailor-made for server-based video applications, D2D communications are expected to have native support in 5G wireless by exploiting the advantage of proximity-based direct mobile-to-mobile communications. NGCNs are expected to explore these emerging communication technologies to develop the more efficient social video sharing architecture needed to improve the video QoE.

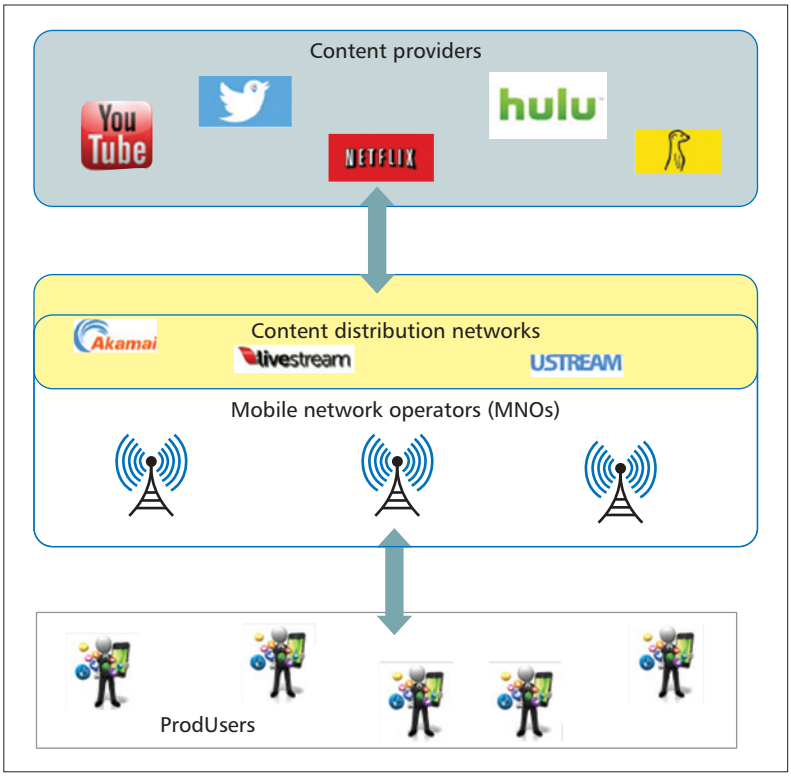


Figure 1. Different players in a typical video content distribution process.

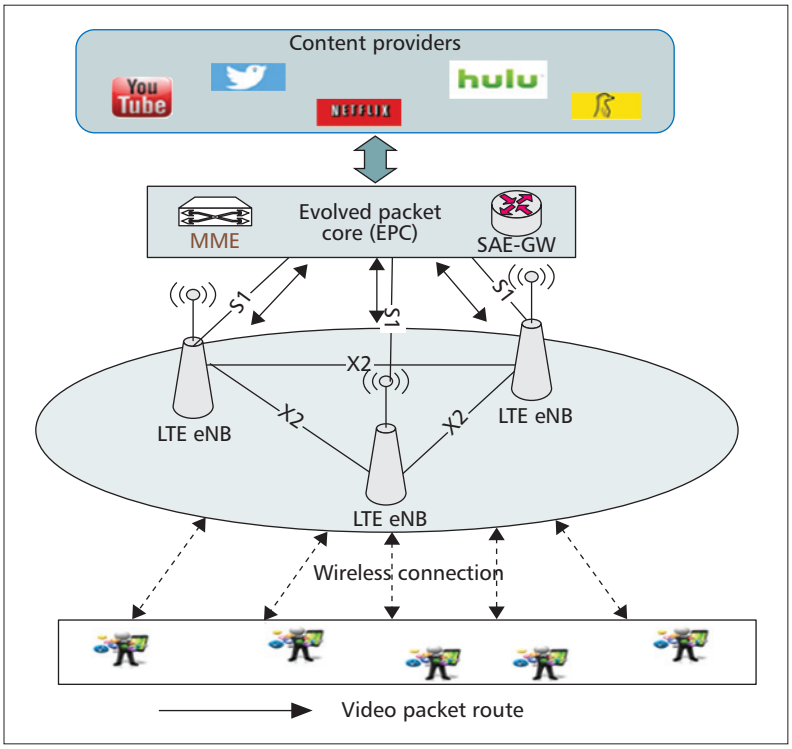


Figure 2. LTE network architecture showing the organization of the network infrastructure elements, like eNB and EPC, and the connection to the servers hosted by the content providers.

Location-Based Social Video Sharing

We present an alternative delivery mechanism for live social video streams that can be supported by MNOs using additional information about users that are subscribed to the MNOs. Depending on the location of the social video receivers, an NGCN can reroute social video streams directly from the nearest network node that is shared by the video sender and receivers. This will help the video traffic bypass the network elements and links, thereby significantly reducing network load, delay, and jitter, and improving the video QoE, as well as cell capacity.

In the proposed video distribution scenarios, as shown in Fig. 3, content providers must share the user identity with the MNO. There is an initial signaling phase, when the MNO must map the user identity to the user's location within the network. Once the mapping is established, the streams can be rerouted using the mapping without reaching the content provider. In this model, the streams are not stored by the content provider for streaming on demand. Meerkat currently uses a similar model, where receivers can only join a live stream but cannot use playback. At present, in LTE networks, during the session establishment phase, gateways use IP and transport layer information, like service type, source and destination address, and port numbers, to map every session to a unique Enhanced Packet System (EPS) bearer with a specific QoS class identifier (QCI) [11]. The bearer information is conveyed from the gateways to the corresponding eNBs and UEs by exchanging radio resource control (RRC) connection setup messages. Once the RRC connection and dedicated bearer are established, every UE (in the uplink) and eNB (in the downlink) map the dedicated EPS bearer into a unique logical channel. This logical channel identification (LCID) is contained in the LTE medium access control (MAC) header. The eNB can uniquely identify a video connection and its QCI using this LCID. Once the identity of the sender and receivers are shared by the content provider, the LCID can be mapped to the identities. Subsequently, a video stream can be rerouted by the eNB without reaching the content provider. This introduces additional IP lookup overhead at the eNBs. However, LTE eNBs, equipped with high-performance processors such as Octeon 8-core processors, are capable of processing high traffic rates with negligible additional latency for IP lookup.

Next we use four different scenarios to illustrate how video receivers' location can be exploited by the MNO for more efficient social video sharing.

- As shown in Fig. 3a, if the social video source and the intended receivers are in the same cell, the video packets from the video source will traverse the uplink to the LTE eNB. The eNB, in turn, will directly transmit the video packets over the downlink to the video receivers. Thus, the system can shorten the route of the packets between the sender and the receiver by eliminating the path from the eNB to the video server (seNB → core network → video server → core network → teNB). This results in significant improvement in delay and jitter of video streams.
- Recently, D2D communication [8, 10] among users in close proximity is emerging as a communication mechanism in 5G standards. As shown in Fig. 3b, social video sharing is expected to efficiently explore this proximity-based D2D communication to improve video delivery.
- On the other hand, if the social video source and receivers are not in the same cell, but in a nearby cell, we can exploit the X2 interfaces of 4G eNBs to efficiently reroute the video packets. As shown in Fig. 3c, after receiving the uplink video packets from the video content generator, using the X2 interface, the source eNB (seNB) can reroute these video packets to the target eNBs (teNBs) of the receiving cell. The teNB, in turn, will transmit the video packets in the downlink to the video receivers. The system can bypass the traversal of video packets along the seNB → core network → video server → core network → teNB routes, thereby reducing video packet delay and jitter.
- Finally, as shown in Fig. 3d, if the social video source and receivers are in distant cells across different EPCs, we can explore LTE's S1 interfaces, as well as inter-EPC (core network) links. Using S1 interfaces, the seNB can route the video packets to the source SAE-GW (in the source core network). The source SAE-GW uses the inter-EPC links to reroute packets to the target SAE-GW (in the target core network). The target SAE-GW, now uses S1-interfaces to reroute the packet to the teNB. Thus, although the video receivers are in distant cells (across different core networks), next generation cellular networks can still avoid the external network,

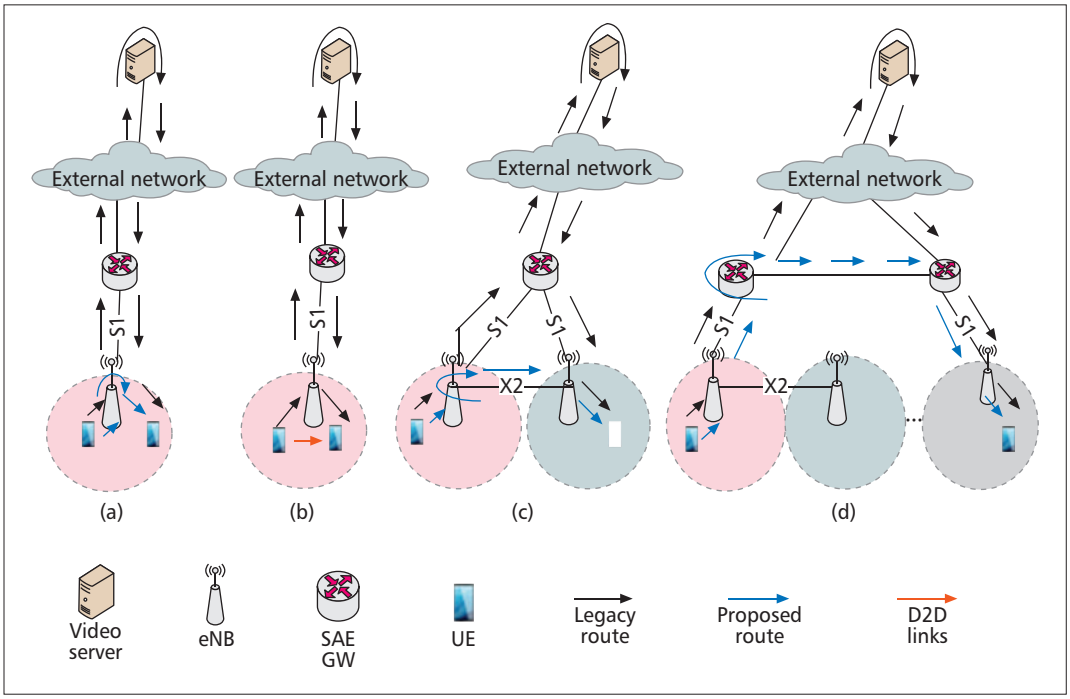


Figure 3. Scenarios in NGCN architecture for distributing social video streams where the sender and recievers belonging to a sender's social group are spread across different cells: a) video source and destination in the same cell; b) video source and destinations with D2D communications links; c) video source and destinations in different cells fo same core network; d) social video source and destinations in different core networks.

The NGCN options involve unicast video transmission between the video content producer and receiver. However, unicast video sharing incurs separate wireless radio resource consumption for every video receiver. Naturally, with a resource constrained cellular systems, this significantly reduces the cell capacity, measured in terms of number of receivers supported.

relieve the video server, and achieve significant improvements in social video sharing performance.

Note that the social network of the video content producer can be geographically spread over a large region, spanning multiple cells and core networks. Hence, in a real system, streaming user generated video content can encompass all four scenarios. While in traditional video transmission schemes the video server is responsible for all video transcoding, we consider that, as roles merge, the MNOs can equip eNBs with basic video transcoding and adaptation features. In this design, as transcoding workload is transferred to the eNBs from the content provider, power consumption at content providers will decrease at the cost of higher consumption at the eNBs. But overall reduction in network traffic should be beneficial in terms of total power consumption. Also, low-power video transcoding techniques should help in reducing the energy footprint for streaming.

Another practical concern relates to the use of location information of users by the MNOs. In the proposed architecture, MNOs do not need to share the location information with the content providers. Thus, no additional vulnerability is introduced, as compared to the privacy issues identified by Hahn *et al.* [12].

Next Generation Multicast for Social Video Sharing

The NGCN options involve unicast video transmission between the video content producer and receiver. However, unicast video sharing incurs separate wireless radio resource consumption for every video receiver. Naturally, with resource constrained cellular systems, this significantly reduces the cell capacity, measured in terms of number of receivers supported. For example, in a 20 MHz 2×2 multiple-input multiple-output (MIMO)-based 4G LTE systems, the theoretical peak data rates are 75 Mb/s in UL and 150 Mb/s in DL. Thus, to share a $\sim$ 2 Mb/s high definition (HD) video, the theoretical maximum cell capacity is upper bounded by $\lfloor 75/2 \rfloor = 37$ video receivers. With multiple senders, spread across different locations of the cell, as well as non-video traffic workload, the actual cell capacity of a real cellular network would be much lower.

In order to improve the cell capacity for video sharing, NGCNs need to explore efficient user-initiated video multicast. Note that eMBMS [9] have already been standardized from Third Generation Partnership Project (3GPP) LTE Release 11.0 onward. However, legacy eMBMS systems currently support only server-based broadcast and multicast services across multiple users. This eMBMS in NGCNs are expected to gradually evolve to incorporate user-initiated video multicast for cell capacity improvement in social video sharing.

As shown in Fig. 4a, in this user-initiated multicast, the video sender and receivers form multiple multicast groups across different cells. Instead of multiple unicast connections, a sender can use a single uplink connection with its source eNB, which will multicast the video packets to the group of receivers within its coverage area. The source eNB also sends the video packets to the target eNBs over X2 interfaces or S1-SAE-

Typically, in a deployed network, the network vendor and the operator perform the cell planning and multicast group design by choosing a suitable worst channel condition. Further optimization is possible by designing multiple multicast groups within cells by grouping UEs with similar channel conditions.

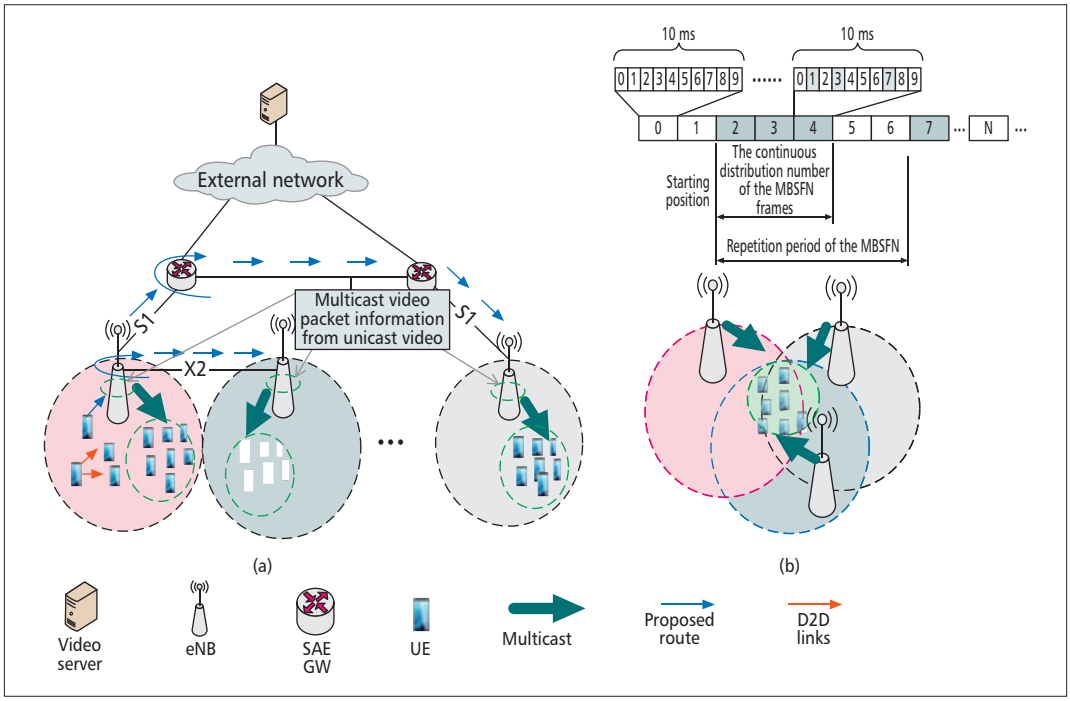


Figure 4. Use of multicast features in NGCNs, which allows efficient infrastructure resource usage to stream live video from a sender to its social group of receivers: a) social video sharing exploring user-initiated multicast; b) exploiting MBSFN for social video sharing.

GW interfaces. The target eNBs multicast the video packets to respective social groups within its coverage area. As multicast inherently shares radio resource among the receiver groups, this will significantly increase the network capacity of social video sharing. However, as legacy point-to-multipoint multicast involves multiple receivers, the eNB typically uses the user with the worst RF condition to determine the transmission rate. Selecting the transmission rate with the worst RF condition, can affect the QoE of UEs with good channel conditions. But, the trade-off is paid by supporting larger number of users in the multicast group. Typically, in a deployed network, the network vendor and the operator perform the cell planning and multicast group design by choosing a suitable worst channel condition. Further optimization is possible by designing multiple multicast groups within cells by grouping user equipments (UEs) with similar channel conditions.

Furthermore, NGCNs are evolving to exploit the MBSFN. As shown in Fig. 4b, MBSFNs enable multiple neighboring eNBs to cooperate for video packet transmissions to the same multicast group. Now, to the video receiver, transmissions received from multiple eNBs appear as a single transmission, subject to multipath propagation. This overcomes the shortcomings of legacy point-to-multipoint multicast by combining simultaneously received signals from different eNBs and transforming the corresponding portion of the destructive interferences into constructive ones (i.e., gain). Naturally, it improves the overall RF conditions experienced by all the social video receivers, especially the receivers at the cell edge. Thus, the difference between the best and worst RF conditions are largely reduced. Improvement in RF conditions enable the eNB to use less radio resource for the same video transmissions, thereby increasing the capacity even more. As shown in Fig. 4b, a 10 ms LTE MAC frame is made up of 10 1-ms subframes. Out of these 10 subframes, up to maximum 6 subframes could be reserved for eMBMS. Now, neighboring eNBs collaborate to form a cooperative downlink transmission frames at a regular interval, called the repetition period of an MBSFN.

Performance Evaluation

In this section we first discuss our simulation platform, major simulation parameters, and assumption. Subsequently, we show the simulation results with "user-initiated" unicast social video sharing. Finally, we extend our simulation to demonstrate user-initiated multicast for social video sharing.

In order to validate our proposed framework, we have developed an OPNET-based realistic urban LTE network model, involving multiple EPCs and eNBs spread across different cities. Physical parameters and an urban macro LTE channel model, specified in 3GPP physical layer (PHY) specifications [13] are used. Table 1 highlights the major LTE system and radio parameters used by us. The cell capacity is evaluated using standard video formats, like standard definition (SD), HD, full HD (FHD), and quadruple HD (QHD), with 1, 2, 4, and 8 Mb/s rates, respectively. Apart from the video bit rate and capacity experiments, we model two users sharing live HD social video streams at 2 Mb/s rate, simultaneously with 8 receivers (i.e., a total of 16 video receivers). These eight receivers are uni-

formly distributed across four scenarios (same cell, D2D, different cell, and different core), mentioned in the previous section, such that each scenario consists of two video receivers. We have considered social video streams of 5 min duration, as shown by Shen *et al.* to be the typical size of social video shares [14]. We also consider a set of new QoE metrics, which are considered to be more effective in capturing user experience [15]. These QoE metrics are explained below:

- *Playback start time* refers to the delay before a video stream starts playback.
- *Total rebuffering time* is defined as the total intermediate buffering time experienced after initiating the video playback session.
- *Rebuffering frequency* calculates the total number of intermediate buffering events or pauses due to buffer overrun that happen during a video playback.

Figure 5a shows the streaming video bit rate supported with two video sources transmitting to an increasing number of social video receivers spread across different cells. With increasing receivers, the supported bit rate starts dropping, thereby degrading the video quality. With our proposed video routing solution for NGCN, it is possible to support video bit rates 30 percent higher than those in the existing network for the same number of receivers. This is possible since the bottleneck between the eNB and the content provider is avoided in our scheme.

Figure 5b demonstrates that intelligent video traffic rerouting can reduce the packet jitter of HD video streams from 70 ms (median value) to 28, 20, 11, and 8 ms for users who are spread across different core networks, different cells, same cells, and D2D users, respectively. Note that when the stream is routed through the content provider, the jitter for users, irrespective of their proximity to the sender, will be similar.

Figure 5c shows that NGCNs can improve HD video playback start time from 5.9 s to 4.8, 3.8, 2.7, and 2.5 s for users across different core networks, different cells, same cells, and D2D users, respectively. Figure 5d shows how total rebuffering time increases with HD video playback. In a 5 min social video, legacy sharing experiences almost 1 min of total rebuffering time. On the other hand, total video rebuffering time reduces to 20 s even when the receiver is farthest from the sender, which is a 66 percent improvement. When the receiver is within the same cell and using D2D communication, the rebuffering time reduces to less than 10 s.

It is important to note that rebuffering time does not indicate the number of times there was a buffer overrun during viewing, and the user had to wait for rebuffering. Rebuffering leads to stalls during playback and is a bigger source of annoyance that can make users abort the stream. Hence, besides total rebuffering time, the number of rebuffering events is also an important video QoE metric. Figure 5e depicts that in the same 5 min HD video playback, a legacy system incurs around 20 rebufferings. On the other hand, depending on the receivers' locations, NGCNs can reduce rebuffering events to as low as six when the receivers can use D2D communication. Both the total rebuffering time and number of rebuffering events reduce with receiver proximity.

LTE radio access network models	
LTE network operating frequency	2 GHz
Channel bandwidth	20 MHz
LTE channel model	Urban macro [13]
Penetration loss (I)	20 dB
Cell radius (R)	500 m
Attenuation factor	$I + 37.6\log_{10}R$
Path loss compensation	0.8
eNB system models	
eNB's max. Tx power	43 dBm [13]
eNB's idle power	0.19 dBm [13]
Mobile's max power	20 dBm
Video Quality and BitRates	
Standard definition	1 Mb/s
High definition	2 Mb/s
Full high definition	4 Mb/s
Quadruple high definition	8 Mb/s

Table 1. LTE radio and system parameters.

Finally, Fig. 5f shows the cell capacity of social video sharing with a predefined video quality for all four types of video mentioned in Table 1. It points out that by exploring user-initiated multicast features, point-to-multipoint eMBMS and MBSFN can increase the cell capacity by almost 9 and 12 times, respectively, while maintaining the same video quality and user perceived QoE metrics for SD, HD, FHD, and QHD quality video. eMBMS achieves this significant capacity gain by transmitting a single uplink stream from a video source to the eNB and subsequently exploiting downlink resource sharing by multicast video packets. MBSFN further increases this capacity by exploring cooperation among multiple neighboring eNBs.

Conclusion

Lack of information sharing across content providers and mobile network operators can lead to suboptimal quality of experience for users. With the growing focus on video streaming in social networks, network traffic is poised to surge significantly. There have been many proposals that aim to improve video QoE using implicit performance indicators, like buffer status and channel conditions. However, cross-provider collaboration can

It is important to note that rebuffering time does not indicate number of times there was a buffer overrun during viewing, and the user had to wait for rebuffering. Rebuffering leads to stalls during playback and is a bigger source of annoyance that can make users abort the stream.

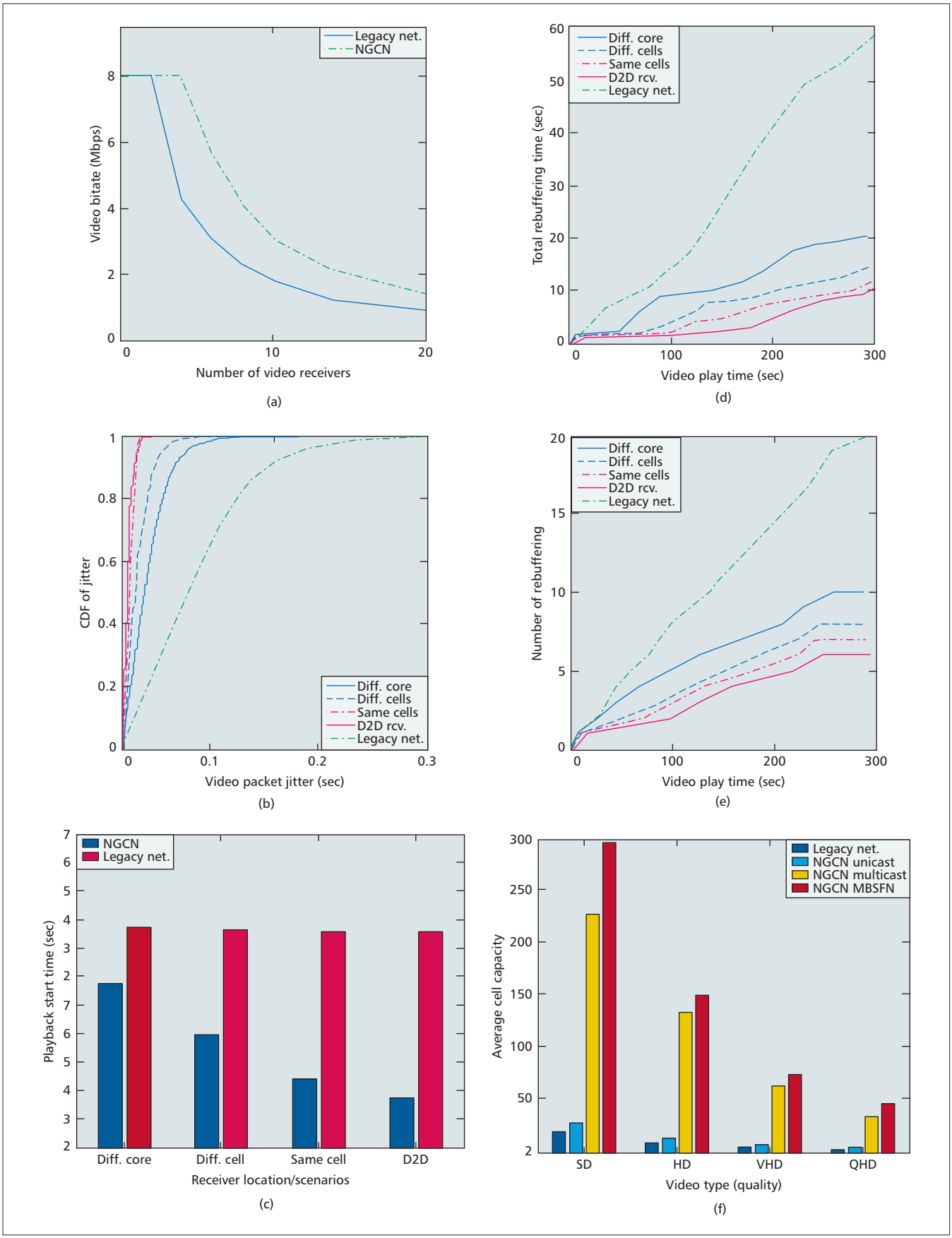


Figure 5. Simulation results showing the impact of location based re-routing of live social video streams. We show how different user perceived QoE metrics, viz. bitrate, jitter, start time, rebuffering time, and rebuffering events, are affected in different scenarios that represent different user locations. a) Video bitrate (quality); b) video jitter; c) playback start time; d) rebuffering time; e) rebuffering events; f) video sharing capacity.

lead to significant gains. In this article, we show that location information of the sender and receiver can be effectively utilized by MNOs for live streaming in mobile-device-centric social network sharing. Using simulation experiments, we show that routing live streams through the MNO infrastructure can significantly improve QoE with respect to supporting higher bit rate streams, lower jitter, reduced video playback start time, and smoother playback. The content provider can act as the controller that initiates the connection and shares the stream information with the MNO. The MNO leverages the proximity information of users to deliver live streams in an effective manner by routing through its nodes, which bypass the content provider network. We have also highlighted how features such as D2D can play an important role in video streaming in next generation cellular networks.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the ministry of Education (S-2015-0849-000) and by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the "IT Consilience Creative Program" (NIPA-2013-H0203-13-1001) supervised by the NIPA (National IT Industry Promotion Agency).

References

[1] J. Erman *et al.*, "Over the Top Video: The Gorilla in Cellular Networks," *Proc. 2011 ACM SIGCOMM Conf. Internet Measurement*, 2011, pp. 127–36.

[2] S. Strover and W. Moner, "The Contours of On-Demand Viewing," Holt and Sanson, Eds., *Connected Viewing: Selling, Streaming, and Sharing Media in the Digital Era*, 2013, pp. 234–54.

[3] A. Bruns, "Produsage," *Proc. 6th ACM SIGCHI Conf. Creativity & Cognition*, 2007.

[4] X. K. Zou *et al.*, "Can Accurate Predictions Improve Video Streaming in Cellular Networks?" *Proc. 14th ACM Workshop on Hot Topics in Networks*, 2015.

[5] W. Jiang *et al.*, "Cooperative Content Distribution and Traffic Engineering in An ISP Network," *ACM SIGMETRICS Perf. Eval. Rev.*, 2009.

[6] B. Frank *et al.*, "Pushing CDN-ISP Collaboration to the Limit," *ACM SIGCOMM Comp. Commun. Rev.*, vol. 43, no. 3, 2013, pp. 34–44.

[7] J. Jiang *et al.*, "Eona: Experience-Oriented Network Architecture," *Proc. 13th ACM Wksp. Hot Topics in Networks*, 2014.

[8] F. Boccardi *et al.*, "Five Disruptive Technology Directions for 5G," *IEEE Commun. Mag.*, vol. 9, no. 1, Feb. 2014.

[9] 3GPP, "Multimedia Broadcast/Multicast Service (Mbms); Protocols and Codecs," Standards Tech. Spec. 26.346, Rel. 11.0.0, 2013.

[10] A. Asadi, Q. Wang, and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," *IEEE Commun. Surveys & Tutorials*, vol. 9, no. 1, Feb. 2015.

[11] 3GPP, "Policy and Charging Control Architecture," Standards Tech. Spec. 23.203, Rel. 11.0.0, 2013.

[12] C. Hahn *et al.*, "A Privacy Threat in 4th Generation Mobile Telephony and Its Countermeasure," *Wireless Algorithms, Systems, and Applications*, Springer, 2014, pp. 624–35.

[13] 3GPP, "Further Advancements for E-Utra Physical Layer Aspects," Standards Tech. Rec. 36.814, Rel. 11.0, 2013.

[14] H. Shen *et al.*, "Socialtube: P2P-Assisted Video Sharing in Online Social Networks," *IEEE Trans. Parallel Distrib. Sys.*, vol. 25, no. 9, 2014.

[15] A. Balachandran *et al.*, "Developing a Predictive Model of Quality of Experience for Internet Video," *Proc. 2013 ACM SIGCOMM*.

Biographies

Abhishek Roy (abhishek.roy@samsung.com) is currently working in the Advanced Technology Group, Networks Division, Samsung Electronics South Korea. He received his Ph.D. in 2010 from Sungkyunkwan University (SKKU), South Korea, and his M.S. in 2002 from the University of Texas at Arlington. His research interests include mobility and resource management aspects of 4G/5G wireless systems. He has served as a Guest Editor and Technical Program Committee member of many international journals and conferences. He has co-authored one book and published 20 international journal papers.

Pradipta De [SM] (pradipta.de@sunykorea.ac.kr) received his Ph.D. from the State University of New York (SUNY), Stonybrook, in 2007. He was a research staff member at IBM Research India until 2012. He is currently an assistant professor at the State University of New York, Korea campus (SUNY-Korea), and jointly appointed as a research assistant professor at SUNY Stonybrook. He directs the Mobile Systems and Solutions (MoSyS) lab at SUNY Korea. His research interests are in mobile computing systems, with a focus on mobile cloud computing.

Navrati Saxena [M] (navrati@skku.edu) is an associate professor in the Electrical Engineering Department, SKKU. She was an assistant professor at Amity University India and a visiting researcher in the University of Texas at Arlington,. She completed her Ph.D. from the Department of Information and Telecommunication, University of Trento, Italy. Her research interests involve 4G/5G wireless and smart environments. She directs the Mobile Ubiquitous System Information Center (MUSIC) of SKKU and serves as a Guest Editor and Technical Program Committee member of international journals and conferences. She has co-authored one book and published 20 international journal papers.

The MNO leverages the proximity information of users to deliver live streams in an effective manner by routing through its nodes, which bypass the content provider network. We have also highlighted how features, like D2D, can play an important role in video streaming in next generation cellular networks.

NCCU Trace: Social-Network-Aware Mobility Trace

Tzu-Chieh Tsai and Ho-Hsiang Chan

ABSTRACT

Delay-tolerant networking (DTN) is a network architecture characterized by the lack of continuous connectivity. Messages are delivered by moving nodes in a store-and-forward manner. In such a network, the mobility models of nodes play an important role in DTNs, because messages can only be delivered when two or more nodes contact each other. In general, mobility models can be categorized into synthetic and trace models. Synthetic models are based on mathematical models that generate the mobility models. Trace models record people's daily movements in the real world; these models faithfully render the actual situation of people's movements in their lives. However, there is a challenge in studying human mobility, and the mobility models of humans affect the performance of routing protocols in DTNs. In this article, we design an Android application to collect the mobility traces of college students in a campus environment, called NCCU Trace Data. We design a mobility model that traces students' movement, and this model can be imported into the ONE simulator to verify routing protocols. More importantly, it can be used to evaluate the performance of a social-based routing method. Finally, we evaluate our routing method and compare it to other routing methods in each mobility model. The simulation results show that our newly designed mobility model is a step closer to the real environment.

INTRODUCTION

Smart handheld devices are quickly emerging in the market, and almost everyone carries one smart handheld device at any time. Most smart handheld devices have the ability to establish an ad hoc connection and exchange data directly with each other through Bluetooth 4.0 and Wi-Fi Direct. In some applications or scenarios (e.g., location-based services or proximity applications), smart handheld devices can exchange data directly without the help of any infrastructure. However, they may not be able to stay connected once they move apart or are intermittently disconnected due to wireless link instability. In such cases, the concept of delay-tolerant networking (DTN) [3, 4, 5, 7, 12, 13] can be applied to face this problem. DTNs make exchanging data among peers located in different places possible, even when nodes do not have direct Internet connectivity.

Although DTNs do not guarantee continuous end-to-end connection between the sender and the receiver, this network environment is even more like the real world situation. In reality, people move about various places as they start their day, including workplaces, restaurants, schools, and so on. During this process, it is very possible that people would want to disseminate data or receive data. However, the contact time of the people they encounter on the road may often be much shorter. Mobility models of people play an important role in DTNs; knowing the state of the network can facilitate and predict future contacts for selecting an appropriate relay node to which to forward data.

Numerous previous works have discussed mobility models. In general, mobility models are divided into two types: synthetic and trace models. Synthetic models are based on mathematical theory to generate mobility models, such as the random walk mobility model and random waypoint mobility model [6]. This type of model tries to present people's real movement; however, it is still questionable. Previous works [9, 10, 15] discuss how to collect people's movement in real traces. Most trace models record people's movements in the real world via smart handheld devices that are equipped with Global Positioning System (GPS). Collecting the trace data from people's movements is not a simple task.

In this article, we build on the MIT reality trace [9] to design a location-aware and behavior-aware Android app: NCCU Trace Data. The Android app installed on the smart handheld devices of National Chengchi University students collects their daily trace data. Through collecting the trace data via our app, we can obtain students' movements on campus. Our experiment is designed to trace the movement of students on a typical school day. Also, our mobility model is closely linked with the social relations such that it is applicable to a social-based routing method for DTN-like simulations. Based on our NCCU Trace Data, we propose a method for interest-based message dissemination, and compare the

The authors are with National Chengchi University.

0163-6804/15/$25.00 © 2015 IEEE

performance of classical routing methods in different mobility models. The simulation results show that the performance of routing methods in our mobility model is closer to the real world situation.

The remainder of this article is structured as follows. We discuss the related work. We describe our approach. We compare different routing methods in different mobility models, which is followed by our conclusion and future direction.

Related Work

First, we review previous works on mobility models. In general, the mobility models can be divided into two types: synthetic and trace.

Synthetic Mobility Model

The random walk mobility model: Each node randomly chooses speed and a destination to be reached from the present location. While the node is moving toward the destination, the node does not stay at any position until it reaches the destination.

The random waypoint mobility model [6]: This mobility model is similar to the random walk mobility model, with the difference that the random waypoint mobility model increases pause time. While the node is moving toward the destination, the node will stay at any position(s) with a pause time before it reaches the destination.

The random direction mobility model [10]: In this mobility model, the node randomly chooses a direction to move along until it reaches the border. After the node stays on the border for a pause time, it randomly chooses another direction in which to move.

As described above, synthetic models try to present the node's movement in real environment. However, it is not sufficient to show people's movements in their daily lives. People often go to places such as workplace, school, and home, and stay there for a certain period of time.

Trace Model

Infocom 06 [15]: The author seeks volunteers who participated in Infocom 06 to do this experiment. Each of 78 attendees carried an iMote device to record Bluetooth connection established between attendees during a four-day conference.

Cambridge [10]: This experiment was proposed by a computer lab at Cambridge. Fifty-four University of Cambridge students were involved and divided into two groups: freshman and sophomore. Each carried an iMote device to record Bluetooth connection with neighboring devices for 11 days.

MIT Reality Trace [9]: Eagle and Pentland proposed an experiment to collect social trace data at MIT. One hundred users participated in the experiment. Seventy-five of them were either students or faculty from the MIT Media Laboratory, and the remaining 25 were incoming students at the MIT Sloan Business School, which is adjacent to the Media Laboratory. Each user carried a Nokia phone that ran the software which recorded any contacts with neighboring Bluetooth devices, application usage statistics, phone state, and so on.

As described above, trace models are not easy to collect, but closely follow the scenario of human movement. Users participating in the experiment were mostly active in the same building, so there was frequent contact between users. However, in a real-world scenario, contacts between nodes may be scarce. For example, students might go to the library, restaurant, or gym after class. Not all users are in frequent contact with another user. In this article, we design an Android app to collect the daily lives of college students. Students were not restricted within a certain location. Instead, these students were uniformly dispersed in various places on campus.

Proposed Approach

In our mobility model, we consider students' movement on campus. Each student can have a sporadic schedule on a daily basis. It is natural for a student to move to any place on campus, and to stay for different lengths of time at each place. In order to collect the data of these different movements, we designed an Android app named NCCU Trace Data.

Chengchi University students who participated in our experiment installed our designed Android app. This allowed us to obtain their movement on campus via our app and system. The participants were not restricted to any specific department, and were free to move around the campus. Therefore, the data we collect will be the real data of students' movement on campus. A total of 115 participants participated in the experiment. Their GPS data, application usage, Wi-Fi access points, and Bluetooth devices in proximity are recorded while the participants were on campus over a period of two weeks.

Since our Android app records personal information, privacy concerns are essential in the context of our technology. In order to safeguard potential privacy threats resulting from leakage of personal data, we provided consent forms for students. Moreover, their personal information was encrypted to ensure further protection.

NCCU Trace Data referred to previous works [9] and was designed to be location-aware and behavior- aware in order to collect the daily trace data of college students. Considering a smart handheld device's power consumption, our app collects data once every 10 min, and uploads the collected data to our server every day. This app runs in the background, and does not interfere with normal usage of the smart handheld device. The following describes what trace data our app collected.

Position: We collect students' location via their smart handheld devices' GPS. If a student happens to be indoors (classroom, library, gym, etc.), our app would switch to Wi-Fi or third/fourth generation (3G/4G) to facilitate positioning, allowing us to obtain the students' movement data in the campus environment.

Wi-Fi access point proximity: Our app records how many Wi-Fi access points are available nearby and their detailed information, such

In our mobility model, we consider students' movement on campus. Each student can have a sporadic schedule on a daily basis. It is natural for a student to move to any place on campus, and to stay for different lengths of time at each place. In order to collect the data of these different movements, we designed an Android app named NCCU Trace Data.

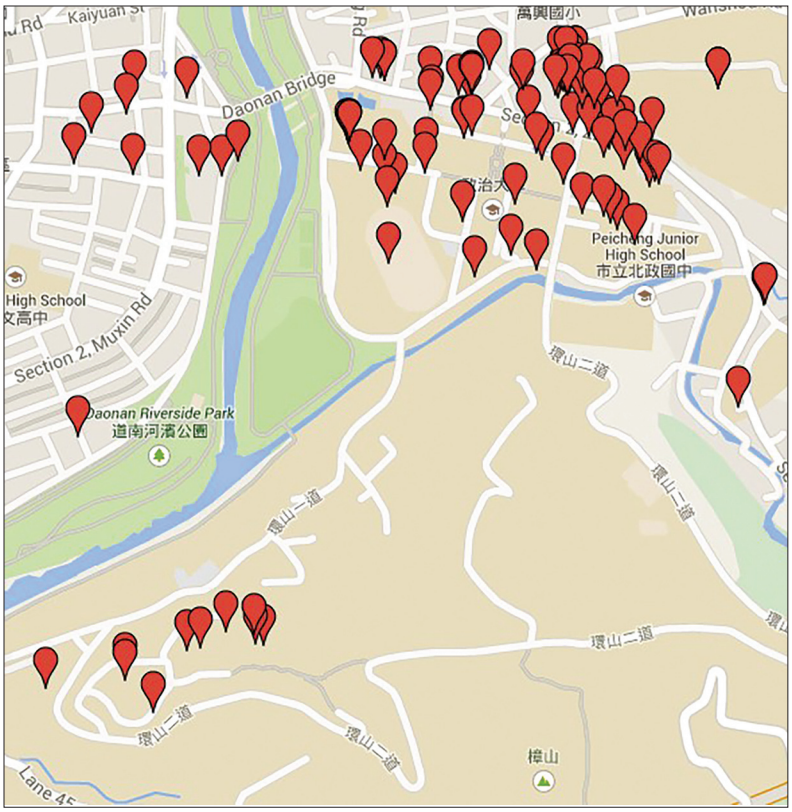


Figure 1. NCCU campus environment.

as medium access control (MAC) address, service set identifier (SSID), and received signal strength indication (RSSI) of each access point.

Bluetooth-based device proximity: The number of available nearby Bluetooth devices and their detailed information, such as MAC address and RSSI, are recorded. With this data, it is possible to discover students' social relations when they are in communication range of Bluetooth devices. Thus, the community relations of students can be obtained.

Students' behavior using smart handheld devices: Our app records how frequently and how long students use apps (Google Map, WhatsApp, Tweet, etc.) in their smart handheld devices. We can find the students' usage behavior on their smart handheld devices through the data our app records.

We filter out movement data that are not on the campus, and also import the collected position data of 115 participants to a map, as shown in Fig. 1.

In our movement model, we use GPS to collect the trace data. GPS requires a strong signal in order to be accurate. Accurate GPS signals may not be received inside most buildings. In other words, the interval between two available GPS data may be more than 10 min (which is our measurement interval). This may mean the device is not moving (stays inside the building) during the period, or it really moves but the GPS data is lost. To this end, we make up the missing data by assuming that the device was moving at the same speed between the two positions during the period.

In addition to our app, we also used questionnaires to collect students' personal information including sex, grade, majors, personal interests, places on campus they would go most frequently, Facebook ID, and the behavior when using Facebook. Our collected questionnaires help us verify that the data collected is very close to a student's real movement on campus.

To further validate our trace data and the personal information collected from questionnaires, we also develop a message dissemination routing method for DTN environments. Traditional routing methods do not consider whether the person is interested in the content of a message. We think it is annoying when one (or a message forwarder) randomly receives messages of no interest. If we want to filter the messages according to their interests, there is no existing realistic data set (with both mobility data and interest data) for us to simulate the behavior. Thus, using our NCCU Trace Data can tackle this kind of problem.

Suppose that one message dissemination application is used in the DTN environment. For example, a student S_A has a message (of art lectures) to disseminate. Those who are interested in this message (of art lectures) are the potential receivers. The goal is to design the message dissemination method to achieve higher delivery ratio or reduce delivery delay. To evaluate the performance, a realistic model for simulation is needed.

We develop an interest-based dissemination method, and utilize our NCCU Trace Data to evaluate the performance and its practice. We consider two scenarios to disseminate the messages: direct contact and indirect contact, described as follows.

Direct contact: When student S_A has a message to disseminate about, for example, art lectures, as the student walks on campus and meets student S_B. They exchange their information of interests. If student S_A thinks student S_B is interested in an art lecture, student S_A delivers the message to student S_B.

Indirect contact: Continued from direct contact, by logging the frequency and interests of others met, student S_B may possibly know who is interested in this art lecture and meets them relatively frequently, although he or she is not interested in this art lecture. Student S_A still sends the message to student S_B in order to let more students know about the art lecture. To this end, we utilize the social relationship information to assume that student S_B is willing to be the forwarder of the message.

Simulation Results

In DTNs, the mobility models greatly affect the performance of routing methods. We evaluated the performance of our interest-based message dissemination method and other traditional routing methods in different mobility models. Here we use the Opportunistic Network Environment (ONE) Simulator [1], which can import our NCCU Trace Data as the mobility model. We also compared the performance of other existing routing methods in our real mobility model and synthetic model. The following performance metrics are used in the simulation:

- Delivery success ratio: The ratio of the number of successfully delivered messages to interested destinations to those delivered to the total number of potentially interested destinations
- Delivery overhead: The average number of relays used for one message successfully delivered to an interested destination
- Delivery delay: The average delay for all messages succesfully delivered to interested destinations

Simulation Methods

In order to achieve a fair comparison of the performance of routing methods, we modified the original versions of the PRoPHET [4] and Spray and Wait [12] routing methods. The original version of them sets only one destination to receive a message. We further modified the message so that it could be delivered to multiple destinations in DTNs. The following summarizes the routing methods with which we compare.

- Epidemic [3]: When a student has a message to disseminate and contacts another student, the message is replicated and forwarded to the other student.
- PRoPHET [4]: When a student has a message to disseminate and contacts another student, if the contacted student encounters any destinations more frequently than the student with a message to send out, the student takes the message and forwards it to other potential students.
- Spray and Wait [12]: This routing method includes two phases: spray and wait. Initially, the student has a message to disseminate that is delivered to multiple destinations, and the student controls the number of messages sent out. In the spray phase, the student forwards half of the copies of the messages when he or she makes contact with another student. This continues until the student has only one message left; then the student switches to the wait phase. In the wait phase, the student delivers the last copy of the message to any destinations he or she encounters.

Simulation Setup

We use the ONE Simulator [1] to show the performance of routing method results. Our NCCU Trace Data can easily be imported to the ONE Simulator, we are able to compare the performance of the routing methods in the different mobility models. In our NCCU Trace Data mobility model, 115 students were simulated for 86,400 s during the students' day on campus. The area is 3764 m × 3420 m, which serves as the main active area of the campus environment. In another mobility model, we chose the random waypoint mobility model for simulation purposes. In both mobility models, the transmission range of each student's smart handheld device is 10 m, and the data transmission rate is 2 Mb/s. The size of each message is 500 kB~1 MB. Furthermore, we consider the interval of creating new messages according to the user's behavior of smart phone usage rather than a random interval. Thus, our messages creation time interval is based on questionnaire results. The simulation setting is shown in Table 1:

Parameter	Value
Simulation times	86400 s
Area	3764 m × 3420 m
Movement model	(1) Random waypoint mobility model (2) NCCU Trace Data
Radio range	10 m
Message size	500 kB~1 MB
Interval of message creation	Student behavior
Data rate	2 Mb/s
Buffer size	500 MB
Time to live	5 hours

Table 1. Simulation setting.

Simulation Results

Simulation results show the performance of various routing methods in different mobility models. We compared our proposed interest-based message dissemination method with Epidemic, PRoPHET, and Spray and Wait routing methods.

In Fig. 2, we can see the delivery success ratio of all the routing methods in the random waypoint mobility model presented a stable curve each day. In the features of this mobility model, Epidemic has the best delivery success ratio over the other three routing methods. This is due to the fact that the Epidemic routing method would forward messages to every student encountered. The method does not consider the cost of sending out different numbers of copies of messages; thus, most of the students possibly have the same message.

The PRoPHET routing method uses previous contact of nodes to calculate who has the higher probability to be in contact with other students. Even if the student has a higher probability of encountering the destination, it is difficult to guarantee that encountering the destinations and then forwarding a message in the random waypoint mobility model can simultaneously occur. In this mobility model, it cannot be justified that the probability of encounter is able to model the chance of meeting the destination nodes.

The delivery success ratio performance of our method lies between PRoPHET and Epidemic.

However, because the delivery success ratio is not close to the real situation, the delivery success ratio of all routing methods cannot show a stable curve. We can see that all routing methods have a higher delivery success ratio in our NCCU Trace Data than that of the random waypoint mobility model. This may be due to the reason that our mobility model is based on the students' class schedule or their extracurricular activities, and students' movement may be connected to their social relations, not only on random movement. Especially on the 4th and 11th days, we can clearly see that the curve obviously

declined in the Spray and Wait and PRoPHET routing methods. The 4th and 11th days fell on a Saturday, and most students do not go to the campus during weekends. This leads to the result that fewer students meet other students on campus, making it difficult to find a student to forward or deliver messages. Furthermore, the Spray and Wait routing method controls the number of copies of messages. In the wait phase, the student may never contact any destination node before the message's time to live limit is exceeded. The PRoPHET routing method is based on the students' movement situation to compute the probability of a given student having a higher chance of encountering more destinations. When fewer students walk on campus, it is difficult to find the appropriate relay for students to forward messages, and the messages hit their time to live limit, similar to the result of the Spray and Wait routing method.

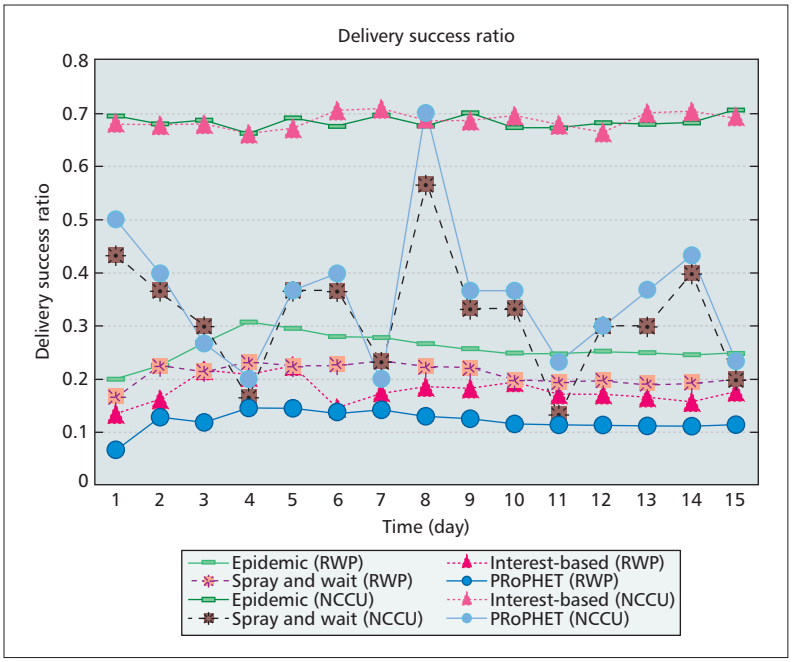


Figure 2. Delivery success ratio.

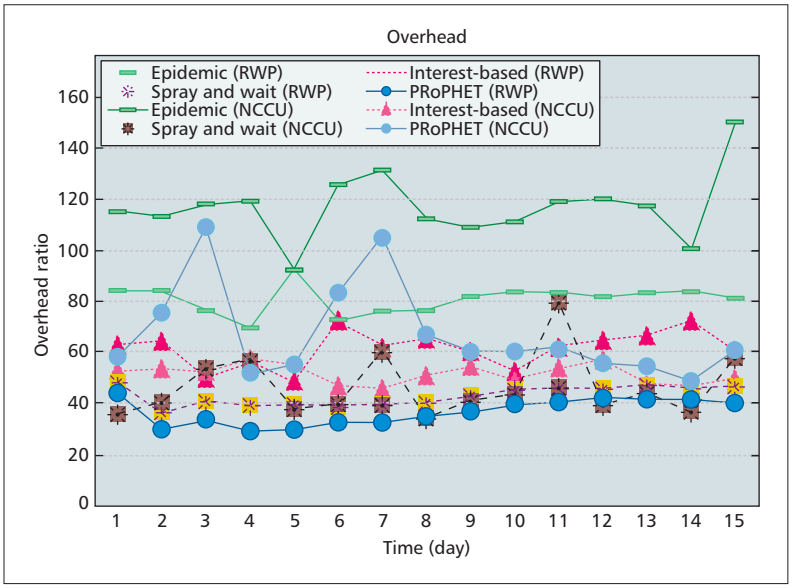


Figure 3. Overhead.

It can be observed from the result that the mobility model does affect the performance of all routing methods.

In Fig. 3, we can see that the overhead of all routing methods is lower in the random waypoint mobility model than that in our NCCU Trace Data. Because the random waypoint mobility model is not movement-focused, most of the routing methods are unable to find the right person to forward the message. On the other hand, the students are frequently in contact with each other according to their social relations in the real environment, while exchanging messages that users are more likely to be interested in. Using our NCCU Trace Data, we can see that students who frequently forward messages have relatively high overhead. It can be noticed that our interest-based message dissemination protocol reduced the message delivery overhead in our NCCU Trace Data, and outperforms the random waypoint mobility model. Since our method is to determine whether a student is interested in the message, it can reduce any unnecessary forwarding of messages.

In Fig. 4, we can see the delivery delay of all routing methods is higher in the random waypoint mobility model than that in our NCCU Trace Data. The main reason is that because the students' movement is toward a randomly chosen direction, it may be difficult to meet any potential destinations. It takes a long period of time for the destinations to receive the message. Our NCCU Trace Data is a real human behavior model. Students tend to make friends who are similar to each other, so they may have the same interests. Friends with similar interests may frequently get along. Besides, message delivery with similar interests increases the success ratio and decreases the delay. The fourth day, which is Saturday, results in long delay time and low delivery ratio. This is because students do not go to school on Saturday. The simulation results show that our NCCU Trace Data is close to real students' movement with social relations.

Conclusions and Future Work

We designed a location- and behavior-aware Android app ,NCCU Trace Data, to collect daily trace data of college students. We describe how we collected the students' movement data and filtered out any data that were not on campus, making a mobility model in the campus environment. Our mobility model has the characteristics of social relations. It can be imported into the ONE simulator to evaluate social-based routing methods. NCCU Trace Data can be downloaded from http://www.cs.nccu.edu.tw-/~d10003/.

At the same time, we propose an interest-based message dissemination method considering users' interest attributes. Finally, we simulated the traditional routing methods and our method in different mobility models to compare their performance. The results show that our mobility model is closer to the real movement situation. It is an ideal trace data to evaluate the performance of social-based routing

methods. Moreover, our proposed interest-based message dissemination method has a better delivery ratio than that of the PRoPHET and Spray and Wait routing methods. Also, our delivery overhead ratio is much lower than that of the Epidemic routing method.

In the near future, we will continue to release further data collected from the experiment, such as Wi-Fi access points in proximity, Bluetooth devices in proximity, and students' behavior using the apps in smart handheld devices. We will also continue to improve the message dissemination routing method. Additionally, we will consider the attributes of buildings in particular, because we believe that students often hold activities in certain buildings according to their attributes. Finally, we will derive a more realistic synthetic mobility model based on our Trace Data.

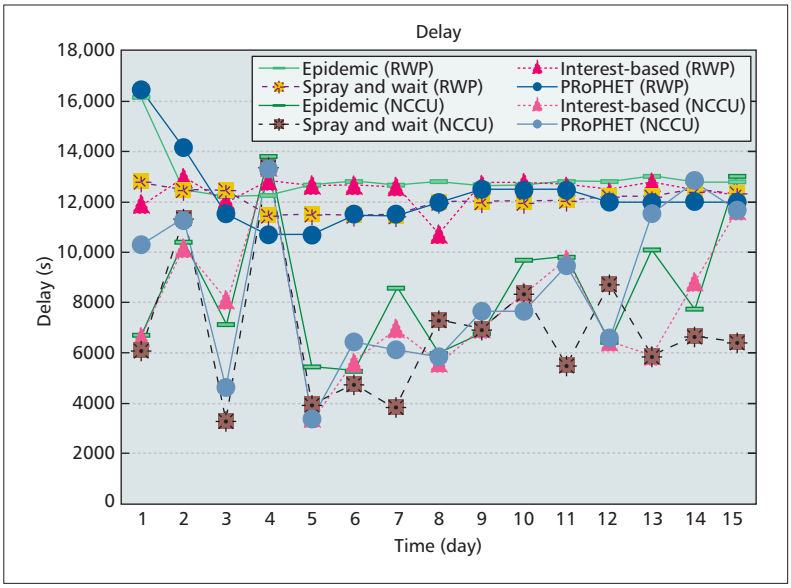


Figure 4. Delivery delay.

References

[1] A. Keränen, J. Ott, and T. Kärkkäinen, "The ONE Simulator for DTN Protocol Evaluation," *Proc. SimuTools*, Mar. 2009.

[2] A. Mtibaa *et al.*, "PeopleRank: Combining Social and Contact Information for Opportunistic Forwarding," *INFOCOM*, 2010.

[3] A. Vahdat and D. Becker, "Epidemic Routing for Partially Connected Ad Hoc Networks," Tech. Rep. CS-2000-06, Duke Univ., July 2000.

[4] A. Lindgren, A. Doria, and O. Schelén, "Probabilistic Routing in Intermittently Connected Networks," LUT, Sweden, *Proc. SIGMOBILE*, vol. 7-3, July 2003.

[5] E. Bulut and B. K. Szymanski, "Friendship Based Routing in Delay Tolerant Mobile Social Networks," *IEEE GLOBECOM 2010*.

[6] C. Bettstetter, H. Hartenstein and X. Perez-Costa, "Stochastic Properties of the Random-Waypoint Mobility Model," *Wireless Networks*, vol. 10, no. 5, 2004, pp. 555–67.

[7] E. Daly and M. Haahr, "Social Network Analysis for Routing in Disconnected Delay-Tolerant MANETs," *Proc. ACM MobiHoc*, 2007, pp. 32–40.

[8] A. Mei *et al.*, "Social-Aware Stateless Forwarding in Pocket Switched Networks," *Proc. IEEE INFOCOM*, Mini-Conference, 2011.

[9] N. Eagle and A. Pentland, "Reality Mining: Sensing Complex Social Systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, May 2006, pp. 255–68.

[10] P. Hui *et al.*, "Pocket Switched Networks and the Consequences of Human Mobility in Conference Environments," *Proc. 2005 ACM SIGCOMM Wksp. Delay-Tolerant Networking*, 2005.

[11] S. C. Nelson, M. Bakht, and R. Kravets, "Contact-Based Routing in DTNs," Univ. Illinois at Urbana-Champaign, *Proc. INFOCOM*, Apr. 2009.

[12] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and Wait: An Efficient Routing Scheme for Intermittently Connected Mobile Networks," *Proc. WDTN '05*, ACM Press, 2005, pp. 252–59.

[13] Y. Zhu *et al.*, "A Survey of Social-Based Routing in Delay Tolerant Networks: Positive and Negative Social Effects," *IEEE Commun. Surveys & Tutorial*, no. 9, Apr. 2012.

[14] T. Camp, J. Boleng, and V. Davies, "A Survey of Mobility Models for Ad Hoc Network Research," *Wireless Commun. and Mobile Computing*, vol. 2, no. 5, 2002, pp. 483–502.

[15] Infocom'06, http://crawdad.org/.

Biographies

Tzu-Chieh Tsai (ttsai@cs.nccu.edu.tw) received his B.S. and M.S. degrees, both in electrical engineering, from National Taiwan University in 1988 and the University of Southern California in 1991, respectively. He was chair of the Department of Computer Science during August 2005~July 2008, and director of the Master's program in digital content and technologies during September 2013~July 2015, at National Cheng-Chi University, Taipei, Taiwan. Currently, he is an associate professor in the Computer Science Department. His recent research work includes ad hoc networks, delay-tolerant networks, mobile commerce, cloud computing, and wearable computing.

Ho-Hsiang Chan (100753503@nccu.edu.tw) received his Master's degree from Shih Hsin University, Taiwan, in 2010. He is currently working toward his Ph.D. degree in the Department of Computer Science at National Chengchi University. His research interests are delay-tolerant networks and wireless communication technology. His recent focus is message dissemination based on people's interests.

Socially Aware Mobile Peer-to-Peer Communications for Community Multimedia Streaming Services

Changqiao Xu, Shijie Jia, Lujie Zhong, and Gabriel-Miro Muntean

ABSTRACT

Mobile multimedia streaming services have become the main reason behind the latest increase in mobile data traffic worldwide. The enormous user base, frequency of data exchange, and demand for high-quality multimedia content open up new challenges to offer good QoS levels in mobile multimedia streaming systems. This article presents an innovative SMMC that integrates performance factors with aspects such as demand, socialization, mobility, and delivery. Simulation results demonstrate that SMMC achieves high content sharing efficiency and increased QoS levels.

INTRODUCTION

The latest advancements in wireless communication technologies and developments in terms of handheld device capabilities enable mobile users to have ubiquitous access to diverse multimedia streaming services. Figure 1 illustrates a mobile multimedia streaming application scenario: users make use of mobile devices equipped with multiple wireless communication interfaces (e.g., WiFi/WiMAX, WAVE, and fourth/fifth generation, 4G/5G) in order to access multimedia streaming data in a heterogeneous wireless network environment. The above-mentioned technological advancements have resulted in significant benefits in terms of both user service diversity and quality of experience (QoE) levels, which in turn have attracted increased numbers of users with additional demands for services at even higher quality. This process has resulted in large amounts of traffic being delivered over the existing limited capacity networks. Peer-to-peer (P2P) networking, and in particular mobile P2P (MP2P) [1], represent a scalable solution, especially useful in scenarios involving large-scale multimedia streaming over the (mobile) Internet (e.g., the SINA NBA P2P-based live video system supports over 5 million online users in China).

At an increasing scale, the management and sharing of multimedia resources are decisive factors for scalability, and quality of service (QoS) and QoE levels. In traditional P2P solutions, structured architectures such as tree and distributed hash table (DHT)-based support fast resource search, but the high maintenance cost due to peer churn severely limits system scalability. Unstructured architectures have high scalability mostly due to the peer-centric overlay maintenance scheme, but flooding-based search introduces high startup delay and wastes network bandwidth. The balance between scalability and resource sharing performance has become the major issue of any P2P architecture.

Virtual community technologies enable nodes with common interests to form communities that have autonomous and flexible structures [2]. Multimedia virtual communities can be defined to describe the boundary of search and management of multimedia resources with similar user characteristics in order to balance system scalability and resource sharing capability. In this context, the major issue is how to construct these multimedia virtual communities, as the relationships between community members who rely on common user interests and variation of user interests lead to fragile inter-member links, and negatively influence the maintenance cost and sharing capacity of such community structures.

The emerging social networks make use of the interaction and socialization of users to describe the relationship between them. The social relationship tightens up the relationship between community members and addresses the problem of member selfishness. For instance, family members gladly share stored content with each other. Socially aware communities have a resilient structure and optimal distribution of resources, enabling high scalability and QoS levels. For the deployment of community-based multimedia systems in wireless mobile networks, there is a need for a socially aware mobile multimedia community-based approach (SMMC), which also addresses the similarity of mobility between mobile nodes. This is because the mobility of members causes dynamic network topology changes and influences the content

Changqiao Xu is with Beijing University of Posts and Telecommunications.

Shijie Jia is with Luoyang Normal University.

Lujie Zhong is with Capital Normal University.

Gabriel-Miro Muntean is with Dublin City University.

0163-6804/15/$25.00 © 2015 IEEE

delivery efficiency. Mobility awareness helps the members discover the desired resource holders that have similar mobility, reducing the jitter of communication distance and supporting communication over fewer hops. Delivery awareness regulates transmission strategies in terms of QoS and variation of communication quality in the transmission path in order to maximize the performance of content delivery and improve the network bandwidth utilization efficiency.

It is expected that SMMC enhanced with the results of investigating performance-related factors (PRFs) including demand, socialization, mobility, and delivery will be highly beneficial in terms of performance, but will also have low maintenance costs. This article presents an innovative design of SMMC that integrates PRFs and analyzes how PRFs influence serving capability (e.g., resource sharing and delivery) and scalability. The article also investigates approaches for PRF estimation and presents an innovative SMMC construction strategy. Challenges of designing SMMC to improve serving capability and system scalability are discussed, and characteristics of PRFs are investigated in the context of their exploitation for the best SMMC design.

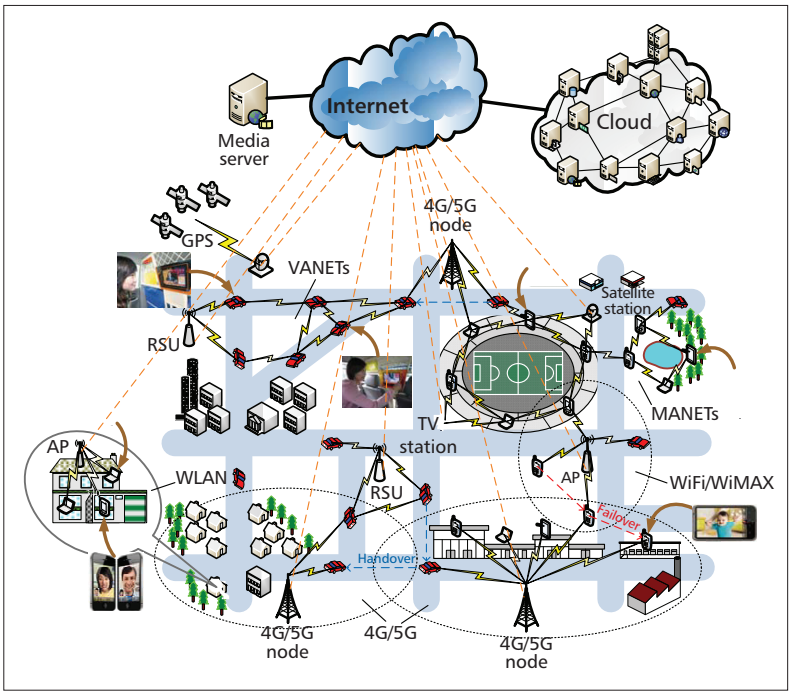


Figure 1. Mobile multimedia streaming services in wireless networks.

SMMC Challenges

SMMCs make use of demand similarity and user socialization to define the relationship between nodes. Demand refers to user requests for specific content, and the desire for certain content drives users to regularly request various content-specific resources. The user demand for multimedia content is embodied in the request behavior. User socialization also indicates a stable relationship, is a natural attribute of humans, and marks the level of relationship between users. The investigation of demand and socialization enables SMMCs to tighten up the relationship between members and tolerate long update periods of member states, which further ensure the scalability and resilience of the community structure.

P2P networks are divided into multiple SMMCs, which include intra-community nodes, which have high similarity in terms of demand and socialization, and inter-community nodes, which are different in terms of the same aspects. Intra-community search for resources has higher probability of success than that of cross-community, so the distinct search coverage in SMMCs can reduce the wait delay.

The closeness of a community member relationship determines the stability and scalability of the SMMC structure. A major issue is how to estimate the relationship in terms of demand and socialization, which are the main components when describing member relationship. A wide range of interests causes users to fetch multiple types of multimedia content. For instance, a sports fan not only enjoys NBA and the soccer World Cup, but may also like to watch the "Transformers" movie. In terms of common interest, classic approaches model the interests in terms of vectors and graphs, and calculate the interest similarity between users. However, threshold-based methods are difficult to adapt to changes in interests. Another problem is how to accurately capture common interests of users and prevent frequent movement between communities due to transient interest variations.

The social relationship has been extracted from user contact online in social networks and offline in real society. In the online social network the similarity of user interests and values is used to build virtual user communities. The relationship in the social network is also hierarchical in terms of the contact frequency between users. The relationship in real society relies on factors such as family and contact frequency to associate levels of strength. Because the users in real society have distinct community characteristics, their relationships are also mapped to virtual communities. The existence of multiple socialization-related factors increases the complexity of the estimation methods for the similarity and intensity of user relationships. Moreover, strong relationships between intra-community members limit resource sharing with inter-community nodes and introduce problems related to community selfishness. For instance, users can limit access to uploaded videos and photographs by the content access authority setting.

Another main goal of multimedia streaming systems is efficient content transfer from content providers to requesters while supporting large-scale applications. Changing network topologies caused by user mobility is one of the main influencing factors for the capacity of content transport. User mobility is characterized by socialization and randomness. The socialization of mobility is embodied in space and time. For instance, a student moves from home to school in the 7:00–8:00 period every day during the school term. The movement range of the student is within the school grounds during the 8:00–17:00 time interval, and the student may return to and stay at home after 17:00. The socialization of mobility ensures a stable mobility trace and range for users. It is easier to recog-

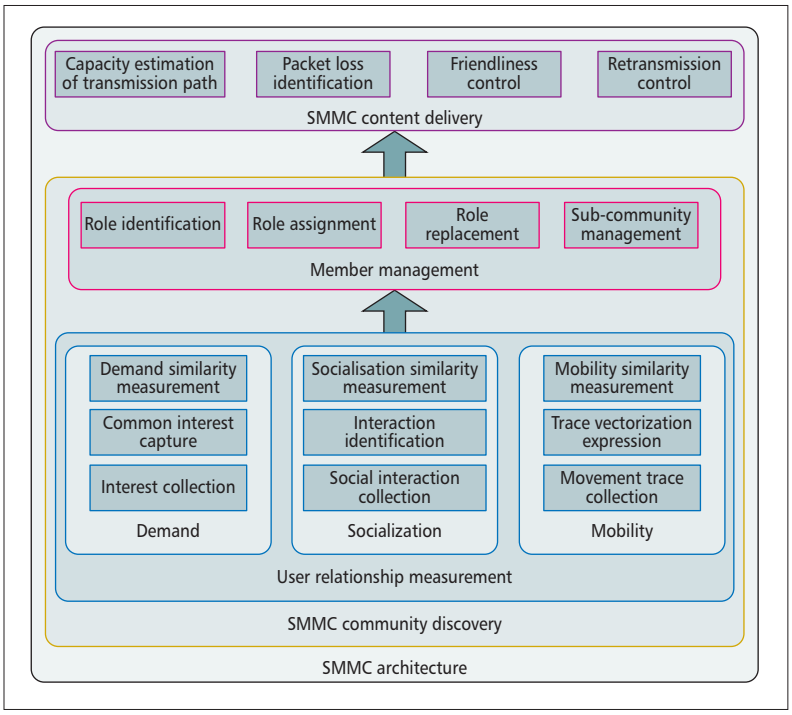


Figure 2. SMMC architecture.

nize the movement behavior of users to pair two communication parties in close geographical locations. Adjacent location provides low transmission delay and prevents variation in the content delivery performance. However, randomness leads to an increase in the difficulty of recognition and prediction of user mobility. For instance, when a student goes to hospital instead of school after 8:00 as s/he is ill or when drivers change dynamically the route according to real-time road conditions affected by traffic jams.

However, it is difficult to maintain content delivery so that it is always performed within close geographical locations. The delivery performance is easily influenced by communication quality variation due to the dynamic transmission path fluctuations. There are multiple factors that reflect communication quality in transmission paths (including packet loss rate and delay), and there are multiple reasons behind the variation of these factors. The accuracy of recognition of the real factors causing the change of communication quality is a precondition of making corresponding delivery adjustments. The heterogeneous network environment introduces high complexity in monitoring transmission paths and adapting delivery. In particular, in order to improve multimedia streaming data transmission performance, the use of concurrent multipath transmission (CMT) further increases the difficulty of monitoring and adjusting the delivery due to the involvement of multiple interfaces of diverse technologies such as 4G, WiFi, and WAVE.

SMMC Architecture Overview

Figure 2 illustrates the SMMC architectural design, which includes community discovery and content delivery. In SMMC community discovery, we perform user relationship measurement in terms of the similarity in:

- *User demand* by interest collection, common interest capture, and demand similarity measurement
- *User socialization* from online and offline social relationships by collection and identification of social interaction and socialization similarity measurement
- *User mobility* in two scenarios by movement trace collection, trace vectorization expression, and mobility similarity measurement

Based on those, we design a multi-role *member management* mechanism, which includes role identification, assignment and replacement, and sub-community management. In SMMC *content delivery*, we propose a network context-aware CMT solution in wireless heterogeneous networks, which includes capacity estimation of transmission path, packet loss identification, friendliness, and retransmission control.

SMMC Community Discovery

User Demand

In order to accommodate scalable sharing of content, having similar demands is a fundamental factor for community discovery. SocialTube makes use of the number of watched videos in common between source node and users to estimate interest similarity levels [3]. However, the similarity estimation method based on the number of watched videos captures with difficulty real common interests, and results in low efficiency of content sharing.

This work focuses on an investigation of multiple factors affecting user playback behaviors including the *number of watched videos*, *video switchover*, and *playback time* for capturing common interests between users. The number of watched videos describes the user interest coverage, but it is difficult to explicitly represent the interest coverage boundary. Video classification makes use of content similarity to build a mapping between video and category, and enables, for a limited category set, denoting interest coverage with explicit boundary and obtaining distance between two users in terms of the watched content (content distance). The video categories of each user are considered as a vector where the vector value is calculated by popularity and number of videos. The content distance between users is expressed as the cosine value of the vectorial angle.

Video switchover denotes a change in user interests (from World Cup to music TV) and marks contact between videos (e.g., TV series) from the aspect of user behaviors. The classic modeling method uses graphs to represent the contact between videos by analyzing playback traces of all users. We use the average weight value of all edges passed by each user to denote the content compactedness of the switchover behavior of each user. The playback time directly reflects the users' interest level in each video and is used to estimate playback state stability. The normalization of playback time involves calculation of the ratio between playback time and video length.

Based on the analysis of three factors related to user demand, we design a fuzzy ant-inspired clustering algorithm to find the user group with similar playback behavior and interests [4]. Each user is mapped into a point in a plane where the

content compactedness and playback time ratio are selected as the coordinate axis of the plane. We make use of fuzzy ant clustering to finalize the initial clustering and perform further refinement by iterative removing of impurities. The refined results without use of thresholds have major advantages including global optimum, high resilience, and explicit boundary between sets.

User Socialization

As already stated, user socialization is a natural attribute of users. Some estimation methods of social relationships make use of common interests to denote social relationship levels (e.g., SPOON [5]). However, these methods rely on access content similarity to construct communities. This is not enough to ensure stable user relationships and promote content sharing.

In online social networks, there are some important factors related to interaction behaviors between users that can reflect social relationship, such as *push*, *attention*, and *forwarding* of multimedia content. Users push some content of interest to one or multiple users and consider that the pushed object may be of interest due to the close relationship between pushers and receivers. We consider the push success rate for a single user and average success rate as push metrics. They represent the influence level of a pusher to a receiver, and a group whose members keep contact with the pusher, respectively. User attention is not only an indication of interest, but also denotes that the users actively accept content shared by other users. We consider the metric of user attention from two aspects:

- As attention is similar to playback of content in terms of interests, the attention related to the content is considered as a parameter of interest estimation.
- Because the users focus on the content shared by other users, there is contact between users, so the attention frequency is used to estimate closeness of contact between users.

Content forwarding has similar consideration as attention.

User relationship in the (offline) real society is a mutual trust foundation to promote resource sharing and cooperation between users. For instance, a user makes use of the storage in mobile devices of his/her parents to cooperatively download video content using a cooperative download algorithm as described in our previous work [6]. However, when users in multimedia systems are mapped into overlay networks, they carry some information only (e.g., interest or played video) into their resource space due to user privacy protection. In this latter context, it is difficult to identify the real relationship between users. We consider that the real relationship metric is translated into an estimation of cooperation capacity between users. For instance, the frequency of resource transmission and cooperative download is used to estimate closeness of contact between users.

Based on the above discussion of metrics for the factors related to users' social relationships, including *push*, *attention*, *content forwarding*, and *cooperation*, we normalize these parameters and use the Euclidean distance in a multi-dimensional space as a closeness metric of social relationships between users. Minimizing this metric enables a user relationship to be more stable and speeds up multimedia content sharing in SMMCs.

Demand similarity	Sociality similarity	Mobility similarity	Rank	Abbreviation
No	No	Yes	Level 1	M1
No	Yes	No	Level 2	M2
Yes	No	No	Level 3	M3
No	Yes	Yes	Level 4	M4
Yes	No	Yes	Level 5	M5
Yes	Yes	No	Level 6	M6
Yes	Yes	Yes	Level 7	M7

Table 1. Member ranking.

User Mobility

User mobility leads to increased complexity of community construction and content delivery in SMMCs. Because the socialization of user mobility enables stable trace and area of movement, the content sharing between users who frequently encounter each other in the same or close area has a distinct advantage in terms of delivery efficiency [7]. We consider the stay time in an area as an estimation parameter of mobility. By observing the variation in stay time in different areas according to time series, we calculate the probability of staying in each area for a certain period of time.

Not all users stay in an area for a long time. Mobile users could benefit from some solutions that make use of user similarity in the mobility pattern to improve content sharing efficiency [8]. We proposed an estimation strategy of similarity of mobility in [4] in terms of a trace composed of a passed area (e.g., passed access point). By using a Markov process to model the movement behavior and predict the moving trace of mobile nodes, the similarity of mobility is based on the matching of predicted results and historical traces to obtain accurate estimation results. When the traces cannot be expressed by passed area (e.g., the communication between mobile users in mobile ad hoc networks), we consider making use of the encountered nodes in the process of movement to define movement traces of mobile users. The mobile nodes record encountered one-hop neighbor nodes with stable movement state during a period of time where the definition of stable movement state of nodes is proposed in our previous work [6]. Because the stable movement state in a time span precisely reflects the relative location of mobile nodes, there are no big errors in the process of location expression relative to the modeling methods with the help of a fixed access point.

Community Construction

Traditional community identification approaches cluster nodes in terms of simple metrics such as similarity level and use thresholds [9]. However, the similarity level obtained by matching com-

The sub-communities in the lower layer have one communicator to maintain contact with the corresponding sub-community in the intermediate layer. When the members search content in community space or move from the current sub-community to other sub-communities, the communicator forwards the request to the connected upper sub-community.

mon characteristics leads to the loss of various user information. For instance, after demand, socialization, and mobility of users are mapped into a multi-dimensional space, if the Euclidean distance is used to calculate the similarity level between users, the impact level on user similarity from the three factors is reduced. Instead, we consider formulating a multi-role mechanism to build the structure of SMMCs. Table 1 shows the rank of community members in terms of similarity in demand, socialization, and mobility. For instance, users with similar demands, socialization, and mobility are defined as members with level 7 (M7); if a user has similar demands and socialization with any M7, s/he is an M6; users who have similar demands and mobility with any one M7 become M5. By parity of reasoning, the rank of other members relies on the similarity with M7 in demand, socialization, and mobility. Although M1, M2, and M4 differ in demand with M3, M5, M6, and M7, the former can be considered as partners for cooperative fetching of resources for the latter. M3, M5, M6, and M7 rely on demand similarity to promote resource sharing.

We design a hierarchical multi-sub-community structure to group and manage members. As Fig. 3 shows, M5, M6, and M7 form three sub-communities, SC5, SC6, and SC7, in the community space, respectively. SC7 is located at the top layer; SC5 and SC6 are in the intermediate layer and have links with SC7. In the lower layer, M1 and M3 form two sub-communities, SC1 and SC2, which keep contact with SC5; SC3 and SC4 are composed of M2 and M4, and maintain the logical links with SC6, respectively. Not only can the hierarchical structure of SMMCs distribute the maintenance overhead of member state to multiple sub-communities, but it can also improve sharing efficiency of content. When the values of factors change, the members only need to join the new sub-community in terms of the corresponding member level. For instance, when a member meets the demand similarity and changes its membership level from 1 to 5, s/he moves from the current sub-community to SC5. If the level of a member changes from 2 to 4, s/he remains in the current sub-community. In particular, because there may be the large number of M1, M2, M3 and M4, the members in the lower layers will be divided into more multiple heterogeneous sub-communities like SC1, SC2, SC3 and SC4, in order to keep appropriate sub-community scale.

The sub-communities in the lower layer have one communicator to maintain contact with the

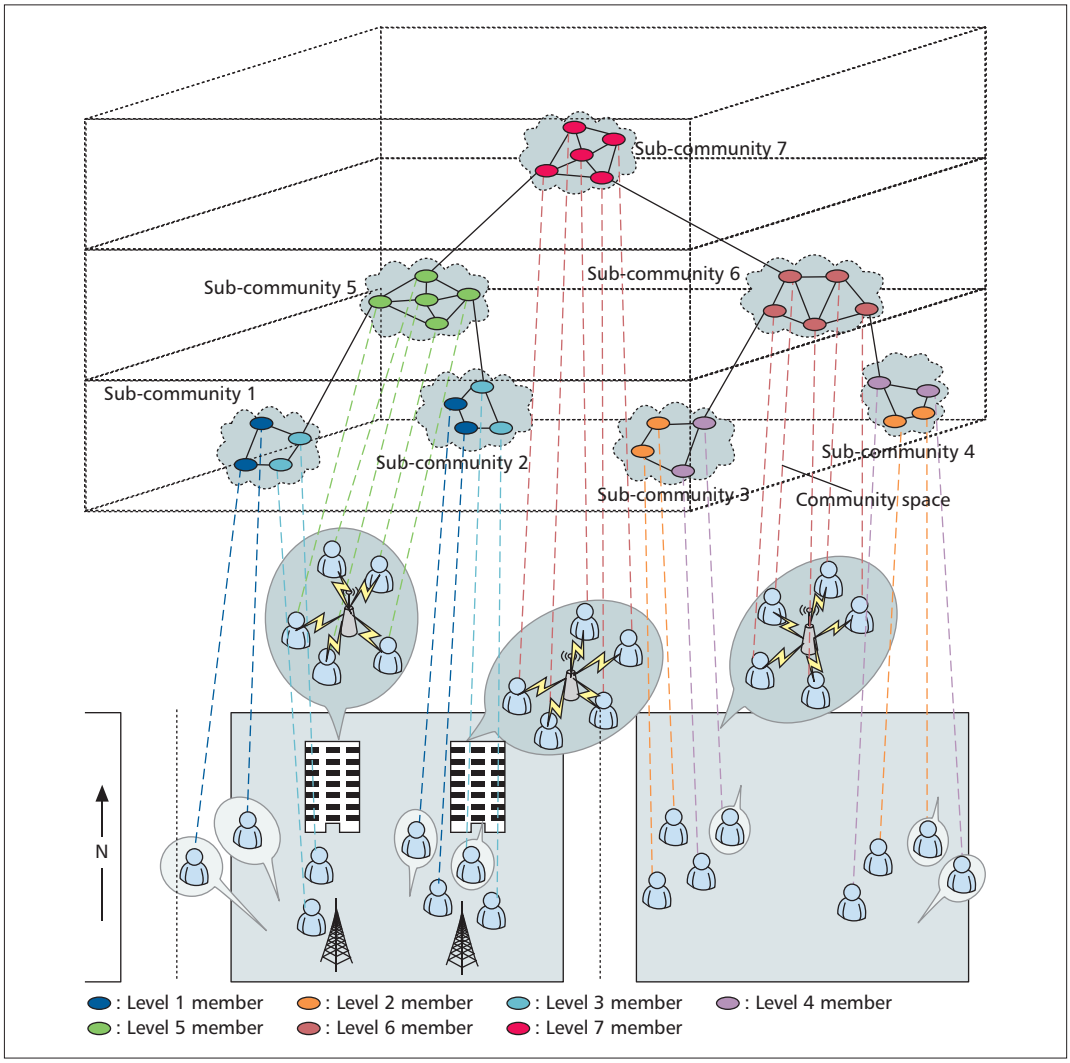


Figure 3. SMMC hierarchical structure.

corresponding sub-community in the intermediate layer. When the members search content in community space or move from the current sub-community to other sub-communities, the communicator forwards the request to the connected upper sub-community. SC5 and SC6 have multiple communicators to maintain contact with the lower-layer sub-communities and SC7, respectively. In particular, when the members build a new sub-community in the lower layer, a member in SC5 or SC6 acts in the communicator role to maintain the new link. The community composed of all sub-communities (e.g., from SC1 to SC7 in Fig. 3) also needs to keep in contact with other communities. Therefore, SC7 includes multiple communicators to maintain contacts with the top-layer sub-communities in other communities. Moreover, when a member can obtain a higher level in other communities or is dissimilar to M7 in demand, sociality, and mobility, s/he leaves the current community. The hierarchical structure maintains the stability of the whole community by making use of local adjustment, and obtains advantages in terms of scalability and sharing efficiency.

SMMC Content Delivery

CMT aggregates the bandwidth of multiple network connections to improve data throughput, and provides backup via multiple transmission paths, also achieving fast retransmission and path switching, and meeting the demand for high-quality multimedia content delivery. However, the dynamic mobile network topologies increase the complexity of path management and data distribution. Moreover, the out-of-sequence data caused by heterogeneous paths results in buffer congestion, which severely influences content delivery efficiency. The current solutions, such as CMT-DA [10], do not accurately estimate the path quality, and their congestion control and flow rate allocation passively respond to network environment changes. Therefore, the multimedia content cannot be carried in real time from provider to requester to ensure smooth playback for the users.

On the other hand, the CMT-based high performance of data transmission by using multiple paths bandwidths is a double-edged sword. It employs the CMT aggressive delivery of content in given network resources, which brings unfairness for other single path transmission protocols such as TCP.

We propose a network context-aware CMT solution for real-time video data delivery in wireless heterogeneous networks to deal with the negative effects of loss and out-of-sequence data, and balance friendliness and efficiency, improving previous work [11]:

- The path capacity is calculated by effective signal-to-noise ratio (ESNR) at the data link layer, available bandwidth, CWND, and a defined friendliness window (FWND), which improves the estimation accuracy of path capacity.
- We identify the reason behind packet loss and classify it to wireless channel errors, network congestion, and path failures, and then take different actions: keep no adjusting CWND, reduce the CWND, and mark the path as "non-active," respectively.
- In order to achieve trade-off between friendliness and efficiency, FWND maintenance relies on a reward method that improves transmission throughput, and a penalize method that reduces the aggressiveness.
- The largest CWND path is selected to retransmit the lost data. Moreover, the receiver adds the current playback point to the ACK packets. The sender compares the latest playback point from the receiver with the time corresponding to lost data and avoids retransmitting the outdated data.

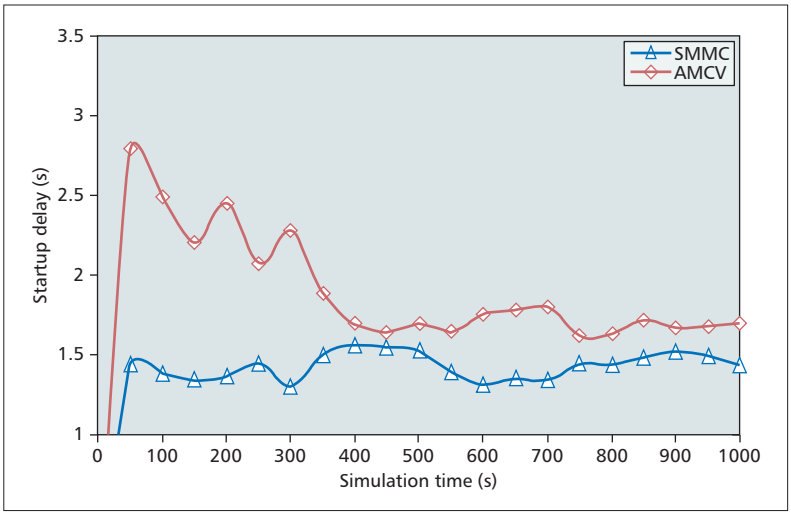


Figure 4. Startup delay during the simulation time.

Performance Evaluation

We compare the performance of SMMC with our ant-inspired mini-community-based solution for video-on-demand services (AMCV) [12] in terms of startup delay and video quality during a stimulation time period of 1000 s. In AMCV, the mobile users who request the same video are grouped into a community. The SMMC transmission protocol is set to SCTP. The two solutions are deployed in a wireless network modeled using NS-2 with the settings described in [6]. Two hundred selected mobile nodes join the system following a Poisson distribution. They request and play videos clips selected from 60 video files of 100 s each, according to 200 viewing logs. Demand similarity in SMMC and video access probability in AMCV are calculated according to historical playback logs. The viewing and historical logs are generated according to [13]. The simulation results include average values of multiple tests.

Figure 4 compares the startup delay, which is defined as the time span from sending a request to receiving streaming data. The blue curve corresponding to SMMC is lower than that of AMCV. Figure 5 compares average peak signal-to-noise ratio (PSNR) of streaming, where PSNR is defined in [4] and denotes the watched quality of the user. Although two curves have the fall trend, SMMC outperforms AMCV. In SMMC, the distinct community boundary and similar demand

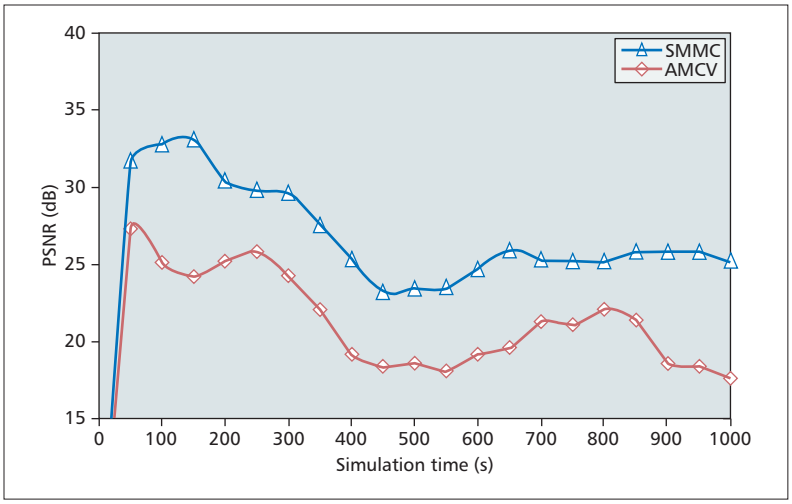


Figure 5. PSNR during the simulation time.

can fast search the resource providers and reduce the number of forwarding request messages. Moreover, because nodes with similar mobility in SMMC employ the network context-aware CMT method to concurrently send the video data, SMMC adapts to the dynamic network environment, and has low transmission delay and high data throughput with little jitter. AMCV does not consider the similarity of demand and socialization, and the change of network topologies brings the negative effects for resource sharing and delivery performance. Therefore, SMMC performs better than those of AMCV.

Conclusion

This article proposes an innovative design for a socially aware mobile multimedia community (SMMC) and discusses challenges in mobile peer-to-peer communications for community multimedia streaming services. Testing results show how, by combining strategies related to demand, socialization, mobility, and delivery, SMMC provides a promising solution for future efficient mobile multimedia streaming systems. Future works will extend SMMC to include user personalization and energy awareness [14].

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61372112, 61501216, and 61522103, and the Beijing Natural Science Foundation (4142037).

References

[1] C. Xu *et al.*, "QoE-Driven User-Centric VoD Services in Urban Multi-Homed P2P-Based Vehicular Networks," *IEEE Trans. Vehic. Tech.*, vol. 62, no. 5, 2013, pp. 2273–89.
[2] H. Bandara and A. Jayasumana, "Community-Based Caching for Enhanced Lookup Performance in P2P Systems," *IEEE Trans. Parallel Distrib. Sys.*, vol. 24, no. 9, 2013, pp. 1698–1710.
[3] H. Shen *et al.*, "SocialTube: P2P-Assisted Video Sharing in Online Social Networks," *IEEE Trans. Parallel Distrib. Sys.*, vol. 25, no. 9, 2014, pp. 2428–40.
[4] C. Xu *et al.*, "Performance-Aware Mobile Community-Based VoD Streaming over Vehicular Ad Hoc Networks," *IEEE Trans. Vehic. Tech.*, vol. 64, no. 3, 2015, pp. 1201–17.
[5] K. Chen *et al.*, "Leveraging Social Networks for P2P Content-Based File Sharing in Disconnected MANETs," *IEEE Trans. Mobile Comput.*, vol. 13, no. 2, 2014, pp. 235–49.
[6] S. Jia *et al.*, "A Novel Cooperative Content Fetching-Based Strategy to Increase the Quality of Video Delivery to Mobile Users in Wireless Networks," *IEEE Trans. Broadcast.*, vol. 60, no. 2, 2014, pp. 370–84.
[7] L. Tu and C.-M. Huang, "Collaborative Content Fetching Using MAC Layer Multicast in Wireless Mobile Networks," *IEEE Trans. Broadcast.*, vol. 57, no. 3, 2011, pp. 695–706.
[8] H.-L. Fu *et al.*, "Group Mobility Management for Large-Scale Machine-to-Machine Mobile Networking," *IEEE Trans. Vehic. Tech.*, vol. 63, no. 3, 2014, pp. 1296–1305.
[9] S. Fortunato, "Community Detection in Graphs," *Phys. Rep.*, vol. 486, nos. 3–5, 2010, pp. 75–174.
[10] J. Wu *et al.*, "Distortion-Aware Concurrent Multipath Transfer for Mobile Video Streaming in Heterogeneous Wireless Networks," *IEEE Trans. Mobile Comp.*, vol. 14, no. 4, 2015, pp. 688–701.
[11] C. Xu *et al.*, "Cross-Layer Fairness-Driven Concurrent Multipath Video Delivery over Heterogenous Wireless Networks," *IEEE Trans. Circuits Sys. Video Tech.*, vol. 25, no. 7, 2015, pp. 1175–89.
[12] C. Xu *et al.*, "Ant-Inspired Mini-Community-Based Solution for Video-on-Demand Services in Wireless Mobile Networks," *IEEE Trans. Broadcast.*, vol. 60, no. 2, 2014, pp. 322–35.
[13] I. Ullah *et al.*, "A Survey and Synthesis of User Behavior Measurements in P2P Streaming Systems," *IEEE Commun. Surveys & Tutorials*, vol. 14, no. 3, 2012, pp. 734–49.
[14] A. Moldovan *et al.*, "Energy-Aware Mobile Learning: Opportunities and Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 1, 2013, pp. 234–65.

Biographies

Changqiao Xu received his Ph.D. degree from the Institute of Software, Chinese Academy of Sciences (ISCAS) in January 2009. He was an assistant research fellow at ISCAS from 2002 to 2007, where he was a research and development project manager in the area of communication networks. During 2007–2009, he worked as a researcher with the Software Research Institute at Athlone Institute of Technology, Ireland. He joined Beijing University of Posts and Telecommunications (BUPT), China, in December 2009. Currently, he is an associate professor with the State Key Laboratory of Networking and Switching Technology, and vice-director of the Next Generation Internet Technology Research Center at BUPT. He has published over 100 technical papers in prestigious international journals and conferences. His research interests include wireless networking, multimedia communications, and next generation Internet technology. He serves as a Co-Chair and Technical Program Committee (TPC) member for a number of international conferences and workshops.

Shijie Jia received his Ph.D. degrees in communications and information system from BUPT in March 2014. He is a lecturer at the Academy of Information Technology, Luoyang Normal University, Henan, China. His research interests include next generation Internet technology, wireless communications, and peer-to-peer networks.

Lujie Zhong received his Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences, in July 2013. She is a lecturer in the Information Engineering College, Capital Normal University, China. Her research interests are in the areas of communication networks, computer systems and architecture, mobile Internet technology, the Internet of things, and vehicular networks.

Gabriel-Miro Muntean [M] received his Ph.D. degree from Dublin City University (DCU), Ireland, for research in the area of quality-oriented adaptive multimedia streaming in 2003. He is a senior lecturer with the School of Electronic Engineering at DCU, co-director of the DCU Performance Engineering Laboratory, and consultant professor with BUPT. His research interests include quality-oriented and performance-related issues of adaptive multimedia delivery, performance of wired and wireless communications, energy-aware networking, and personalized technology-enhanced learning. He has published over 250 papers in prestigious international journals and conferences, has authored three books and 16 book chapters, and has edited six other books. He is an Associate Editor for *IEEE Transactions on Broadcasting* and *IEEE Communications Surveys and Tutorials*, and a reviewer for other important international journals, conferences, and funding agencies. He is a member of the IEEE Broadcast Technology Society.

When Crowdsourcing Meets Mobile Sensing: A Social Network Perspective

Pin-Yu Chen, Shin-Ming Cheng, Pai-Shun Ting, Chia-Wei Lien, and Fu-Jen Chu

ABSTRACT

Mobile sensing is an emerging technology that utilizes agent-participatory data for decision making or state estimation, including multimedia applications. This article investigates the structure of mobile sensing schemes and introduces crowdsourcing methods for mobile sensing. Inspired by social networks, one can establish trust among participatory agents to leverage the wisdom of crowds for mobile sensing. A prototype of social-network-inspired mobile multimedia and sensing application is presented for illustrative purposes. Numerical experiments on real-world datasets show improved performance of mobile sensing via crowdsourcing. Challenges for mobile sensing with respect to Internet layers are discussed.

INTRODUCTION

Wireless sensor networks (WSNs) explore avenues to collect and use information from the physical world by deploying low-cost tiny sensor nodes on the ground, in the air, under water, on bodies, in vehicles, and inside buildings. With sensing, processing, and communication capabilities, networked sensor nodes cooperatively collect information on entities of interest, and WSNs have emerged as a promising technology with numerous and various applications. As shown in Fig. 1, sensor nodes locally collect information and then forward the sensed result over a wireless medium to a remote static sink, where it is fused and analyzed in order to determine the global status of the sensed area. In order to successfully gather sufficient information, a static sink could send a mobile agent to collect data from individual sensor nodes by following a trajectory spanning all the nodes (Fig. 1).

To accomplish large-scale sensing, the WSN evolves not only at the sink side (e.g., mobile agents), but also at the sensor node side. Mature mobile networks consisting of mobile devices with advanced processing and communication capabilities become a possible sensing infrastructure of WSNs. By exploiting the rich set of embedded sensors (e.g., camera, gyroscope, GPS, accelerator, light sensor, and digital compass) on mobile phones as sensor nodes, a new paradigm of WSNs is realized, which is known as *mobile sensing* [1–9]. As shown in Fig. 1, mobile sensing utilizes crowd-sensed information for data analysis and decision making due to penetration of mobile devices as well as human mobility and ubiquity. It relies on the wisdom of crowds [10] to successfully infer the information of interest and accomplish its tasks. The data from mobile crowds (users, sensors, robots, etc.) can be either numerical or categorical, depending on applications. Examples of crowd-sensed data include numerical environmental measurements such as temperature and air conditions [3, 5], personal activities such as daily life patterns and events [2], interactions among people such as crowd density [7] and common interests [6], categorical recommendations such as ratings for nearby restaurants [1], and user experience/quality feedback on wireless mobile multimedia applications [1].

It is worth noting that many multimedia applications lie within the scope of mobile sensing, since extracting and analyzing the information sensed or generated from the crowds is one of the core goals for many multimedia applications in order to attract users' attention. Better prediction of users' interests leads to longer multimedia stickiness, and hence more revenues can be expected. Modern multimedia applications often pull user-centric information from the crowds and offer personalized contents (e.g., the next video to watch). Typically, location and social network information are widely used for targeted advertisement and recommendation. Therefore, the major challenge is to efficiently and accurately extract user-centric information from the crowds and identify users of high similarity for improved content delivery.

In recent years, many machine learning tasks and business models have leveraged the wisdom of crowds to acquire crowdsourced data for discriminating unknown objects. The website *Galaxy Zoo* asks visitors to help classify the shapes of galaxies, and the website *Stardust@home* asks visitors to help detect interstellar dust particles in astronomical images. Business models such as *Amazon Mechanical Turk* (*MTurk*) and *CrowdFlower* provide crowdsourcing services at low prices. For *MTurk*, a minimum of US$0.01[1] is paid to a labeler/worker when he/she makes a click (i.e., generates a label) for an item. Despite

Pin-Yu Chen and Pai-Shun Ting are with the University of Michigan

Shin-Ming Cheng is with National Taiwan University of Science and Technology.

Chia-Wei Lien is with Amazon Corporate LLC.

Fu-Jen Chu is with the Georgia Institute of Technology.

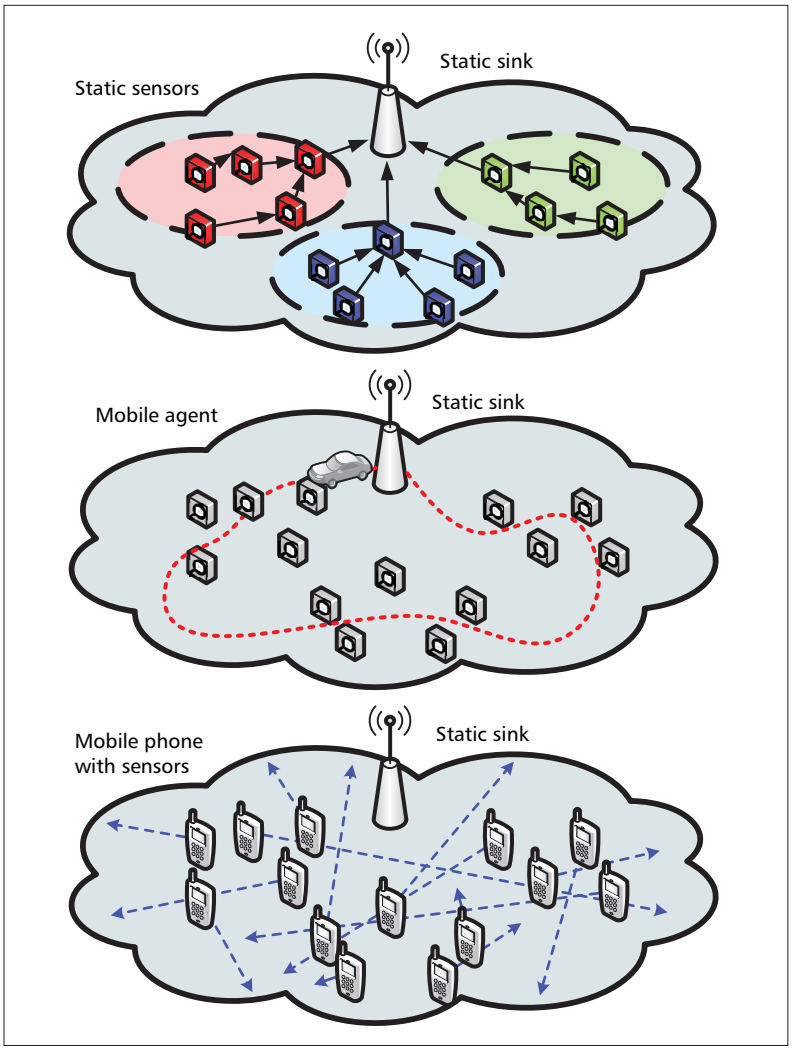


Figure 1. Evolution from wireless sensor networks to mobile sensing.

the low costs for acquiring crowdsourced data, one of the major challenges of mobile sensing is rooted in dealing with noisy and potentially erroneous data [4, 8, 11]. These undesired data can originate from environmental/object uncertainties (e.g., channel noise and difficulty of object discrimination) or user intentions (e.g., fraudulent recommendations and irresponsible user clicks). Consequently, identifying trustworthy data and reliable agents becomes an essential task in mobile sensing [4, 5, 8].

Utilizing the concept of trust in social networks, this article proposes trust-based data analysis approaches for crowdsourcing in mobile sensing schemes. In addition to scraping trustworthy data from the bulk, these approaches aim to identify reliable agents for performance enhancement. A weight of trust is built on the reliability of each agent for mobile sensing via a limited number of training queries. For spectrum sensing, these approaches can be implemented by broadcasting reference signals for reception power calibration. For annotation, these approaches can be implemented by uploading some items with known answers. For multimedia, these approaches can be associated with user behaviors based on the provided contents.

[1] https://requester.mturk.com/pricing

This article summarizes mobile sensing network paradigms and introduces several trust-based crowdsourcing methods for mobile sensing. We also illustrate a prototype application of a social-network-inspired mobile multimedia and sensing scheme. Numerical experiments on real-world datasets show that mobile sensing can benefit from crowdsourcing for improved performance. Potential challenges of social-network-based mobile sensing with respect to mobile multimedia Internet layers are discussed. Thus, this article sheds new light on integration of social networking and mobile sensing, and applications therein.

Mobile Sensing Paradigms

In mobile sensing, people share and distribute sensed information via physical proximity or social relations over portable sensors. As illustrated in Fig. 2, a mobile user plays the roles of both *querier* and *collector*, who request and provide information, respectively. A querier can simultaneously be a collector if he/she also participates in mobile sensing. Network structures of mobile sensing can be classified into two categories: the *direct* and *indirect* paradigms, which are described as follows.

Direct Mobile Sensing Paradigm

A direct mobile sensing paradigm involves direct communication between a querier and crowds (i.e., the collectors). Typically, it is achieved by adopting current device-to-device (D2D) communication technologies, such as WiFi Direct, ZigBee, Bluetooth Low Energy (BLE), and near field communication (NFC). In such a case, the *store-carry-forward* behavior facilitates information delivery in an ad hoc fashion. That is, sensed information may be stored in a sensor node in the absence of immediate connectivity to any other node, and relayed to other sensor nodes at encounters. Examples include the following.

Proximity sensing in mobile social networks (MSNs): It supports social platforms among physically proximate mobile users. For instance, one can simply scan the environment for discoverable Bluetooth devices to analyze crowd density and crowd flow direction [7]. By exploiting P2P communications, one can further make new social interactions with nearby devices. A popular example is sensing "potential friends with similar interests nearby." To enjoy such new activities, mobile users have to provide their own interests for profile matching by broadcasting their personal profiles to all nearby users, and then comparing their personal profiles and other users' profiles for friend matching [6].

Cooperative spectrum sensing in cognitive radio networks (CRNs): Unlicensed secondary users (SUs) sense the surrounding environment and exploit spectrum holes unoccupied by licensed primary users (PUs) for secondary transmission with minimal interference to PUs. To achieve better spectrum management and enhance radio resource utilization, a querier could exploit observations on local spectrum vacancy from surrounding SUs (i.e., crowds). The empowerment of cooperative spectrum sensing improves the throughput of wireless

communications and reduces potential interference among heterogeneous systems.

Indirect Mobile Sensing Paradigm

In this paradigm a querier and crowds are indirectly connected through a communication system in a centralized or distributed fashion. Typically, access points in a WLAN and a base station in a cellular network or WiMAX are exploited as communication paths in the former case. In the latter case, a querier/collectors could download/upload data from/to nearby relays via localized communication technologies such as WiFi-direct, BLE, or NFC. Examples follow.

Environmental measurements: Collectors provide local measurements (e.g., temperatures, air pollution indexes) to a querier via an existing cellular infrastructure for event detection or state estimation. Consequently, the current environment can be understood and improved [3, 5]. For example, the PEIR project [3] exploits sensors in mobile phones to build a system that tracks the impact of individual actions on carbon emissions.

Personal activity sharing: A collector shares his/her daily life patterns, activities (e.g., sports), health (e.g., heart rate, blood pressure) with his/her friends using online social networks. For example, by automatically classifying events in people's lives via sensors on mobile phones, CenceMe [2] enables selective event sharing among friends using Facebook or Twitter.

Online recommendation: Crowds (e.g., data collected from proximal users or users of high similarity) provide recommendations to a user-centric query, such as the best seafood restaurant within two miles, or the next video to watch for multimedia applications. For example, MicroBlog [1] encourages users to record multimedia blogs manually or automatically (via sensors). Moreover, the blogs from collectors in the same area are integrated to enrich the contents. Consequently, a querier can browse multimedia blogs at a selected region for relevant information.

Annotation: Crowds (e.g., machines, people) annotate labels, such as scenery labels for a picture or comments and interactions for multimedia contents, for an item requested by a user. One typical example is the Amazon Mechanical Turk (MTurk) service.

LifeTie: A Prototype Social-Network-Inspired Mobile Multimedia and Sensing Application

For further illustration, in this section we introduce a social-network-inspired mobile multimedia and sensing application. In Taiwan, mountainous areas are hikers' paradise. Hikers are used to tie trail marking ribbons on trees for direction guidance. It is a matter of life and death to clearly know one's own location, especially at night. However, trail marking ribbons have several disadvantages such as misinterpretation, lack of detailed information, and environmental pollution caused by overuse. To ensure hikers' safety while overcoming the aforementioned drawbacks, we propose a mobile sensing system named LifeTie.

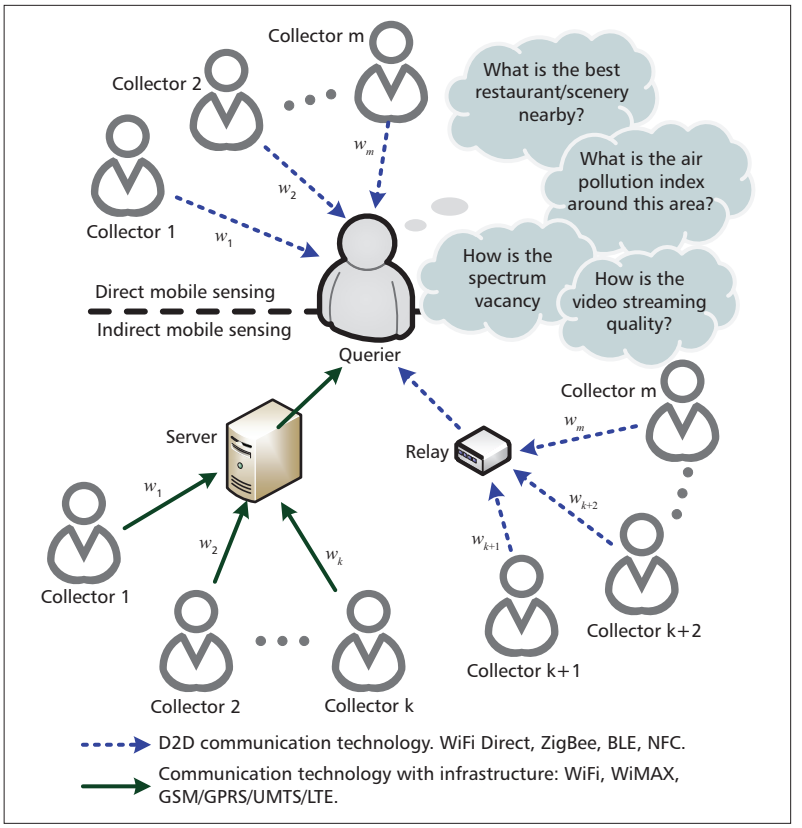


Figure 2. Network architecture of mobile sensing.

Unlike traditional WSNs deployed in mountains for wildlife tracking and ecological monitoring, LifeTie acts as an annotation platform where users can exchange their sensing results. The integration of tail marking ribbon and NFC technology replaces the traditional marking method with a smartphone app and achieves the purposes of information exchange, rescue, and search.

The main concept of LifeTie is to exploit NFC tags as the enabler for information exchange among people. To achieve that, NFC tags shall be attached around a mountain, thus creating an infrastructure. Hikers can trigger NFC tags nearby (typically in the range of only a few inches) using their NFC-enabled mobile phones. Then the NFC-enabled mobile phones can read/write data from/to the NFC tags. With embedded memory, LifeTie could handle a vast amount of data, which opens up the possibility of multimedia information and advances toward integration of social network and mobile sensing. The prototype of LifeTie is shown in Fig. 3. The features and functionalities of LifeTie are summarized as follows.

Features

Flexible: The shape of LifeTie was inspired by zap-straps. It can easily be tied to tree branches.

Recognizable: With the color of fluorescent orange and the addition of reflective stickers, it provides direction guidance even at night time.

Reliable: Polypropylene (PP) is used in LifeTie for its bendability and durability. An NFC tag is integrated in the internal parts of a strap to resist extreme weather conditions in mountains.

Affordable: The cost of an NFC tag is low.

Figure 3. LifeTie mobile sensing system — the physical device and its mobile multimedia interface.

Power saving: An NFC tag is not powered by electricity. As a result, no replacement for LifeTie is required, and therefore even lower deployment cost can be achieved.

Functionalities

Navigation and warning: Via the corresponding app, hikers can check out LifeTie's guestbook right after triggering NFC tags. Depending on the current environmental conditions, hikers can leave comments or draw a simple map to make a notification or assist navigation. Some useful warning icons (e.g., cliff, snake, wasps, slippery conditions) and guiding icons (e.g., cave, camp) are provided when drawing the map to enable diverse multimedia contents. The updated surrounding information facilitates following hikers.

Tracking and rescue: When a hiker is lost in a mountain and finds LifeTie, he/she can check regular comments to see if there is any shelter nearby. Moreover, he/she can leave urgent comments highlighted in red. If he/she can leave such information on several LifeTies, rescuers can easily identify a rough search area according to the positions of deployed LifeTie and timestamps of the comments. As a result, the rescuing operations become more effective and efficient.

Trust-Based Crowdsourcing Methods for Mobile Sensing

This section provides an overview of weight (trust) assignment methodologies on agents for crowd-sensed data in mobile sensing. The utility of these methods is investigated in the next section, and the challenges toward practice are addressed after that. For crowdsourcing-empowered mobile sensing, a user (or an intermediate system) evaluates a weight of trust w_i for the ith agent and fuse information from agents via weighted combination of agents' observations for the user's query. As shown in Fig. 4, the collected data from the agents can be viewed as a matrix with rows representing agents and columns representing observations associated with queries. The training queries refer to queries with known answers and are used for weight evaluation. The number of agents is denoted by m.

The final output for mobile sensing is the combination of each agent's observation multiplied by the associated weight. Here we discuss several crowdsourcing methods involving different weight evaluation approaches. These methods can be classified into two categories, unsupervised and supervised, separated by the need for training queries.

Unsupervised Crowdsourcing Methods

Majority votes: Majority votes adopts uniform trust among all agents (i.e., the weight $w_i = 1/m$ for all i) and selects the observation on which most agents agree as the final output. This method may lead to poor performance when the majority of agents have incorrect observations or some observations are maliciously manipulated.

Probabilistic inference: Probabilistic inference assumes that each observation made by an agent is statistically independent and imposes a statistical model to infer the weights from observations. One popular method is the weight evaluation method based on an expectation maximization (EM) algorithm [5, 12].

Supervised Crowdsourcing Methods

Supervised crowdsourcing methods aim to find the optimal weight of each agent by solving the optimization problem as

$$\text{minimize}_{\mathbf{w}}\ \text{cost}(\text{training-queries}, \text{final-output}) + \lambda \cdot R(\mathbf{w}),$$

where $\mathbf{w} = [w_1, w_2, \ldots, w_m]$ is the vector of weights, $R(\mathbf{w})$ is a regularization function for $\mathbf{w}$, and $\lambda \geq 0$ is the regularization parameter for $R(\mathbf{w})$. Here we introduce several supervised crowdsourcing methods.

Weighted averaging: Weighted averaging is a heuristic weight evaluation approach which assigns a weight that is proportional to the accuracy of each agent in the training queries. Let q_i be the fraction of correct queries responded by agent i. The weight of agent i is the normalized accuracy $w_i = q_i / \Sigma_{i=1}^{m} q_i$.

Exponential weighted algorithm: The exponential weighted algorithm adopts exponential cost function and zero regularization parameter ($\lambda = 0$) and sequentially adjusts the weight of each agent from the training queries. Interested readers can refer to [13, references therein] for more details.

Support vector machine: The support vector machine adopts the Hinge loss function as its cost function and assumes the regularization function $R(\mathbf{w}) = \Sigma_{i=1}^{m} w_i^2$ and positive regularization parameter (i.e., $\lambda > 0$). The support vector machine aims to find a separating hyperplane that best discriminates the training queries in the data sample space, and the weight of each agent can be determined by the resulting separating hyperplane. Interested readers can refer to [13, references therein] for more details.

Professional search: Inspired by social networks where problems are often resolved by professionals, professional search aims to assign weights to only a few agents that have outstanding accuracy in the training queries [14]. Professional search adopts the Hinge loss function as its cost function and assumes the regularization function $R(\mathbf{w}) = \Sigma_{i=1}^{m} |w_i|$. This regularization function is known as a surrogate function that promotes sparsity in $\mathbf{w}$ (i.e., most of the weights are zero), and hence the professionals hidden in the crowds can be selected for mobile sensing.

Numerical Experiments

In this section we use two crowd generated datasets to investigate the performance of the crowdsourcing methods in the previous section. For the first dataset, each agent only participates in some fraction of queries and hence resembles the dynamic participatory nature in mobile sensing. For the second dataset, almost every agent responds to each query, but none of the agents have correct answers to all queries, which resembles the imperfect sensing capability in mobile sensing. In both scenarios, crowdsourcing methods can improve the query classification performance by identifying trustworthy agents.

For crowdsourcing methods involving a regularization function R, we use a leave-one-out-cross-validation (LOOCV) approach [13] to determine the optimal regularization parameter λ, by swiping λ from 0 to 200 to select the optimal value that leads to minimum training error. A one-to-all classification approach is used for multiple (more than two) categorical datasets (e.g., the exam dataset).

Text Relevance Judgment

The text relevance judgment dataset is provided by the Text Retrieval Conference (TREC) crowdsourcing track in 2011,[2] where 689 agents (participants) are asked to judge the relevance of paragraphs excerpted from a subset of articles with given topics. Each agent then generates an observation, either "relevant" or "irrelevant," for an article. It is worth mentioning that this dataset is sparse in the sense that on average each agent only read roughly 26 out of 394 articles. For supervised crowdsourcing methods we use roughly 10 percent (40 training queries) of articles to evaluate each agent's weight. The rest of the data samples are used to test the accuracy of crowdsourcing algorithms, and the results are summarized in Table 1. It is observed that supervised methods can achieve higher accuracy than unsupervised methods via training queries. Also note that the support vector machine and professional search outperform other methods since their main objective is to assign more weights on the trustworthy agents/data samples possessing eminent discriminant capability.

Science Exam Dataset

The science exam dataset is collected by the authors[3] and contains 40 questions. Each question has four choices, and the correct answer is one of these four choices. There are 183 agents (students) taking the exam and producing observations (their answers). Unlike the TREC dataset, this exam dataset is dense in the sense that almost every student has provided an answer for each question. We use 10 questions as training queries and the rest to test the accuracy. The results are summarized in Table 2. The baseline accuracy (random guess) is 25 percent. None of the methods can achieve accuracy as high as in the TREC dataset due to the following facts:

- The observations of the TREC dataset only have two categories, whereas the observations of the exam dataset have four categories, which renders the latter more difficult to discriminate.
- The exam is challenging since the majority votes method leads to low accuracy, and there is no student who answers all questions correctly. The best student only has 70 percent accuracy.

Nonetheless, crowdsourcing methods such as the support vector machine and professional search can still attain relatively good accuracy by identifying reliable agents.

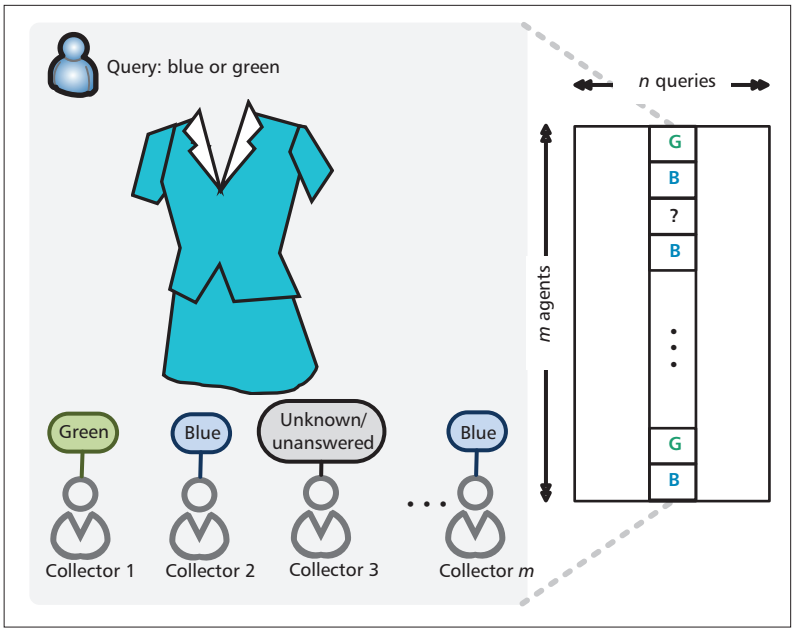


Figure 4. Illustration of crowdsourced data.

Ongoing Challenges: Aspects from Mobile Multimedia Internet Layers

In this section we discuss some ongoing challenges toward integration of social networking and mobile sensing, particularly in aspects of the mobile multimedia Internet. Issues corresponding to each layer of the mobile multimedia Internet are specified as follows.

Application Layer

Conflict between privacy and trustworthiness: Although such integration of social networking and mobile sensing is exciting and promising, the collection and sharing of personal information related to human activity introduces the concern of privacy, where participants are reluctant to reveal any sensitive personal information (e.g.,

[2] https://sites.google.com/site/treccrowd/2011

[3] The science exam dataset can be downloaded from the first author's website, https://sites.google.com/site/pinyuchenpage/.

The amount of collected information and the number of participants are proportional to the degree of volunteering. When the amount of collected information is insufficient, the sensing results might not be precise. Thus, mobile sensing needs incentive and mechanism design to encourage people to participate.

	Unsupervised		Supervised			
Methods	Majority votes	Expectation maximization	Weighted averaging	Exponential weighted algorithm	Support vector machine	Professional search
Accuracy (%)	79.38	78.81	83.05	80.51	83.33	84.46

Table 1. The TREC 2011 dataset. Supervised methods attain higher accuracy than unsupervised methods via acquiring a few training queries for weight (trust) assignment.

	Unsupervised		Supervised			
Methods	Majority votes	Expectation maximization	Weighted averaging	Exponential weighted algorithm	Support vector machine	Professional search
Accuracy (%)	46.67	50	46.67	26.67	50	53.33

Table 2. The science exam dataset. Despite the fact that no students answer all questions correctly, professional search can still achieve more than twice the accuracy of random guess (25 percent accuracy).

time, location, pictures, sound, acceleration, and biometric data). As a result, it is crucial to design an approach to collecting sufficient information from participants without violating their privacy [8]. Specifically, authentication shall be supported to identify legal mobile users and adversaries. Moreover, anonymity shall be preserved to hide sensitive information by using technology such as *k*-anonymous or pseudonyms. These avenues prevent adversaries traversing a relationship between users' contributions and identities.

In particular, trust-based crowdsourcing methods aim to preserve trustworthiness in harsh environments where malicious participants may deliberately feedback fraudulent data. Obviously, to counteract this effect we need to observe contributions made by each user for a period of time and hence evaluate his/her trustworthiness. However, it may conflict with the privacy consideration where actual attribute values of a specific user are obscured, and the links between multiple contributions from the same user are broken. How to acquire linkability across multiple contributions from the same user while preserving privacy is a challenging issue [5].

Data integrity on complicated multimedia content: By manually recording via users or automatically collecting via sensors, a huge amount of information can be retrieved in mobile sensing. The multimedia contents generated from the retrieved materials via applications like MicroBlog [1] contain abundant but complicated information, which burdens data integrity. Unlike the example raised in Fig. 4 where we just need to identify the color of cloth, multiple parts in one clip or video might lead to a distinct conclusion that is highly dependent on a viewer's ideology. Extra meta information shall also be included to increase data integrity like user reviews (e.g., iMDB, Youtube) or user preference (e.g., Netflix).

Incentives for participation: Mobile sensing requires participants to spend their time, attention, and mobile phones' battery power for contributing data. Obviously, the amount of collected information and the number of participants are proportional to the degree of volunteering. When the amount of collected information is insufficient, the sensing results might not be precise. Thus, mobile sensing needs incentive and mechanism design to encourage people to participate [15].

Network Layer

Data retrieval in a distributed environment: In distributed scenarios (e.g., proximity sensing, spectrum sensing, and annotation), a querier can only retrieve information from localized collectors, which might lead to biased inference results. To overcome this issue, current researchers propose to enable information relay for each individual. Leveraging human mobility and store-and-forward features, the amount of data collected from crowds grows substantially via information exchange, thereby improving the accuracy of estimation. In addition, in the indirect mobile sensing paradigm involving data retrieval and analysis from distributed systems (e.g., data storage servers), distributed computation is known to be one of the big data challenges.

The limitations of D2D communications: In the direct sensing paradigm, a querier could communicate with collectors via D2D communications like WiFi Direct, ZigBee, BLE, and NFC. However, current D2D communication technologies typically require manual mutual authentication when making a connection, which is unfavorable for automatic data collection and device connectivity. Moreover, mobile sensing applications should be capable of integrating the features of different D2D communication technologies (e.g., transmission ranges, transmission rates, and power consumption) for ubiquitous multimedia content delivery.

Link Layer

Unreliable link due to mobility and interference: The mobile nature of users and agents may hinder the performance of mobile sensing due to change in location, environment, and participatory agents. The highly dynamic positions of agents incur ever

changing received interference, thereby affecting link reliability. To accommodate this effect, a crowdsourcing-aided mobile sensing method should possess adaptivity and robustness in such a dynamic situation in order to identify inadequate agent participation and obsolete data collection.

Physical Layer

Trade-off between power consumption and sensing accuracy: Apparently, the higher the sensing accuracy attained, the greater the amount of energy is consumed in mobile devices with limited power due to the increase in data acquisition frequency, which might violate the design rationale of sensing paradigms. Consequently, power consumption fairness and energy-efficient scheduling for participatory devices should be considered jointly, and new throughput measures should be studied to balance the trade-off between power consumption and sensing accuracy.

Conclusion

This article proposes to incorporate crowdsourcing methods for mobile sensing and introduces several crowdsourcing methods for evaluating the weight of trust among agents. The direct and indirect mobile sensing network paradigms are discussed. A prototype of social-network-inspired mobile multimedia and sensing application is illustrated toward integration of social network and mobile sensing. Numerical experiments on real-world datasets show that mobile sensing can benefit from crowdsourcing methods for performance improvements. Ongoing challenges of integration of social networking and mobile sensing are addressed through the aspects of mobile multimedia Internet layers. This article therefore paves new avenues to various mobile applications and future mobile technology development.

Acknowledgment

This research is sponsored by the Ministry of Science and Technology, Taiwan, under the contracts of 103-2221-E-011-008-MY3 and 104-2221-E-011-051. We would also thank to Kai-Ching Wang, Wei-Ju Kao, and Ting-Wen Lin for their efforts on implementing the mobile sensing system LifeTie.

References

[1] E. Gaonkar *et al.*, "Macro-Blog: Sharing and Querying Content through Mobile Phones and Social Participation," *Proc. ACM MobiSys 2008*, June 2008, pp. 174–86.
[2] E. Miluzzo *et al.*, "Sensing Meets Mobile Social Networks: The Design, Implementation and Evaluation of the CenceMe Application," *Proc. ACM SenSys 2008*, Nov. 2008, pp. 337–50.
[3] M. Mun *et al.*, "PEIR, The Personal Environmental Impact Report, as a Platform for Participatory Sensing System Research," *Proc. ACM MobiSys 2009*, June 2009, pp. 55–68.
[4] R. Ganti, F. Ye, and H. Lei, "Mobile Crowdsensing: Current State and Future Challenges," *IEEE Commun. Mag.*, vol. 49, no. 11, Nov. 2011, pp. 32–39.
[5] C. Xiang *et al.*, "PassFit: Participatory Sensing and Filtering for Identifying Truthful Urban Pollution Sources," *IEEE Sensors J.*, vol. 13, no. 10, Oct. 2013, pp. 3721–32.
[6] R. Zhang *et al.*, "Privacy-Preserving Profile Matching for Proximity-Based Mobile Social Networking," *IEEE JSAC*, vol. 31, no. 9, Sept. 2013, pp. 656–68.
[7] J. Weppner *et al.*, "Participatory Bluetooth Scans Serving as Urban Crowd Probes," *IEEE Sensors J.*, vol. 14, no. 12, Dec. 2014, pp. 4196–4206.
[8] D. He, S. Chan, and M. Guizani, "User Privacy and Data Trustworthiness in Mobile Crowd Sensing," *IEEE Wireless Commun.*, vol. 22, no. 1, Feb. 2015, pp. 28–34.
[9] H. Liu, Y. Zhou, and Y. Zhang, "Estimating Users' Home and Work Locations Leveraging Large-Scale Crowd-Sourced Smartphone Data," *IEEE Commun. Mag.*, vol. 53, no. 3, Mar. 2015, pp. 71–79.
[10] J. Surowiecki, *The Wisdom of Crowds*, Anchor, 2005.
[11] R. Snow *et al.*, "Cheap and Fast — But Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks," *Proc. Conf. Empirical Methods in Natural Language Processing*, 2008, pp. 254–63.
[12] V. C. Raykar *et al.*, "Learning from Crowds," *J. Machine Learing. Research*, vol. 11, Aug. 2010, pp. 1297–1322.
[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, 2001.
[14] P.-Y. Chen *et al.*, "Supervised Collective Classification for Crowdsourcing," IEEE GLOBECOM Wksp. 2015, San Diego, CA; arXiv:1507.06682.
[15] D. Yang *et al.*, "Crowdsourcing to Smartphones: Incentive Mechanism Design for Mobile Phone Sensing," *Proc. ACM Mobicom 2012*, Aug. 2012, pp. 173–84.

Biographies

Pin-Yu Chen [S'10] received his B.S. degree in electrical engineering and computer science (undergraduate honors program) from National Chiao Tung University, Taiwan, in 2009, and his M.S. degree in communication engineering from National Taiwan University, Taipei, in 2011. He is currently working toward his Ph.D. degree in electrical engineering and computer science at the University of Michigan, Ann Arbor. His research interests include network science, interdisciplinary network analysis, graph clustering and community detection, statistical graph signal processing, cyber security, and their applications to data analysis and communication systems. He is a member of the Tau Beta Pi Honor Society and Phi Kappa Phi Honor Society, and was a recipient of the Chia-Lun Lo Fellowship. He was also the recipient of the IEEE GLOBECOM 2010 GOLD Best Paper Award, an IEEE ICASSP 2014 NSF travel grant, and an IEEE ICASSP 2015 SPS travel grant.

Shin-Ming Cheng received his B.S. and Ph.D. degrees in computer science and information engineering from National Taiwan University in 2000 and 2007, respectively. He was a postdoctoral research fellow at the Graduate Institute of Communication Engineering, National Taiwan University, from 2007 to 2012. Since 2012, he has been with the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, as an assistant professor. His current research interests include mobile networks, wireless communication, information security, and complex networks. He was a recipient of the IEEE PIMRC 2013 Best Paper Award and the 2014 ACM Taipei/Taiwan Chapter K. T. Li Young Researcher Award.

Pai-Shun Ting is a Ph.D. student in the Department of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor. His current research concerns design automation and mathematical analysis of stochastic logic networks. He earned his M.S. degree from the University of Michigan in 2013 and his B.S. degree from National Taiwan University in 2011, both in electrical engineering. Before joining the Ph.D. program in 2014, he worked on machine learning, image processing, and computer graphics as a research assistant at the National Institute of Informatics, Chiyoda-ku, Tokyo, Japan. He is currently a student member of the University of Michigan's Advanced Computer Architecture Laboratory.

Chia-Wei Lien received her M.S. degree in electrical engineering systems from the University of Michigan in 2014 and a B.S degree in electrical engineering from National Taiwan University. She joined Amazon.com, Inc. in 2014 and has been working on supply chain optimization modeling Amazon worldwide network planning. Her fields of interest include statistical signal processing, artificial intelligence, and machine learning.

Fu-Jen Chu received his B.S. degree in electrical engineering from National Taiwan University in 2011 and his M.S. degree in electrical engineering systems from the University of Michigan in 2014. He is currenly pursuing his Ph.D degree in robotics and intelligent machines at the Georgia Institute of Technology, Atlanta. His research interests include computer vision, machine learning, and neural networks, and he is also interested in control systems and multi-agent problems.

The highly dynamic positions of agents incur ever-changing received interference, thereby affecting the link reliability. To accommodate this effect, a crowdsourcing aided mobile sensing method should possess adaptivity and robustness in such a dynamical situation in order to identity inadequate agent participation and obsolete data collection.

Pervasive Data Sharing as an Enabler for Mobile Citizen Sensing Systems

Waldir Moreira and Paulo Mendes

ABSTRACT

Today, users can use their personal devices for a wide range of applications and services, such as controlling other devices, monitoring human physiological signals, and accessing information while on the move. Due to the communication and sensing capability of personal and wearable devices, their pervasive deployment and use may lead to an improvement of social and personal welfare by exploiting novel mobile citizen sensing systems. However, the pervasiveness of such large-scale sensing systems is only possible if devices are able to share sensing data independent of the available communication infrastructure, their location, and applications making use of the collected data. This article describes a set of paradigms that should be considered to build pervasive data sharing systems, and proposes a node architecture to implement them.

INTRODUCTION

Advancements in computing and communication technologies have resulted in mobile devices (e.g., smartphones, watches, glasses) encompassing a diversified set of communications (e.g., Bluetooth, Wi-Fi, third generation or 3G) and sensing (e.g., proximity, location, body signals) capabilities. Such devices have a direct impact on the evolution toward an Internet able to accommodate citizens' needs in real time by supporting the development of mobile citizen sensing (MCS) systems.

MCS systems are able to sense, process, and share sensing data produced by personal and wearable devices. By being able to exploit data collected throughout people's daily routines, MCS systems may have higher impact on different sectors of society (e.g., health care and civil protection) than wireless sensor networks made of specialized devices deployed for specific applications.

In order to support large-scale sensing applications (e.g., MIT SENSEable City Lab, Intel Urban Atmospheres project), a general-purpose MCS system should be able to exchange sensing data among trustworthy devices that need information about users' behavior and social context. In order to be truly pervasive, MCS systems should be supported by a networking system that allows the exchange of data among trustful mobile devices based on any communication opportunity, independent of the intermittent presence of 3G coverage or open WiFi networks. Such a pervasive data sharing (PDS) system should ensure high delivery data rates with low latency and cost, even when facing difficult networking conditions, while being agnostic of the location of devices.

MCS systems have various applications in real life, from improvement of daily life routines to the implementation of emergency/rescue applications [1]. For instance, in a disaster prevention scenario, users in a national park may use the MCS system installed in personal devices to collect information (e.g., photos, temperature and humidity readings) related to a potential fire situation. By making use of the PDS system installed in trusted personal devices carried by pedestrians and drivers, the collected sensing data can be made available to local authorities (e.g., firefighters), even in the absence of network infrastructure.

This article fills a gap in the literature: although there are proposals for data-centric networking over the Internet at large and delay-tolerant networking for challenging scenarios, there are no proposals for data-centric networking over challenging scenarios based on trusted devices that are willing to cooperate. The scientific contributions of this article are:

- Analysis of networking requirements and assumptions of pervasive MCS systems
- Definition of a set of design paradigms for the development of PDS systems
- Identification of a set of functional building blocks for the instantiation of the proposed design paradigms
- Description of a node architecture to implement the proposed set of PDS paradigms, based on the combination of data-centric networking and opportunistic communications among trusted and cooperative devices

NETWORKING IN MOBILE CITIZEN SENSING SYSTEMS

Sensing data in MCS systems (Fig. 1) is produced by different entities based on specific requirements, and may:

The authors are with COPELABS/University Lusófona Lisbon.

0163-6804/15/$25.00 © 2015 IEEE

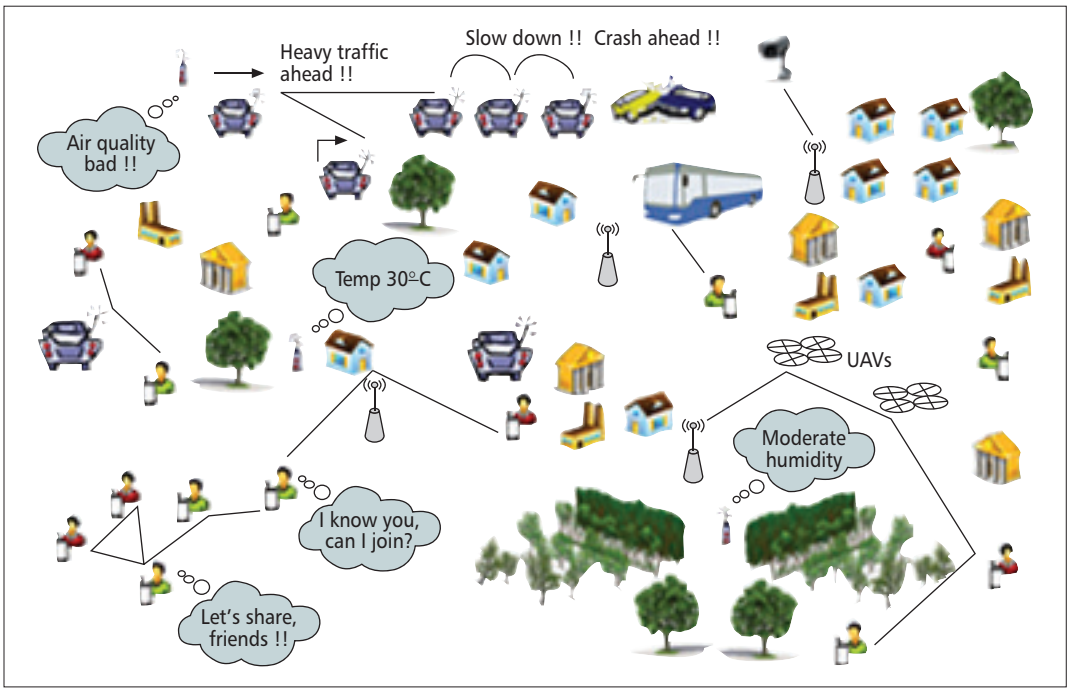


Figure 1. Communication in a sample MCS system: large-scale urban sensing scenario.

Despite the increased capabilities of personal devices, MCS systems cannot assume that their users are willing to share such capabilities. Thus, egoistic behavior must be overcome by incentive models to encourage users in engaging in the cooperative networking process.

- Span and be collected in different areas (i.e., in-building, citywide, forests)
- Be of different types (i.e., physiological, images, location, temperature)
- Be used for a variety of purposes (i.e., connected health, smart small grids, people-centric sensing applications)
- Be combined with data coming from different sources for a more general purpose (e.g., temperature and humidity combined to identify a fire hazard situation)

Due to the inherent pervasive and opportunistic nature of MCS systems, sensing should be deployed based on a networking system that is able to exploit any communication opportunity taking place among mobile nodes. As nodes are worn/carried by people and people have social relationships and data interests,[1] sensing data may be exchanged considering their social interactions and/or common interests [2].

Despite the increased capabilities (e.g., processing, storage) of personal devices, MCS systems cannot assume that their users are willing to share such capabilities. Thus, egoistic behavior must be overcome by incentive models to encourage users to engage in the cooperative networking process.

Furthermore, mobile communications should only rely on trusted devices that are willing to store and share data by taking advantage of communication opportunities: a malicious user may easily display cooperative behavior with the intention of accessing the users' sensing data. Thus, trust mechanisms must be in place to provide users with secure data exchange in dynamic scenarios.

When facing a large number of independent devices, it is necessary to rely on a data-centric networking approach [3, 4] in order to create a robust communication system independent of device location, while being aware of users' data interests. By focusing on the produced data and not on the device, the networking system is expected to scale while keeping the required robustness.

Considering these characteristics/requirements, next we analyze a set of paradigms that should be followed to design truly deployable PDS solutions to support MCS systems.

Design Paradigms for Pervasive Data Sharing

Current literature does not provide any guidelines for building PDS systems able to cope with the networking requirements of MCS systems. Hence, in this section we propose four paradigms to be considered in the design of any PDS system able to allow MCS systems to support applications of different sectors of society, such as health care, education, industry, and government agencies.

To start with, there is the need to create the conditions for realistic deployment of MCS systems. This means that a PDS system should operate based on a set of cooperative (paradigm #1) and trustful (paradigm #2) devices. Data exchange among trustful devices in pervasive scenarios is only possible by exploiting any communication opportunity (paradigm #3), which should ensure good performance. The efficiency of the PDS system is then ensured by the application of paradigm #4, which allows the system to scale, independent of the location of devices, reaching the desirable level of data robustness and reliability. We believe that using data-centric networking together with social-based opportunistic communications may facilitate the development and deployment of PDS systems: since nodes, which are constantly collecting and consuming data, are carried by users, they can per-

[1] Data interests in the context of this work can refer to the analyzed sensing data, the node's contextual data, or the information required by the sensing process (baseline data and learning models).

In a pervasive scenario, the PDS system should rely on the possibility to build trust on the fly: devices create trust circles based on the reputation of those with whom they interact and the impact that those devices had on previous interactions. This way, users can establish communication with desirable trust levels and safely share data.

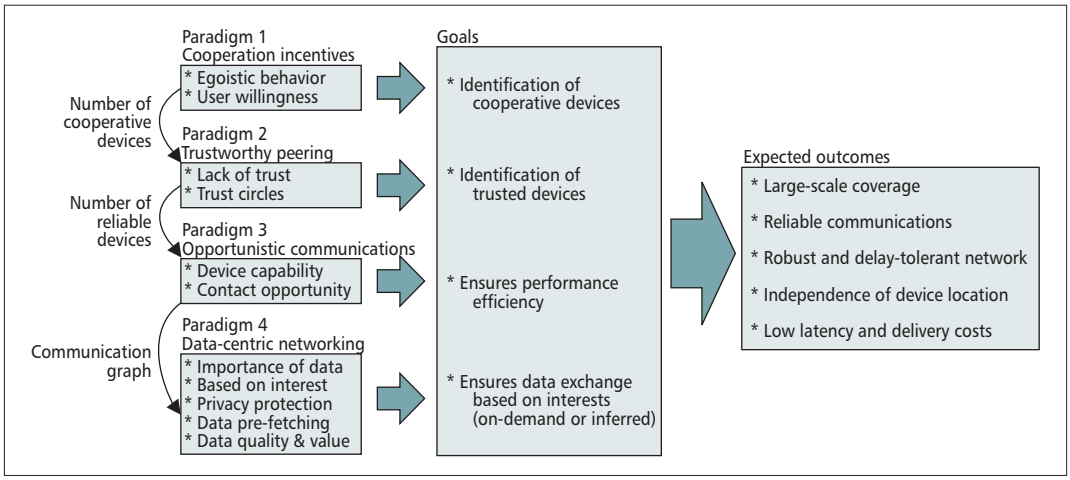


Figure 2. Design considerations for a PDS system.

vasively exchange such data through users' social interactions and data interests. Moreover, the exchange of data can be done with high probability, low cost, and low latency, while being agnostic about its location, which is beneficial in dynamic mobile scenarios.

Figure 2 shows how the four paradigms should come together, observing important correlation aspects that must be taken into consideration while building a PDS system. The rest of this section provides a description of each paradigm, including an analysis of current technical solutions that can be used to implement each one of them.

Paradigm #1: Cooperation Incentives

One may argue that paradigm #1 is not required given the fact that users are expected (i.e., willing) to cooperate at all times. This assumption must be handled with care: although devices have significant resources, their owners will certainly not share all the available resources to support pervasive data sharing [5]. Hence, a PDS system must encompass mechanisms to encourage cooperation. This can be done, for instance, by employing virtual currency or based on a reputation mechanism [6]. By employing virtual currency mechanisms, the user is rewarded for sharing sensing, storage, and communication resources. Such rewards can later be exchanged by the user for other resources, such as free Internet access. In reputation-based cooperation, the user's reputation increases inside the system as long as he/she cooperates with others. Such increased reputation makes the user very reliable and trustworthy in the system.

On the other hand, this paradigm may encompass a more aggressive approach aiming to penalize egoistic users, for instance, by discouraging other users from carrying/relaying their data.

Paradigm #2: Trustworthy Peering

Independent of users' willingness to share data, certain users may not trust all available neighbors to carry or relay their data. Paradigm #2 adds to the PDS system the ability to overcome this lack of trust, avoiding potential hazardous operational situations. For instance, data may be generated by malicious users with the intent of harming the functioning of the PDS system (e.g., a misbehaving user could create fake data causing the triggering of false alarms). This situation can be mitigated by identifying users who are trustworthy. Therefore, in a pervasive scenario, the PDS system should rely on the possibility to build trust on the fly: devices create trust circles (i.e., sets of trusted users) based on the reputation of those with whom they interact, and based on the impact that those devices had on previous interactions. This way, users can establish communication with desirable trust levels and safely share data [5].

The creation of trust circles can be complemented by using communication mechanisms that allow data sharing based on the notion of social relationships [7, 8] and shared interest [9] aspects (cf. paradigm #3): within their trust circle, users may find it easier to share data with users with whom they share social relationships and data interests.

Paradigm #3: Opportunistic Communications

With paradigms #1 and #2, PDS systems are built based on a set of cooperative and trusted users who are considered for data exchange. Paradigm #3 allows PDS systems to explore the networking capabilities of trustful devices to exchange data by making the most out of the different contact opportunities between devices. Note that paradigm #3 should not be employed prior to paradigms #1 and #2: we cannot assume that every user, despite having the potential to be a carrier/disseminator, is willing to cooperate or can be trusted.

With paradigm #3, users are able to cooperate and/or can be trusted, and are able to exchange data based on the probability of meeting suitable carriers over any wireless interface (e.g., Bluetooth, Wi-Fi) instead of relying only on the probability of finding an open Wi-Fi access point or the availability of expensive 3G connectivity. This is what we refer to as the *right-here-right-now* approach, where a carrier can be a device or even an open Wi-Fi access point con-

figured with a set of data interests that can allow the immediate exchange of data.

Different solutions have emerged to allow this right-here-right-now approach [2]. Since devices are used by people who happen to have social relationships and diverse data interests, information may be exchanged considering the social interactions existing between users and/or the common interests they share [8–10].

Nevertheless, of importance to the success of PDS systems is the trade-off between the delivery probability and latency experienced by such an opportunistic PDS system, as well as the robustness of the overall sensing system: modern control theory is largely based on the abstraction that data is shared over perfect communication channels, which is not realistic to assume in large-scale urban sensing systems.

Paradigm #4: Data-centric Networking

While paradigm #3 ensures opportunistic operation over trustworthy pervasive scenarios, paradigm #4 guarantees that the PDS system reduces the network load by exchanging data based on users' interest and privacy considerations, and pre-fetching data near interested users.

The efficiency of a data-centric approach relies on selective management of data to tackle users' interests and privacy concerns. By understanding the relevance that data has to the user and to the entities interacting with him/her, network (e.g., bandwidth) and device (e.g., battery) resources can be spared by reducing the number of data transmissions.

Since users are normally concerned about their privacy, the PDS system needs to be able to selectively share data based on privacy concerns, that is, when the device does not have the computational power needed to analyze the collected sensing data. Still, a mechanism to ensure data privacy and anonymous sensing operations must not rely on registration authorities or centralized task services [11].

Besides the need to selectively handle data, the robustness and reliability of a PDS system depends on the availability of useful data. This can be ensured by pre-fetching data based on data interest inference, improving resource utilization and the availability rate of data to the sensing applications.

Functional Blocks for Pervasive Data Sharing

This section overviews relevant functional blocks that can be used to implement each of the proposed four paradigms, as well as their implicit/explicit alignment with one another.

Paradigms #1 and #2 refer to users' engagement in the PDS cooperation process as trustworthy peers. Such engagement can happen based on either:

- The trust level among users
- Rewarding those who engage in cooperation

A suitable cooperation framework may consider:

- A reciprocity-based incentive mechanism that takes into account users' reputation (i.e., levels of trust) to allow cooperation in scenarios where nodes know each other
- A reward-based incentive mechanism that encourages nodes to cooperate by allowing the exchange of virtual currency, which overcomes the lack of trust [6, 12]

Such a cooperation framework matches the requirements of a pervasive networking scenario: cooperation happens independent of how trusted the environment is. This is a desired feature for a PDS system given the different contexts in which users will find themselves (e.g., their known communities, another country).

Regarding the trust framework, different mechanisms to properly reflect trust associations between users may be used [5]: users are uniquely identified by means of virtual identities based on cryptography to reduce impersonation and non-repudiation issues; users can explicitly set how they trust unknown users (i.e., dispositional trust). Trust computation is given by the different trust associations, and may be influenced by local (e.g., user reputation) and external (e.g., presence of malicious users in the vicinity) aspects.

These trust mechanisms allow the creation of trust circles to allow reliable, trusted communications that may be more efficient than hard-coded security (e.g., public keys) in pervasive dynamic scenarios.

One can clearly see that solutions for paradigms #1 and #2 are closely aligned and intertwined: cooperation among trusted users happens easily. However, such alignment is not as clear for the solutions employed in the remaining paradigms.

Regarding paradigm #3, there are several opportunistic forwarding solutions that may be exploited in a PDS system, ranging from flooding approaches to more elaborate ones encompassing different social features (i.e., common communities, shared interests, popularity, dynamic social behavior of users) [2].

By considering social features, solutions for paradigm #3 implicitly relate to paradigms #1 and #2 to some extent: users can easily cooperate and are prone to trust others with the same interests or who belong to the same community, for instance. Nevertheless, paradigms #1 and #2 are required for real application deployment as users may still be reluctant to exchange data even with those with whom they share interests, communities, or some other level of social affinity. Additionally, solutions for paradigm #3 must be used to allow the system to be aware of the dynamics of users' social behavior, as this is beneficial for opportunistic communications in dense urban scenarios [13].

With this in mind, solutions such as Bubble Rap [7], dLife [8], and SCORP [9] fit paradigm #3: these proposals consider how users are socially connected, the communities in which they belong and how important they are in the system, and take into account the users' interests on the data traversing the network. The features of these social-aware and content-based solutions are aligned with the characteristics of nodes operating over a large-scale urban sensing scenario: users are very dynamic, focused on their interest on sensing data, desire anytime/anywhere data exchange capability, and have relevant social interactions.

Besides the need to selectively handle data, the robustness and reliability of a PDS system depends upon the availability of useful data. This can be ensured by pre-fetching data based on data interest inference, improving resource utilization and the availability rate of data to the sensing applications.

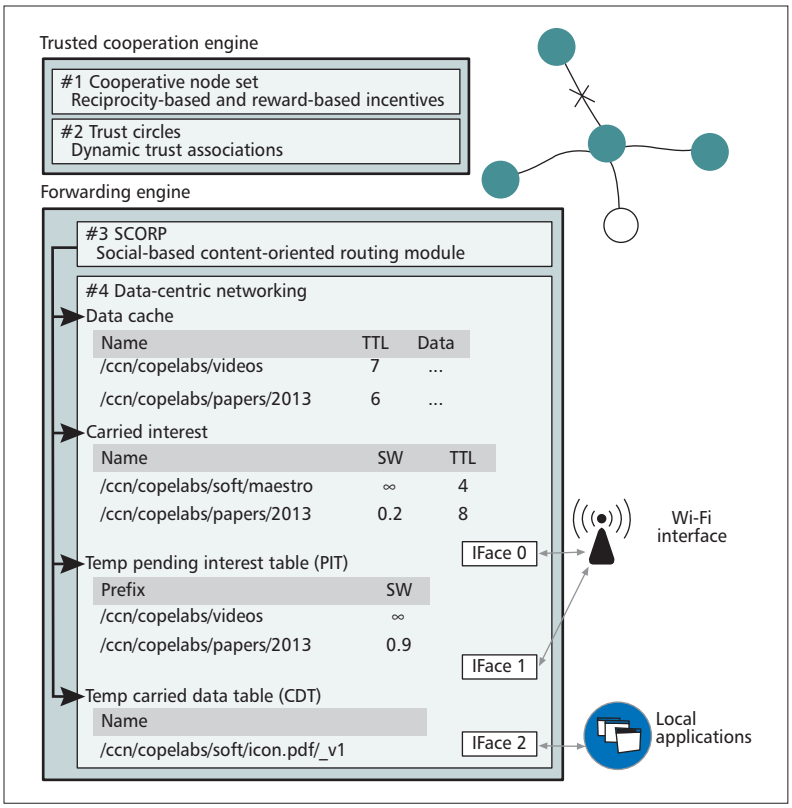


Figure 3. PDS node architecture for mobile citizen sensing.

Finally, there are several frameworks that are centered on the data and could be employed to implement paradigm #4, such as the publish-subscribe Internet routing paradigm (PSIRP), data-oriented network architecture (DONA), named data networking (NDN), content-centric networking (CCN), and network of information (NetInf). Each of these proposals has its own particularities (e.g., employ their own naming scheme) and look into different content-centric aspects (e.g., naming, security, routing) emphasizing a few of these aspects according to the application to which they have been devised.

One can observe that solutions for paradigm #4 are mostly aimed at improving data dissemination in fixed networks (i.e., the Internet at large) with few examples supporting ad hoc, vehicular, and Internet of Things networking. This also makes such solutions not aligned with paradigm #3, since these approaches assume that nodes are always connected. Moreover, features of paradigm #4 such as the ability to increase the quality and value of the data by means of data aggregation mechanisms, and privacy are still novel research issues when it comes to this paradigm [1, 14].

Thus, concerning an MCS scenario, paradigm #4 could potentially be based on a CCN/NDN-like framework given its decentralized feature. While in the other frameworks nodes rely on specific entities (i.e., rendezvous points in PSIRP, name resolution system in NetInf, resolution handlers in DONA) to handle user interests, queries, and responses, CCN/NDN is more straightforward, requiring nodes to just manifest their interests for retrieving content [15].

This feature of CCN/NDN is interesting as it allows easy integration with paradigm #3 (i.e., nodes easily exchange interest lists and desired content based on such lists), and can cope with the dynamism of user behavior (i.e., users want to send/retrieve content anytime/anywhere).

Node Architecture for Pervasive Data Sharing

Following the proposed set of paradigms, design guidelines, and analysis of the functional blocks for each paradigm, Fig. 3 presents the proposed node architecture to deploy pervasive data sharing systems.

Generally speaking, in pervasive networking scenarios all nodes may forward data: nodes should not only be able to collect/consume data, but they should also be able to forward such data/interest requests toward the intended destination and/or source of data.

The proposed PDS node architecture comprises two engines: trusted cooperation and forwarding. The former is responsible for implementing the solutions for paradigms #1 and #2. The trusted cooperation engine informs the current node of which neighboring peers are willing to cooperate (represented by filled circles on the right in Fig. 3) and can be considered part of its trust circles (non-crossed links between them in Fig. 3). As mentioned earlier, cooperative behavior can be encouraged with either reciprocity-based or reward-based incentives, and trust associations between users are built as they interact in the system.

It is worth noting that a node may assume cooperative behavior, but still may not be trusted (filled circle with crossed link in Fig. 3). The PDS system must be able to identify such misbehaving/malicious nodes in order to avoid them.

The forwarding engine is responsible for implementing paradigms #3 and #4. Once the current node is acquainted with its neighboring peers (trusted cooperative users), it may start sharing data. From the social-aware and content-based proposals mentioned earlier, opportunistic communication shall be driven by SCORP. This is due to the fact that it considers not only users' interests, but also how these users are socially connected. It is worth mentioning that SCORP measures the social weight (SW) of a node toward specific interests [9].

In the proposed node architecture, data-centric networking is done based on persistent and temporal data structures, as shown in Fig. 3. Persistent structures are:

- The *data cache* (a.k.a. content store), which holds data carried by the node and is of interest to itself and/or other nodes
- The *carried interest* table, which keeps track of the data interests of the node, as well as its SWs toward interests of other nodes with whom it socially interacts

Generally speaking, the *data cache* remains the same as the CCN/NDN content store, but it limits the amount of stored data. This is because devices have different capabilities, and users may not be willing to share all of their storage on behalf of others. Thus, stored data can be

removed according to the rate of interaction a node has with others who are interested in the carried content. Moreover, since the *carried interest* table includes the interests of the node and those of others, SW is assigned a value of infinity to identify the interests of the current node and facilitate content exchange.

It is important to say that data and interest can be generated by local applications (cf. interface 2 in Fig. 3) or received via a networking interface (cf. interfaces 0 and 1). Also note that in any of the two persistent structures, an entry has a defined time to live (TTL): data TTL defines the time usefulness of data, as tagged by the source; interest TTL defines the time period over which some node is interested in a specific type of data. By erasing carried data and interests based on TTL, we ensure that only useful data is transported, which is a desired scalability feature in large-scale urban sensing scenarios.

The temporary structures are:

- The temporary pending interest table (PIT), which lists the interest carried by a neighboring node
- The temporary carried data table (CDT), which replaces the forwarding information base found in CCN/NDN

The PIT also has an SW field to describe the SW between the neighboring node and specific interests it has come across. PIT keeps track of the data missing from *data cache*, and does not forward an interest request as it would normally do with content-centric approaches. CDT is populated with information about the data that neighboring nodes are currently carrying. The entries in these tables are temporary, since they are erased as soon as the contact with a neighboring node is broken. Finally, both PIT and CDT do not map interest/content to interfaces as would normally happen in content-centric approaches. This is because the PDS node is more concerned with mapping the interest/content to nodes in which it is socially well connected. One can observe that these changes allow our PDS system to be easily built based on well-known networking paradigms to support MCS in large-scale urban scenarios. In order to illustrate the operation of the proposed node architecture, let us exemplify the interaction between two devices, *A* and *B*, from the perspective of device *A* (the same operation is taking place at device *B* concurrently). When devices *A* and *B* meet, device *B* sends two metadata lists concerning:

- The data it currently carries (*data cache*)
- Its interests as well as its SWs toward the interests of other devices it has interacted up to this point (*carried interest*)

These lists populate the CDT and PIT on device *A*, respectively.

Device *A* then uses the recently updated PIT to quickly determine whether device *B* is interested, and/or if device *B* is socially well connected to other devices with interest in the data carried by device *A*.

For every piece of information in its data cache, device *A* shall forward actual data if it is not in its updated CDT (i.e., missing from device *B*'s data cache) and either:

- Device *B* is interested in it (*/ccn/copelabs/videos* is forwarded since SW = ∞, cf. PIT in Fig. 3).
- Device *B*'s SW toward that specific interest (obtained from the PIT) is greater than the SW of device *A* toward such interest as specified in its own *carried interest* list (*/ccn/copelabs/papers/2013* is forwarded since device *B*'s SW = 0.9 > device *A*'s SW = 0.2, cf. PIT and *carried interest* in Fig. 3).

It is worth mentioning that the number of interests in a PDS system can be significantly high given its granularity (e.g., one user likes cars, while another likes a specific model from a specific manufacturer), which may affect scalability. Thus, the PDS node architecture can employ mechanisms to mitigate such an effect by creating metadata lists that include only SW to socially relevant interests and with useful TTL.

Conclusions

Today, we are surrounded by a panoply of devices that produce a massive amount of data and have a diversified set of communications (i.e., Bluetooth, Wi-Fi, 3G), and sensing (i.e., activity, proximity, sound, location, human body signals) capabilities. Based on such properties, mobile devices may support mobile citizen sensing applications, which can have an impact on different sectors of society.

Due to its pervasive nature, we cannot assume that the control of MCS systems can be based on a networking system where data is transmitted over perfect communication channels. Instead, the communication between mobile devices should be supported by a data sharing system able to ensure high data exchange probability, low latency, and low cost, even when facing challenging networking scenarios.

The creation of MCS systems, based on pervasive data sharing, requires a constructive engineering approach. This article gives an introduction to networking requirements of MCS systems and proposes four design paradigms for PDS systems. These paradigms are derived by looking at different user-centric and data-centric networking approaches, and extracting common features that contribute to the efficiency of pervasive data sharing in large-scale sensing systems. We believe that these paradigms represent basic building blocks, so we suggest a process for the design of PDS solutions, and show that PDS systems can easily be devised based on the combination of data-centric networking and opportunistic communication among trusted and cooperative wireless devices. We hope that these contributions will stimulate further research to allow the deployment of MCS systems.

Today, we are surrounded by a panoply of devices that produce a massive amount of data and have a diversified set of communications and sensing capabilities. Based on such properties, mobile devices may support mobile citizen sensing applications, which can have an impact in different sectors of society.

Acknowledgment

The research leading to these results has received funding from the European Union (EU) Horizon 2020 research and innovation programme under grant agreement No 645124 (Action full title: Universal, mobile-centric and opportunistic communications architecture, Action Acronym: UMOBILE). This article reflects only the authors' views, and the Community is not liable for any use that may be made of the information contained therein. Acknowledgments are also due to the CitySense project from COPELABS.

These paradigms represent basic building blocks, and so we suggest a process for the design of pervasive data sharing solutions, and show that PDS systems can be easily devised based on the combination of data-centric networking and opportunistic communication among trusted and cooperative wireless devices.

References

[1] J. Crowcroft *et al.*, "D.2.1. End-User Requirements Report," UMOBILE Project Deliv., June 2015.
[2] W. Moreira and P. Mendes, "Social-Aware Opportunistic Routing: The New Trend," *Routing in Opportunistic Networks*, I. Woungang *et al.*, Eds., Springer Verlag, May 2013.
[3] V. Jacobson *et al.*, "Networking Named Content," *Proc. CoNEXT*, Dec. 2009, pp. 1–12.
[4] L. Zhang *et al.*, "Named Data Networking," *SIGCOMM Comp. Commun. Rev.*, vol. 44, no. 3, July 2014, pp. 66–73.
[5] C. Ballester Lafuente *et al.*, "Trust Management in Uloop," *User-Centric Networking, ser. Lecture Notes in Social Networks*, A. Aldini and A. Bogliolo, Eds., Springer, 2014, pp. 107–19.
[6] P. Mendes *et al.*, "Cooperative Networking in User-Centric Wireless Networks," *User-Centric Networking*, ser. Lecture Notes in Social Networks, A. Aldini and A. Bogliolo, Eds., Springer, 2014, pp. 135–57.
[7] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble Rap: Social-based Forwarding in Delay-Tolerant Networks," *IEEE Trans. Mobile Computing*, vol. 10, no. 11, Nov. 2011, pp. 1576–89.
[8] W. Moreira, P. Mendes, and S. Sargento, "Opportunistic Routing Based on Daily Routines," *Proc. WoWMoM*, June 2012, pp. 1–6.
[9] —, "Social-Aware Opportunistic Routing Protocol Based on Users Interactions and Interests," *Ad Hoc Networks*, ser. Lecture Notes of the Inst. Comp. Sciences, Social Informatics, and Telecommun. Engineering, M. H. Sherif *et al.*, Eds., Springer, 2014, vol. 129, pp. 100–15.
[10] W. Moreira and P. Mendes, "Social-Aware Forwarding in Opportunistic Wireless Networks: Content Awareness or Obliviousness?" *Proc. WoWMoM*, June 2014, pp. 1–6.
[11] M. Shin *et al.*, "Anonysense: A System for Anonymous Opportunistic Sensing," *Pervasive and Mobile Computing*, vol. 7, no. 1, 2011, pp. 16–30.
[12] A. Bogliolo *et al.*, "Virtual Currency and Reputation-Based Cooperation Incentives in User-Centric Networks," *2012 8th Int'l. Wireless Commun. and Mobile Computing Conf. (IWCMC)*, Aug 2012, pp. 895–900.
[13] W. Moreira and P. Mendes, "Impact of Human Behavior on Social Opportunistic Forwarding," *Ad Hoc Networks*, vol. 25, Part B, 2015, pp. 293–302; New Research Challenges in Mobile, Opportunistic and Delay-Tolerant NetworksEnergy-Aware Data Centers: Architecture, Infrastructure, and Communication.
[14] N. Fotiou *et al.*, "A Framework for Privacy Analysis of icn Architectures," *Privacy Technologies and Policy*, ser. Lecture Notes in Computer Science, B. Preneel and D. Ikonomou, Eds. Springer, 2014, vol. 8450, pp. 117–32.
[15] G. Xylomenos *et al.*, "A Survey of Information-Centric Networking Research," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 2, May 2014, pp. 1024–49.

Biographies

Waldir Moreira (waldir.junior@ulusofona.pt) holds Bachelor's (2005, UofL/UNAMA, Canada/Brazil) and Master's (2008, UFPA, Brazil) degrees in computer science, and a Ph.D. degree (2014, UM/UA/UP, Portugal) in telecommunications. His research career started at GERCOM/UFPA in 2006, passing by INESC-Porto in 2009. Currently, he is with COPELABS, working on different national and European projects, and is an invited professor at Universidade Lusófona (ULHT/ULP). His research interests and publications are related to wireless ad hoc, mesh, social-aware, cooperative, opportunistic, and information-centric networking and routing.

Paulo Mendes is vice-director of the Cognition and People Centric Computing Research Center (COPELABS). His interests are in the field of self-organized networked systems and pervasive data sharing. In 2004 he got his Ph.D. from the University of Coimbra, having developed his thesis work at Columbia University. After that, he started his research career at NTT DoCoMo Research Labs. He has more than 80 scientific articles, and 14 international patents protect his inventions.

CALL FOR PARTICIPATION

18TH CONFERENCE OF OPEN INNOVATIONS ASSOCIATION FRUCT
SAINT PETERSBURG, RUSSIA, 18–22 APRIL 2016

OVERVIEW

FRUCT is the largest regional cooperation framework forum of open innovations between academia and industry. FRUCT conferences are attended by the representatives of more than 25 FRUCT member universities from Russia, Finland, Denmark, India, Italy, and Ukraine; industrial experts from EMC2, Intel, Nokia, and Skolkovo; and a number of guests from other companies and universities.

The conference is an R&D forum for the most active students, academic experts, industrial researchers, and influential representatives of business and government. The conference invites world-class academic and industrial researchers to give lectures on the most relevant topics, and provides an opportunity for student teams to present progress and results of their R&D projects, meet new interesting people, and form new R&D teams. The conference program consists of three to five intensive (half or full day) training sessions on the most promising technologies, plus three days of the main conference.

We warmly welcome all university research teams to participate in the conference, present your research, and join the FRUCT Association. IEEE members and representatives of Russian and Finnish universities are entitled to large discounts. Registration to the conference is open at http://www.fruct.org/conference18.

LIST OF CONFERENCE TOPICS

- Location-based services, navigation, logistics management, e-tourism solutions
- Mobile healthcare, e-health solutions, well being, fitness, automated diagnostics
- IoT, smart spaces, future services: proactivity, context analysis, data mining, and big data services
- Cross-platform software, innovative mobile services, new approaches to application design, innovative UX
- Smart systems and embedded networks
- Energy-efficient design and peripherals integration
- Mobile security, personal and business privacy
- Modern network architectures, air interfaces, and protocols; emerging wireless technologies
- Mobile multimedia, video services, and solutions
- Communications systems integration and modeling

TYPES OF SUBMISSION

Depending on the type and maturity level, please submit your work into one of the following three categories:

1. Full paper (minimum 6 pages, maximum 12 pages)
2. Extended abstract (minimum 200 words, maximum 5 pages)
   Submission deadline: 4 December 2015
   Notification of acceptance: 5 February 2016
   Camera-ready deadline: 4 March 2016
3. Poster/Demo proposal: optionally you can submit executive summary (minimum 200 words, maximum 3 pages)
   Submission deadline: 2 April 2016

Please follow paper templates that you can find at http://www.fruct.org/cfp.

FEES

RG-01 - FRUCT/ComSoc Members, before April 1, 2016 - FREE
RG-02 - Discounted registration for IEEE Members, before April 1, 2016 — €80
RG-04 - Discounted full registration, before April 1, 2016 — €200
RG-05 - Late registration for IEEE Members — €100
RG-07 - Late full registration — €250

PUBLICATION

All submitted full papers will be peer reviewed by the technical committee. Accepted full papers and extended abstracts will be published in the Proceedings of FRUCT Conference (ISSN 2305-7254). We wait confirmation that all accepted full papers will be published in IEEE Xplore (Scopus indexed), which has been a regular practice for the past five conferences of the series. Selected papers will be recommended for CPCI indexing (Web of Science) and to journals.

CONTACTS

Paper templates, conference news, and other relevant details are available at http://www.fruct.org/conference18. If you have some questions that are not covered at the conference web page, feel free to send email to info@fruct.org.

INTEGRATED CIRCUITS FOR COMMUNICATIONS

Charles Chien

Zhiwei Xu

In this issue of Integrated Circuits for Communications, we have selected two papers that mark recent progress in the communications semiconductor industry that is enabling spectrum sensing and near-field communications.

Communication system-on-chip (SoC) technology has advanced significantly over the past 10 years resulting in highly integrated radio SoC chips for standards such as wireless Wi-Fi, GPS, Bluetooth (BT), 4G cellular, digital TV (DTV), and cable modems. While most radio SoCs are narrowband (e.g., Wi-Fi, GPS, BT, and cellular), future trends favor broadband software defined radio (SDR) and cognitive radio (CR). SDR promises to enable low-cost multi-standard radios, while CR can greatly enhance spectral reuse via spectral sensing. CR can be exploited to improve capacity in commercial systems. IEEE 802.11af and IEEE 802.22 are two emerging commercial standards based on CR that improve the system capacity by dynamically utilizing unused TV channels (i.e., TV white space) in terrestrial broadcast systems. Advanced radio concepts based on CR require spectrum sensing technology to robustly detect and identify existing transmissions and thereby be able to utilize the unused spectrum effectively without interfering with existing communication links. However, while theoretical solutions to such a sensing device abound in the literature, a practical low-cost and low-power implementation has been illusive to attain.

In the article "Wideband Blind Signal Classification on a Battery Budget," the authors introduce an SoC implementation of a broadband spectrum sensor in low-cost 40 nm complementary metal oxide semiconductor (CMOS) technology. The authors give an in-depth overview from the system down to circuit implementation. Particularly interesting is their time-interleaved analog FFT/ADC architecture that enables wideband operation up to 500 MHz, as well as, their highly efficient DSP architecture that enables greater than 12 GOPS/mW energy efficiency in the blind signal identification which allows the sensor to detect and process up to 32 concurrent non-overlapping signals. The integrated sensor chip consumes around 90 mW, and the core circuit occupies 1.7 mm^2. While the chip's power consumption will still need to come down for primetime use in battery operated devices, it marks a significant step toward that goal.

The second topic covered in our series deals with near-field communications. Since Charles Walton filed the first patent in 1983 on the radio frequency identifier (RFID), many near-field communication (NFC) systems have been developed to support data communication between two units in close proximity (~10 cm). However, until recently adoption of NFC was slow. Today, NFC is enjoying newfound popularity due to its integration into smartphones to support secure wireless transactions. Further riding the wave of growing demand for the Internet of Things (IoT), NFC represents a potential technology to realize point-to-point connection between smartphones and wearable devices, as well as other leaf nodes in the IoT universe. Two significant challenges still face NFC before it can be widely adopted for IoT applications. The first deals with low cost, and the second deals with flexible form factor to enable integration in wearable products.

In the article "Flexible Thin-Film NFC Tags," the authors introduce a potential solution to meet both of these challenges by using thin-film transistor technologies instead of CMOS. The authors first briefly review and compare three thin-film transistor (TFT) technologies: low temperature polycrystalline silicon (LTPS) TFTs, metal oxide TFTs, and amorphous metal oxide TFTs. The authors then evaluate four different inverter topologies suitable for NFC. Several small-scale chips using these topologies have been fabricated, featuring ultra-low power consumption ($< 1\ \mu W$) and demonstrating feasibility for application in NFC tags. Most importantly, these tags can potentially be integrated onto different materials and open the door to future sensor fusion, IoT, and indoor positioning applications.

We would like to take this opportunity to thank all the authors as well as reviewers for their contributions to this Series. Future issues of this series will continue to cover circuit technologies that are enabling new emerging communication systems. If readers are interested in submitting an article to this Series, please send your paper title and an abstract to either of the Series Editors for consideration.

BIOGRAPHIES

CHARLES CHIEN (charles.chien@creonexsystems.com) is the president and CTO of CreoNex Systems, which focuses on technology development for next generation communication systems. Previously, he held key roles at Conexant Systems, SST Communications, and Rockwell. He was also an assistant adjunct professor at the University of California Los Angeles. His interests focus mainly on the design of system on-chip solutions for communication systems. He has published in various journals and conferences, and has authored a book, *Digital Radio Systems on a Chip*.

ZHIWEI XU (xuzhw@yahoo.com) is currently with Zhejiang University, working on cognitive radios, high-speed ADC, and mmWave ICs. He has held industry positions with SST Communications, Conexant Systems, NXP, and HRL Laboratories, where he developed wireless LAN and SoC solutions for proprietary wireless multimedia systems, CMOS cellular transceiver, Multimedia over Cable systems, TV tuners, software defined radios, and analog VLSI. He has published in various journals and conferences, three book chapters, and 12 granted patents.

Wideband Blind Signal Classification on a Battery Budget

Ramesh Harjani, Danijela Cabric, Dejan Markovic, Brian M. Sadler, Rakesh K. Palani, Anindya Saha, Hundo Shin, Eric Rebeiz, Sina Basir-Kazeruni, and Fang-Li Yuan

ABSTRACT

A wideband signal sensor is an essential component to enable cognitive radio and dynamic spectrum access techniques, providing real-time detection and modulation classification in a wideband environment of interest. The problem is challenging, requiring a processing suite incorporating detection, estimation, and classification, with stringent power objectives to enable widespread use in untethered battery powered devices. This article provides an overview of an integrated system-on-chip extremely low-power solution, including a wideband mixed-signal front-end, an algorithm suite that incorporates a blind hierarchical modulation classifier, and an ASIC implementation that employs dynamic voltage-frequency scaling and parallel processing that achieves measured energy efficiency ranging between 11.9 GOPS/mW and 13.6 GOPS/mW for full channel feature extraction, resulting in power consumption of 20.1~22.6 mW depending on the number of signals and signal bandwidth. The system bandwidth is selectable at 5, 50, and 500 MHz; in the 500 MHz case an efficient analog 8-point FFT channelizer relaxes the A/D requirement. The sensor can blindly detect and process up to 32 concurrent non-overlapping signals, with a variety of signal characteristics including single- vs. multi-carrier discrimination, carrier detection and estimation, and modulation classification.

Ramesh Harjani, Rakesh K. Palani, Anindya Saha, and Hundo Shin are with the University of Minnesota.

Danijela Cabric, Dejan Markovic, Eric Rebeiz, Sina Basir-Kazeruni, and Fang-Li Yuan are with the University of California, Los Angeles.

Brian M. Sadler is with the Army Research Laboratory.

INTRODUCTION

Blind signal classification is an important component of emerging commercial and military wireless networks, supporting cognitive radio and dynamic spectrum access. This article focuses on a joint project undertaken by the University of Minnesota, University of California, Los Angeles, and the Army Research Laboratory to design a low-power wideband signal classifier integrated circuit (IC), as part of the Defense Advanced Research Projects Agency (DARPA) CLASIC program. The project goals, shown in in Fig. 1, are to blindly identify the carrier frequency, modulation scheme, constellation size, and multiple access scheme in a bandwidth up to 500 MHz; and to do so in an extremely tight power budget.

In particular, we describe the critical algorithms for efficient blind signal classification and the necessary computationally efficient hardware developed to accomplish this. Signal classification can be thought of as the precursor to signal demodulation. For both single-carrier and multi-carrier waveforms, demodulation requires time and frequency synchronization, and relies on knowledge of the physical layer characteristics of the waveform. Although synchronization is a standard step in any communication system such as Long Term Evolution (LTE) or WiFi, the underlying signal characteristics such as synchronization pilot sequences, in addition to more basic parameters such as the signal's bandwidth, are assumed to be known to the intended receiver. These known signal characteristics are at the heart of modern synchronization strategies. Blind synchronization, on the other hand, requires a more robust solution with limited prior information.

The focus on low-energy operation is critical for battery-driven devices, and must be addressed in a comprehensive design strategy. As energy consumption grows with computational complexity, the design problem can be formulated as follows: *how can a wideband radio sensor detect, synchronize, and classify a variety of physical waveforms while minimizing computational complexity?* This leads to the natural question of which parameters are needed for signal classification, and what are the required estimation accuracies for proper signal classification and demodulation?

Modulation classification is highly dependent on the particular waveform, and the specific assumptions of what is known or not known a priori. Modulation classifiers can be roughly split into two categories: likelihood-based and feature-based classifiers. As their name indicates, likelihood classifiers [1] utilize a likelihood function estimated from the received signal, and the classification decision is made by comparing the likelihood ratio against a threshold. The resulting solution is optimal in the Bayesian sense as it minimizes the probability of false classification. However, these classifiers suffer from a range of

0163-6804/15/$25.00

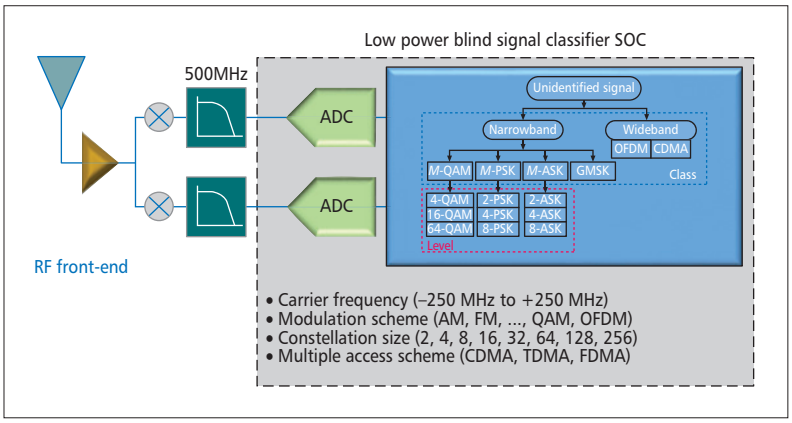


Figure 1. Low power blind signal classifier IC system goals.

practical challenges. First, the optimal solution is typically computationally expensive, which is unappealing for low-energy applications. Second, likelihood-based classifiers require the distribution of the received waveform as a function of the unknown parameters of the received signal, and this is often unavailable in the context of blind signal classification. On the other hand, feature-based classifiers [2, 3, 4] rely on estimation of features of the received waveform, and a classification decision is made based on the observed features. Although feature-based classifiers are not necessarily globally optimal, their computational complexity is typically much lower than likelihood-based classifiers and, when designed properly, can achieve near-optimal classification performance [5].

Survey of Modulation Classifiers

In this section, we provide an overview of different blind modulation classification techniques.

Likelihood-Based Modulation Classification

The probability density function (PDF) of the received waveform conditioned on the waveform's modulation is at the heart of likelihood-based modulation classification [1]. Depending on the model chosen for the unknown parameters of the received waveform, various likelihood-based classifiers can be derived such as the generalized likelihood ratio test (GLRT), the average likelihood ratio test (ALRT), and the hybrid likelihood ratio test (HLRT). Unknown parameters may be modeled as random variables with known distributions, or as deterministic unknowns. In both cases, the method generally requires the distribution of the received waveform as a function of all the unknown parameters, and so does not readily handle cases of unknown waveform characteristics. Likelihood-based classifiers also generally require multi-dimensional maximization or integration, and thus are not typically well suited for real-time low-energy applications due to high computational complexity.

Feature-Based Modulation Classification

Feature-based classifiers [4, 6, 7] extract one or more features of the received waveform and use these features in modulation classification algorithms. Feature extraction is linked with sampling rate, such as oversampling or sampling that is synchronized with the digital symbol rate. A wide variety of feature-based classifiers operate on the received information symbols under various impairments. The features range from the symbols' instantaneous amplitude and phase to statistical metrics such as hierarchical cumulant-based classifiers [4]. Distribution-distance-based classifiers that rely on a goodness-of-fit (GoF) test statistic have been suggested as an alternative to the cumulant-based level classification, and these have been shown to require fewer samples to achieve accurate modulation level classification [2, 8].

Both cumulant-based and GoF feature-based classifiers require accurate time and frequency synchronization, which cannot be achieved without first estimating the waveform carrier frequency and symbol rate. Other feature-based classification algorithms operate on over-sampled signals. These include correlation-based and cyclostationary-based classifiers that rely on second-order statistical features of the received waveform [3, 9]. The temporal structure of the received waveform can also be exploited. Temporal features are typically added by design to aid the underlying communication scheme. For example, orthogonal frequency-division multiplexing (OFDM) typically employs a cyclic prefix (CP) prepended to every OFDM symbol, and these are collected into frames. Single-carrier block coded modulations also have temporal features as they are modulated on a block-basis. The cyclic autocorrelation (CAC) function is a statistic that can be estimated to detect the presence of cyclostationary features [3] in an oversampled received waveform. Different modulation classes can be differentiated via a cyclostationarity test because their CACs possess cyclic peaks at different cyclic frequencies α, which are a function of the symbol rate ($1/T$) and the carrier frequency (f_c).

Proposed Hierarchical Modulation Classifier

To achieve our goals of blind classification and low power consumption, we have designed an algorithmic hierarchical classification tree. The design hierarchy is based on both the level of prior information needed for the particular decision and the overall computational complexity. Blocks that do not require prior information about the signal being classified are processed first [7]. This design methodology dictates the ordering of the classification tree, as shown in Fig. 2. The processing is carried out over time, with increasing total energy consumption, and can be terminated at any step in order to save energy and comply with the desired level of information reported by the system. Next, we describe the classification tree, including the optimization needed to meet the given accuracy and energy requirements.

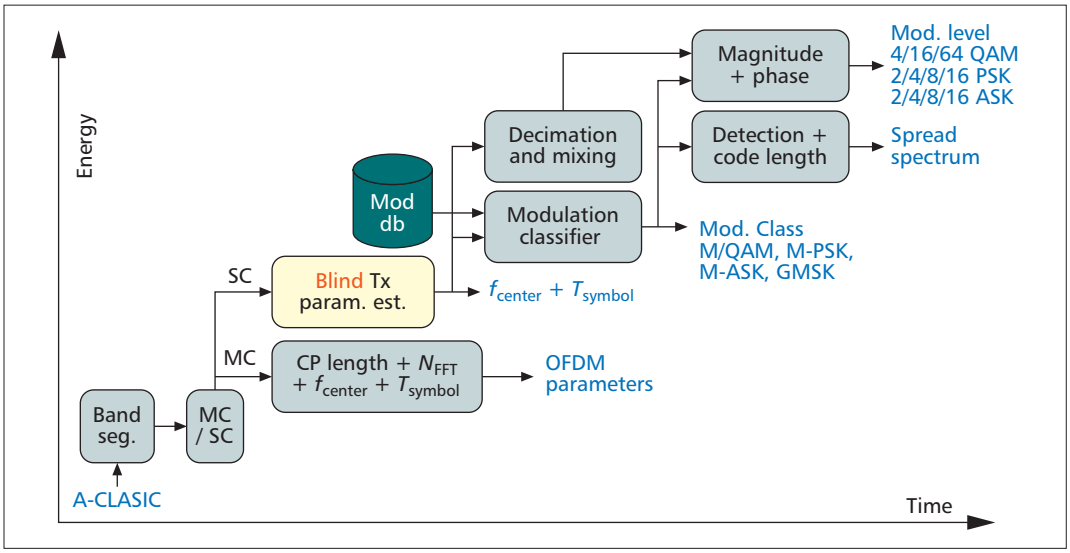


Figure 2. Top-level breakdown of the hierarchical processing kernels of the proposed processor.

After positively classifying an OFDM signal, its CP length, frame length, number of subcarriers, and carrier frequency can then be blindly estimated. OFDM parameter estimation can be performed by exploiting the signal's cross-correlation properties.

We start the overall process with frequency domain detection and segmentation to isolate signal occupied bands. Our system has a 12.5 kHz frequency bin resolution. We then start the classification tree with the multi-carrier (MC) vs. single-carrier (SC) decision. Although OFDM waveforms could be detected by exploiting their temporal structure, there are two main challenges to this approach, the first being that the periodicities may not be readily detectable in the presence of large frequency and time offsets, which are typically present in blind classification applications. Second, single-carrier block-coded modulations also exhibit periodicities of a similar nature and can therefore cause signal misclassification. An alternative approach is to exploit the Gaussianity of multi-carrier waveforms. As a result of the inverse fast Fourier transform (IFFT) operation, an OFDM waveform is composed of a large number of independent and identically distributed random variables. Via a central limit theorem effect, the amplitude of multi-carrier waveforms is well approximated as a Gaussian random process. The C_{42}-based MC classifier is based on the fourth-order cumulant, which is a form of Gaussianity test that is robust to frequency and timing offsets. After positively classifying an OFDM signal, its CP length, frame length, number of subcarriers, and carrier frequency can then be blindly estimated. OFDM parameter estimation can be performed by exploiting the signal's cross-correlation properties.

As a result of the 12.5 kHz resolution of the band segmentation, coarse estimates of the symbol rate and carrier frequency obtained after the band segmentation have estimation errors on the order of thousands of parts per million. As shown in [10], the features used for modulation-type classification degrade under large estimation errors of the cyclic frequencies. To address this issue, preprocessing is performed to yield fine estimates of the transmit parameters by utilizing cyclostationary-based symbol rate and carrier frequency estimation blocks.

Symbol Rate and Carrier Frequency Estimation

We note that all SC modulation classes considered in this work exhibit a cyclostationary feature at cyclic frequency $\alpha = 1/T$. Therefore, detecting the presence of this cyclostationary feature inherently provides an estimate of the symbol rate. The coarse estimate of the symbol rate from the band segmentation can be used to set a search window $\mathcal{W}_T$, within which the cyclic peak at the symbol rate will be located. The detection of the cyclostationary feature at $1/T$ is therefore obtained by solving the following optimization problem [7]:

$$\max_{\alpha_i \in \mathcal{W}_T} \left| \sum_{n=0}^{N_T-1} |x[n]|^2 e^{-j2\pi\alpha_i n T_s} \right|, \quad (1)$$

where N_T is the number of samples per CAC computation used to estimate the signal's symbol rate.

Given that not all classes have a cyclostationary feature related to their carrier frequency, the CAC cannot be directly used to estimate it. Estimation of the carrier frequency can be performed by detecting the cyclic feature at $\alpha = 4f_c$ after squaring the incoming samples. We denote the search window by $\mathcal{W}_f$ within which the cyclic peak at $4f_c$ occurs. The estimation is therefore obtained by solving the following optimization problem:

$$\max_{\alpha_i \in \mathcal{W}_f} \left| \sum_{n=0}^{N_f-1} x[n]^4 e^{-j2\pi\alpha_i n T_s} \right|, \quad (2)$$

where N_f is the total number of samples per CAC computation used to estimate the signal's carrier frequency. In both Eqs. 1 and 2, the optimization is carried out by discretizing the cyclic feature variable to the desired level of accuracy. Note that with increasing number of samples over which the CAC is computed, the noise is suppressed, and the features of interest become more prominent. Performance therefore depends on the available sample size and received signal SNR.

The rcK algorithm computes the CDF of the received symbols at points that achieve maximum deviation among the different theoretical CDFs under comparison test. By doing so, the rcK algorithm does not require estimating the entire CDF, but still performs better than the Kolmogorov-Smirnov test [8] and achieves near-ML classification performance.

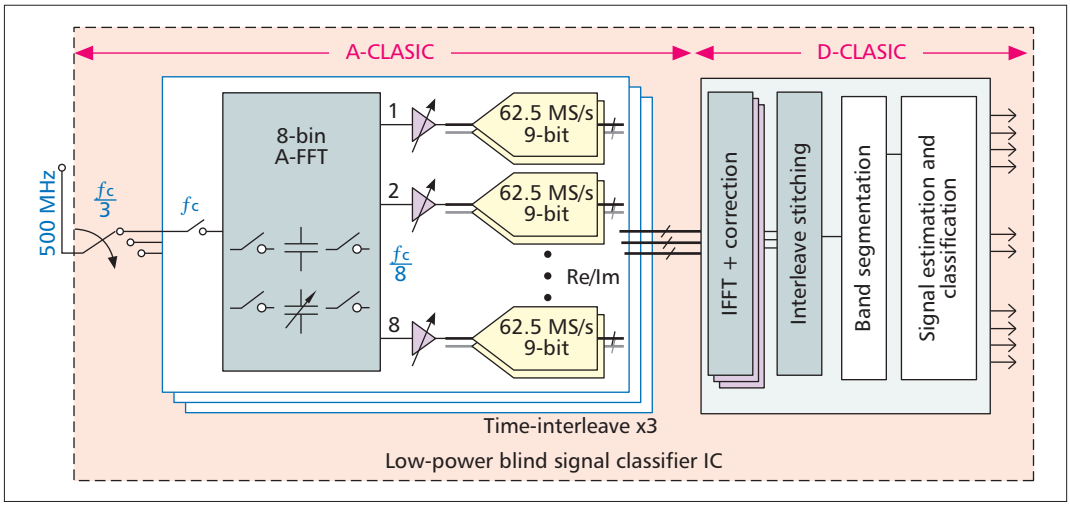


Figure 3. Overall architecture of the blind signal classifier SOC, a 500 MHz 8-channel digitizer with low-power digital signal estimation and classifier.

Modulation-Type Classification

After estimating the signal parameters, the next step is to compute the CAC at cyclic frequencies within the union of possible cyclostationary features, resulting in a a six-dimensional feature vector $\mathbf{F}$ obtained by computing the CAC at lag $\nu = 0$ and the cyclic frequencies $[1/T, 2f_c - 1/T, 2f_c - 1/2T, 2f_c, 2f_c + 1/2T, 2f_c + 1/T]$ [11]. Because each element in the feature vector $\mathbf{F}$ is proportional to the received signal power, we normalize the feature vector to unit power, and compare the normalized feature vector $\overline{\mathbf{F}}$ to asymptotic normalized feature vectors $\overline{\mathbf{V}}_i$, $i \in [1, 2, 3]$ for each of the classes considered. For example, the normalized asymptotic feature vector for M-quadrature amplitude modulation (QAM)/M-phase shift keying (PSK) signals is $\overline{\mathbf{V}}_1 = [1, 0, 0, 0, 0, 0]$ as only one cyclic feature is present at the symbol rate. The resulting normalized feature vector is compared to each feature vector $\overline{\mathbf{V}}_i$, and the classifier picks the modulation class $\hat{C}$ with a feature vector closest in the least square sense [11], that is, $\hat{C} = \arg \min_{i \in [1,2,3]} \| \overline{\mathbf{F}} - \overline{\mathbf{V}}_i \|^2$, where $i \in [1, 2, 3]$ covers the exhibited cyclostationary features by the considered modulation classes. In contrast to the preprocessing estimation steps described earlier, the only degree of freedom in the design of the modulation type classifier is the number of samples N_c required to compute each of the six CACs that form the feature vector. Given the SNR of the received signal and the estimation accuracies of the preprocessing steps, N_c is chosen accordingly to meet the desired classification probability. The six CACs are computed sequentially to enable a high degree of digital hardware reuse.

Modulation-Level and Spread Spectrum Classification

After estimating the modulation type and obtaining fine estimates of the symbol rate and carrier frequency, the waveform is filtered and decimated to sample the underlying information symbols. As part of the modulation level classification, we adopt a reduced complexity Kuiper (rcK) GoF test, as it was shown to be more energy efficient than a cumulant-based approach, especially when choosing among different modulation levels within the same modulation type [2]. The rcK algorithm computes the CDF of the received symbols at points that achieve maximum deviation among the different theoretical CDFs under comparison test. By doing so, the rcK algorithm does not require estimating the entire CDF, but still performs better than the Kolmogorov-Smirnov test [8] and achieves near-ML classification performance. For direct-sequence spread-spectrum signal classification, we employ the variance of the autocorrelation at various lags to test for the presence of an underlying spreading code [7].

Hardware Implementation

The overall system architecture is shown in Fig. 3. First, in A-CLASIC the signal feeds an 8-channel analog complex-valued FFT channelizer, with 8 × 62.5 MHz channels for a total 500 MHz bandwidth. The analog FFT coefficients are amplified to compensate for the reduction in peak amplitude due to the reduced number of signals in band, and then digitized via eight I/Q ADCs. These digital samples are then fed to the D-CLASIC module. An IFFT and correction step simultaneously performs calibration and the IFFT operation to yield full rate time domain samples. The full band signal is then band segmented in up to 8192 sub-bands, and signal detections are fed to the signal estimation and classification algorithms. Further details on the channelized wideband ADC (A-CLASIC), and the digital band segmentation and signal classifier (D-CLASIC), are given next.

Wideband High-Resolution ADC: A-CLASIC

The performance of wideband analog-to-digital converters (ADCs) is limited by two primary factors. First, sampling clock jitter becomes dominant as the sampling speed increases. Second, in a general setting the number of in-band signals -

increases proportionately with the bandwidth, which yields an increase in the peak-to-average power ratio (PAPR) in the time domain; therefore, a higher ADC dynamic range is needed to adequately capture the signals. The increased PAPR is often associated with multi-carrier systems; the PAPR increases with the number of carriers.

To illustrate the effects of jitter noise, consider a sinusoidal input signal, $y = A\sin(\omega t)$, which is sampled with a clock the jitter variance of which is given by ΔT_s. The additional noise introduced can be approximated by $\Delta y = A\omega\cos(\omega t)\Delta T_s$. We note that the impact of the jitter noise is proportional to the derivative of the input signal frequency, and the impact is highest near the zero crossings of the original sinusoidal signal where the slope is maximum. The jitter impact can be mitigated by reducing either the clock variance or the input frequency before sampling. It is difficult to reduce the clock jitter beyond a certain limit, and currently the on-chip practical jitter limit is roughly 100–200 fs. Alternately, we might break up the wideband input signal into several narrow bands using mixers and filters before sampling, referred to as frequency interleaving. Now each sampler sees lower frequency signals, reducing the impact of jitter. However, this approach introduces phase noise arising from the local oscillators feeding each mixer for down-conversion (an effect similar to jitter that we wish to avoid) and mixer spurs.

Channelization (frequency interleaving) is also desirable to reduce the PAPR, because each sub-channel will have fewer signals and therefore lower PAPR, which relaxes the dynamic range requirement of the sub-channel ADCs. Each sub-channel will have its own variable gain amplifier (VGA) to fill the dynamic range of the ADC, with lower PAPR implying a higher adaptive gain requirement for the VGAs.

Frequency Interleaving: To deal with the above issues, we introduce an analog frequency interleaving architecture that avoids the use of mixers and down-conversion of each sub-channel. The design consists of a discrete time (analog) linear phase filter to channelize the wideband signal, followed by a VGA and ADC in each sub-channel. A gain correction is applied in the digital domain during reconstruction of the full rate signal sample stream. Using linear phase filtering simplifies the reconstruction process. The overall process of digitizing using frequency interleaving is generally a hybrid filter bank ADC, where the channelization (or analysis filter) is followed by reconstruction (or synthesis filter).

The M channel ADCs add quantization noise. With orthogonal channels, the noise power adds during reconstruction leading to factor of M increase in the quantization noise when the VGA gains are set to unity. To maintain the same overall SNR as a single wideband ADC the individual VGA gains should be set to $/\overline{M}\cdot$ This is the minimum gain required in the VGA to suppress the overall quantization noise due to channelization. However, recall that the filtering process has reduced the number of signals in each channel, hence reducing the per channel signal amplitude. Thus, any increase in the VGA gain results in a reduction of the dynamic range requirement for the ADCs.

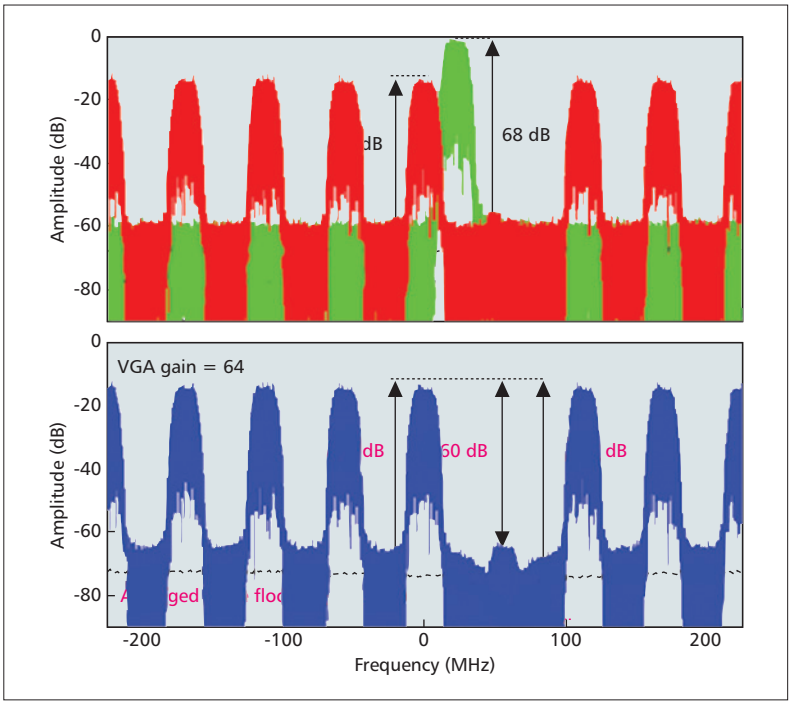


Figure 4. Simulation comparing; a) conventional single 8-bit 450 MS/s ADC; b) our channelized architecture.

Analog-FFT-Based Analysis Filter Implementation: For our hardware filter bank implementation, a decimation-in-time 8-point A-FFT was designed using charge reuse techniques for low power consumption [12]. In this structure, signal processing is performed via passive operations (charge sharing, charge stealing, and wire swapping). For the 8-point A-FFT in a prototype, the sampling frequency of each channel after the A-FFT is decimated by 8 so as to fold each of the channels to DC, which enables use of low-speed VGAs and ADCs. The A-FFT is followed by a bank of VGAs to amplify the channelized signals to the maximum signal range of the ADC. This operation enables even small signals to be digitized with the full ADC dynamic range without worrying about signal saturation that may have resulted from a large signal. After A/D conversion, the amplitude of each signal is recovered by the inverse VGA bank (I-VGA), that is, if the gain in the channel is large due to the presence of only small signals, the inverse gain at the I-VGA is also large such that the quantization noise is reduced. After the I-VGA, signals from all the channels are combined via a digital IFFT, yielding a full rate sample stream. As a result of the IFFT operation, the quantization noise in each channel is Sinc-function shaped and shifted to the corresponding band at the full rate, f_s.

Figures 4a and 4b show simulation results for a single ADC and an 8-channel hybrid filter bank ADC each with 20-MHz-wide 16-QAM modulation signals. The difference in amplitude between the largest and smallest signal is 50 dB.

For both the single ADC and the A-FFT the input peak-to-peak amplitude is 1 V. The simulation was performed with 8-bit ADCs and with VGAs having a maximum gain of 64. The output spectrum for a single modulated signal is shown in green, and the output spectrum for a multi-signal input with seven large signals and one small signal in red, all using a single ADC. For the single modulated signal, the signal-to-noise floor ratio is 68 dB, but for the multi-band signal it decreases to 55 dB due to increased PAPR. Additionally, the small signal in the sixth channel is completely lost in this case. In Fig. 5b we note that after A/D conversion with the proposed channelization scheme, the average noise floor for both the large and small signals is lower than for the single ADC. In particular, the noise floor close to the large signal is 60 dB below the peak signal amplitude resulting in a 5 dB SNR improvement. For the small signal, the noise floor is 73 dB below the peak signal amplitude, resulting in an 18 dB SNR improvement.

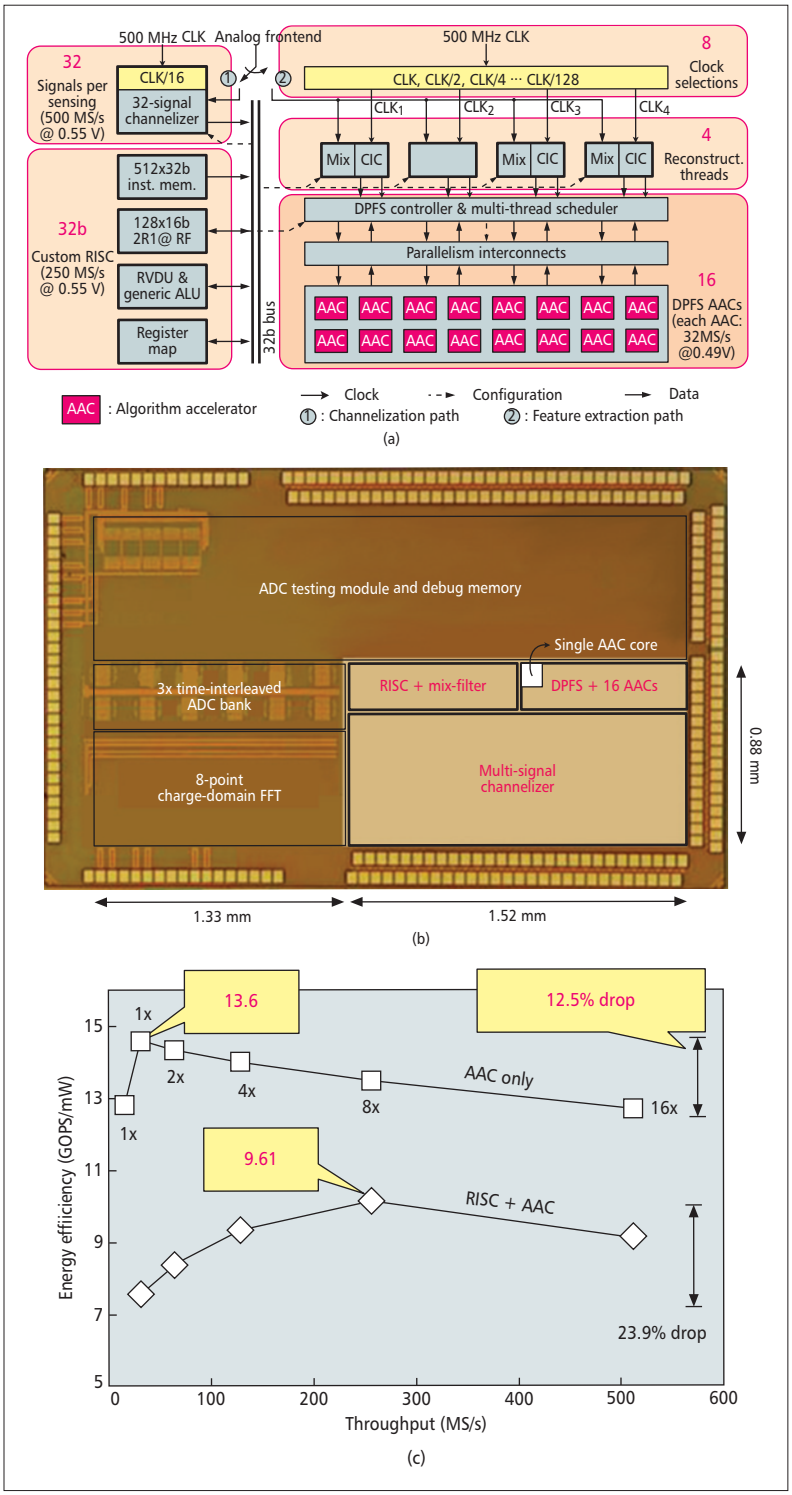


Figure 5. a) D-CLASIC multi-signal classification digital processor architecture; b) A-CLASIC and D-CLASIC 40 nm classification chip micrograph; c) chip performance measurement showing throughput-agnostic energy efficiency.

For continuous operation, three copies of the A-FFT are time-interleaved. This 3X time-interleaving becomes necessary due to the A-FFT processing time and the time needed for VGAs to settle. With 8 channels, a total of 48 VGAs and ADCs (3 copies of 8 channels, complex-valued I/Q signals) were implemented in the analog domain. In the digital domain, 48 I-VGAs, calibration, and an IFFT block are used for signal reconstruction. The hybrid filter bank ADC was implemented in TSMC's 40 nm GP process. The analog area was 1.3 mm^2, and the digital area is 0.39 mm^2. The microphotograph for the combined A-CLASIC and D-CLASIC chip is shown in Fig. 5b. The total power consumption for the hybrid filter bank ADC is 90.4 mW. The 3 copies of the analog FFT and the state machine consume 14.4 mW, and the 48 copies of the VGA and the ADC consume 30 mW and 40 mW, respectively. Power consumption for the digital is 6 mW.

Low-Power Digital Signal Classifier: D-CLASIC

The classification flow consists of three major steps:

1. Power spectral density (PSD) channelization to detect the signal in the frequency domain and estimate its carrier frequency and bandwidth
2. Signal reconstruction that down-converts the signal
3. Feature extraction and classification that implements the hierarchical decision tree

Signal bandwidth (F_s) determines the sampling frequency and thereby dictates throughput and the clock rate for the feature extraction step. As mentioned previously, the blind classification framework uses various statistics to blindly estimate the carrier frequency (F_c), bandwidth (F_s), modulation type, and modulation level. The CAC, for example, estimates F_c and F_s via spectral peaks in the cyclic frequency domain.

Energy-Efficient Real-Time CAC: To realize energy-efficient real-time F_c and F_s extraction, we have demonstrated a hierarchical power spectral density (PSD) sensing with CAC [13]. The challenge of multi-signal feature extraction is to concurrently handle signals of arbitrary bandwidth while ensuring peak energy efficiency for all of them. The hierarchical approach allowed us to achieve 59× lower energy compared to an exhaustive search. (A 10 MHz signal in a 500 MHz channel was used to establish the

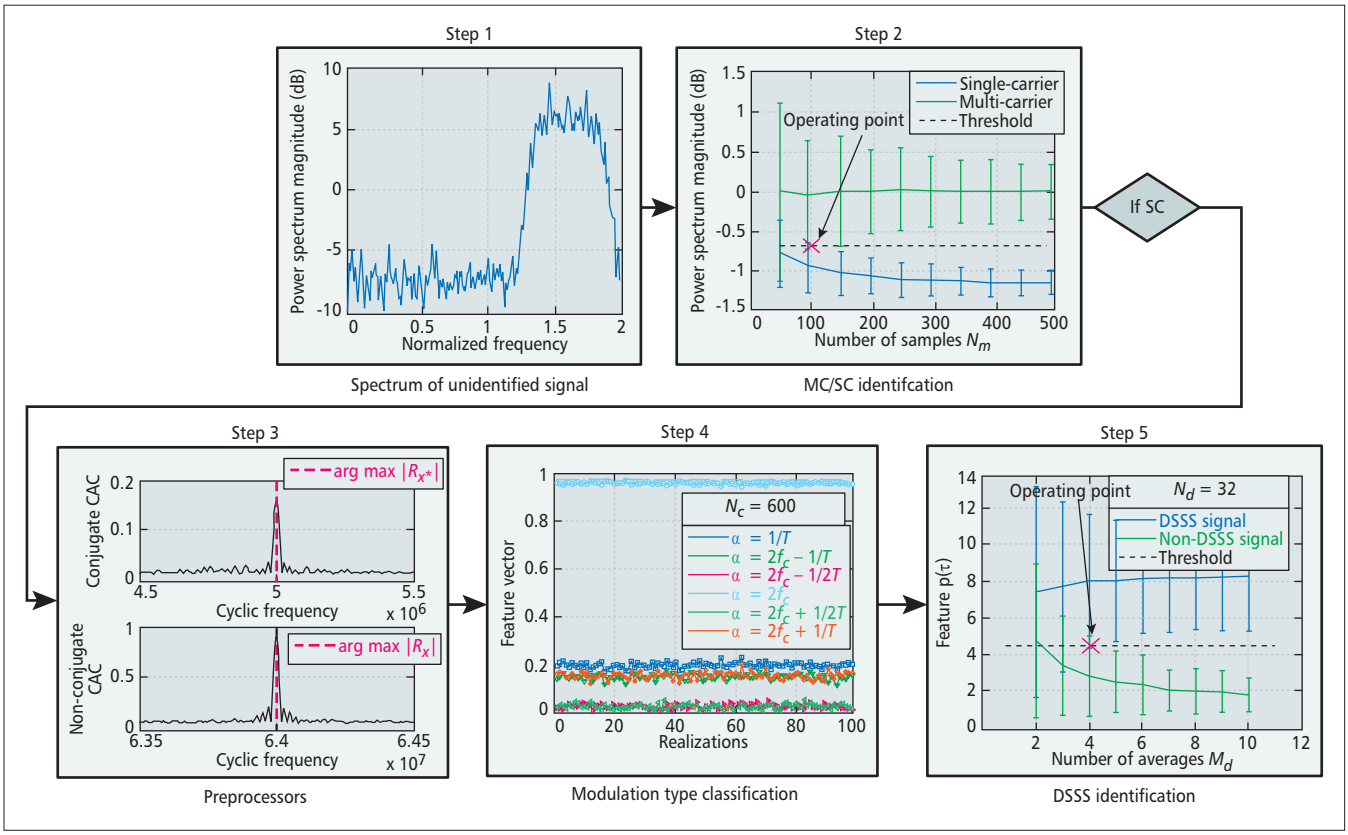


Figure 6. Classification example for a 5 MHz BPSK DSSS signal.

baseline for exhaustive search). We further improved on this to reduce the average energy consumption by another 2.1× by optimizing the micro-architecture and circuit blocks [14], and we introduced a dynamic parallelism-frequency scaling technique to minimize circuit energy. We target up to 500 MHz bandwidth with 12.5 kHz of frequency resolution and 0 dB minimum SNR, with desired detection probability > 95 percent and false alarm rate < 0.5 percent. In the chip we were able to distinguish five modulation classes including multi-carrier OFDM, single-carrier MSK, QAM, PSK, and direct sequence spread spectrum (DSSS), achieving < 2 ms processing time per signal [14].

Throughput-Agnostic Energy Efficiency: To run the circuit at peak energy efficiency, a dynamic voltage-frequency scaling (DVFS) technique is usually employed. Given fixed hardware, the DVFS can only minimize the energy at a fixed throughput, since it is bounded by a single efficiency-frequency (E-F) curve. As the throughput requirement changes, DVFS has to change the supply voltage and frequency, resulting in a sub-optimal energy efficiency. From an implementation perspective, some designers realize DVFS by using two (or a few) voltage domains and employ power gates to switch between them. The number of on-chip voltage domains is limited, so in this case DVFS does not scale well with high throughput resolution. Alternative designs realize DVFS by using an on-chip voltage regulator, but slow voltage adaptation is impractical for the desired frequent throughput changes in our case, with each signal being classified within a few milliseconds. To match the high-resolution throughput possibility at optimal energy efficiency, we propose a dynamic parallelism-frequency scaling (DPFS) technique that allows the hardware to switch between multiple E-F curves. The voltage is fixed, so we do not need to worry about the power-grid overhead. Another benefit of DPFS is that one can target very fine throughput resolution by changing parallelism rather than changing voltage. The parallelism can be dynamically changed by the DPFS controller within a few clock cycles, enabling very fast adaptation to a new optimal efficiency based on real-time throughput requirements. Thus, the sensitivity of energy efficiency to varying signal bandwidth is significantly lowered by DPFS compared to DVFS.

Multi-Signal Classification System on Chip: Our SoC (Fig. 5a) front-end consists of an 8-point charge-domain FFT and a 3× time-interleaved 10b ADC bank. The digital baseband comprises a multi-signal channelizer to sense up to 32 concurrent non-overlapping signals with up to 125 MHz bandwidth in a 500 MHz channel. A 16×-parallel FFT architecture is employed, so the channelizer can run at a reduced voltage (0.55 V in 40 nm complementary metal oxide semiconductor, CMOS, technology) while achieving 500 MS/s throughput. Following the channelizer is a 1–128× clock divider for 8 different clock selections, and 4 mixer-CIC filter pairs to enable up to 4 concur-

The overall low-power design makes this blind signal classifier particularly well suited for untethered battery operation. To the best of our knowledge, this is the first integrated realization of a blind signal classifier that is geared to operate on a battery budget.

rent signal reconstructions based on the estimated carrier frequency and signal bandwidth. The 4 independent post-reconstructed signals are sent to the DPFS controller to share 16 power-gated algorithm accelerator array (ACC) units for feature extraction. All the blocks are managed by a 32b custom RISC with domain-specific instructions and flow control for classification tasks.

Details of the DPFS multi-core scheduling can be found in [14]. The blind classification SoC (Fig. 5b) was fabricated in a 40 nm high-VT process occupying 1.17 mm^2 in the analog front-end and 1.34 mm^2 in the digital baseband.

Power Measurement Results for D-CLASIC: Performance measurements (Fig. 5c) indicate energy efficiency ranging between 11.9 GOPS/mW and 13.6 GOPS/mW (16b equivalent operations) for full-channel feature extraction. The 12.5 percent efficiency variation over > 10× throughput range validates DPFS operation. By including the 0.55 V RISC, the digital baseband achieves 9.61 GOPS/mW (6.7 pJ/sample) at 256 MS/s. Since the RISC does not have DPFS, the system-level efficiency variation was 24 percent. The DPFS helped us achieve more than 2.7× lower efficiency variation than DVFS, and the use of high-VT process enabled 1.6–2.1× better energy efficiency than previous work due to a 3× lower leakage for near-threshold operation. As a result, the chip only consumes between 20.1 and 22.6 mW of power depending on the number of signals and signal bandwidth, enabling wideband blind signal classification on a battery budget. Furthermore, our scheduling also enabled up to 4× shorter classification time per signal [14].

Signal Classification Example

In this section we consider the classification of a binary PSK (BPSK) DSSS signal that is spread with a code of length 8. The DSSS signal has a symbol rate of 5 MHz, and is centered at 125 MHz at an SNR of 10 dB. After detecting the presence of the signal in the band segmentation, the CIC filter down-converts the signal to a center frequency of 16 MHz and decimates, resulting in 4 samples/symbol. Figure 6 shows the output of each of the algorithms discussed in this section. In the first block of the classification tree, the C_{42} cumulant is computed and compared against a threshold. We show that setting $N_m = 90$ samples is sufficient to separate SC and MC classes with a probability of 95 percent. Assuming the DSSS signal is correctly classified as SC, its transmit parameters will be computed using Eqs. 1 and 2. Using $N_T = N_f = 400$ samples, the preprocessing estimates the symbol rate and carrier frequency of the DSSS signal. Using these estimates, the modulation type classifier computes the normalized feature vector $\overline{\mathbf{F}}$, which is compared to the theoretical normalized feature vector for BPSK signals plotted in solid lines in step 4 of Fig. 6 for different realizations of the feature vector. Finally, after being classified as a BPSK signal, the variance of the autocorrelation function $\rho(\tau)$ is computed and compared against a threshold. Given the oversampling ratio of 4, the peak of $\rho(\tau)$ will occur at lag $\tau = 8 \times 4 = 32$. Therefore, detecting the presence of this peak inherently asserts the presence of the DSSS signal and estimates the code length.

Conclusions

In this article we describe an integrated sensor for blind radio signal analysis, incorporating a low-power channelized data converter architecture to digitize wideband signals, and ultra-low power techniques to realize the necessary computational hardware for blind signal classification. This enables extraction of the carrier frequency, signal bandwidth, modulation type, and modulation level with little prior information. The prototype front-end hardware uses an 8-point charge-domain FFT and a 3× time-interleaved 10b ADC bank. The integrated digital processing is able to sense up to 32 concurrent non-overlapping signals with up to 125 MHz signal bandwidth in the observed 500 MHz channel. The overall low-power design makes this blind signal classifier particularly well suited for untethered battery operation. To the best of our knowledge, this is the first integrated realization of a blind signal classifier that is geared to operate on a battery budget.

References

[1] W. Headle and C. da Silva, "Asynchronous Classification of Digital Amplitude-Phase Modulated Signals in Flat-Fading Channels," *IEEE Trans. Commun.*, vol. 59, no. 1, Jan. 2011, pp. 7–12.

[2] P. Urriza, E. Rebeiz, and D. Cabric, "Optimal Discriminant Functions Based on Sampled Distribution Distance for Modulation Classification," *IEEE Commun. Lett.*, vol. 17, no. 10, Oct. 2013, pp. 1885–88.

[3] P. Sutton, K. Nolan, and L. Doyle, "Cyclostationary Signatures In Practical Cognitive Radio Applications," *IEEE JSAC*, vol. 26, no. 1, Jan. 2008, pp. 13–24.

[4] A. Swami and B. M. Sadler, "Hierarchical Digital Modulation Classification Using Cumulants," *IEEE Trans. Commun.*, vol. 48, no. 3, Mar. 2000, pp. 416–29.

[5] O. Dobre *et al.*, "Survey of Automatic Modulation Classification Techniques: Classical Approaches and New Trends," *Commun. IET*, vol. 1, no. 2, Apr. 2007, pp. 137–56.

[6] M. Aslam, Z. Zhu, and A.-K. Nandi, "Automatic Modulation Classification Using Combination of Genetic Programming and KNN," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, Aug. 2012, pp. 2742–50.

[7] E. Rebeiz et al., "Energy-Efficient Processor for Blind Signal Classification in Cognitive Radio Networks," *IEEE Trans. Circuits Sys. I*, vol. 61, no. 2, Feb. 2014, pp. 587–99.

[8] F. Wang and X. Wang, "Fast and Robust Modulation Classification via Kolmogorov-Smirnov Test," *IEEE Trans. Commun.*, vol. 58, no. 8, Aug. 2010, pp. 2324–32.

[9] B. Ramkumar, "Automatic Modulation Classification for Cognitive Radios Using Cyclic Feature Detection," *IEEE Circuits Sys. Mag.*, vol. 9, 2nd qtr. 2009, pp. 27–45.

[10] E. Rebeiz, P. Urriza, and D. Cabric, "Experimental Analysis of Cyclostationary Detectors under Cyclic Frequency Offsets," *Proc. IEEE Asilomar Conf. Signals and Sys.*, 2012.

[11] E. Rebeiz and D. Cabric, "Low Complexity Feature-Based Modulation Classifier and its Non-Asymptotic Analysis," *Proc. IEEE GLOBECOM*, 2011.

[12] H. Shin *et al.*, "An Eight Channel Analog-FFT Based 450MS/s Hybrid Filter Bank ADC With Improved SNDR for Multi-Band Signals in 40nm CMOS," *IEEE Custom Intergrated Circuit Conf.*, Sept. 2015.

[13] F.-L. Yuan, T.-H. Yu, and D. Markovic, "A 500mhz Blind Classification Processor for Cognitive Radios in 40nm CMOS," *Proc. Symp. VLSI Circuits Dig, Technical Papers*, June 2014, pp. 1–2.

[14] F.-L. Yuan *et al.*, "A Throughput-Agnostic 11.9-13.6GOPS/mW Multisignal Classification Soc for Cognitive Radios in 40nm CMOS," *Proc. IEEE VLSI Circuits Symp.*, June 2015, pp. 10–13.

BIOGRAPHIES

RAMESH HARJANI [S'87, M'89, SM'00, F'05] is the Edgar F. Johnson Professor in the Department of Electrical and Computer Engineering at the University of Minnesota. He received his Ph.D. degree from Carnegie Mellon University in 1989, his M.S. degree from the Indian Institute of Technology, New Delhi, in 1984, and his B.S. degree from the Birla Institute of Technology and Science, Pilani, in 1982, all in electrical engineering. Prior to joining the University of Minnesota, he was with Mentor Graphics Corp. in San Jose, California. He co-founded Bermai, Inc, a startup company developing CMOS chips for wireless multimedia applications in 2001. He has been a visiting professor at Lucent Bell Labs, Allentown, Pennsylvania, and the Army Research Labs, Adelphi, Maryland. His research interests include analog/RF circuits for wired and wireless communications.

DANIJELA CABRIC received her Dipl. Ing. degree from the University of Belgrade, Serbia, in 1998, and her M.Sc. degree in electrical engineering from the University of California, Los Angeles (UCLA) in 2001. She received her Ph.D. degree in electrical engineering from the University of California, Berkeley, in 2007, where she was a member of the Berkeley Wireless Research Center. In 2008, she joined the faculty of the Electrical Engineering Department at UCLA, where she is now an associate professor. She received the Samueli Fellowship in 2008, the Okawa Foundation Research Grant in 2009, the Hellman Fellowship in 2012, and the National Science Foundation Faculty Early Career Development (CAREER) Award in 2012. She serves as an Associate Editor of *IEEE Transactions on Cognitive Communications and Networking*, and was the TPC Co-Chair of the 8th International Conference on Cognitive Radio Oriented Wireless Networks 2013.

DEJAN MARKOVIĆ is a professor of electrical engineering at UCLA. He is also affiliated with the UCLA Bioengineering Department as a co-chair of the Neuroengineering field. He completed his Ph.D. degree in 2006 at the University of California, Berkeley, for which he was awarded the 2007 David J. Sakrison Memorial Prize. His current research is focused on implantable neuromodulation systems, domain-specific architectures, embedded systems, energy harvesting, and design methodologies. He co-founded Flex Logix Technologies, a semiconductor IP startup, in 2014. He received an NSF CAREER Award in 2009. In 2010, he was a co-recipient of the ISSCC Jack Raper Award for Outstanding Technology Directions. Most recently, he received the 2014 ISSCC Lewis Winner Award for Outstanding Paper.

BRIAN M. SADLER [S'81, M'81, SM'02, F'07] received his B.S. and M.S. degrees from the University of Maryland, College Park, and his Ph.D. degree from the University of Virginia, Charlottesville, all in electrical engineering. He is a Fellow of the Army Research Laboratory (ARL) in Adelphi, Maryland. His research interests include information science, networked and autonomous intelligent systems, sensing, and mixed-signal integrated circuit architectures.

RAKESH K. PALANI is currently working at Broadcom Corporation, Irvine, California. He received his B.Tech. in electrical engineering from the National Institute of Technology, Kurukshetra, India in 2007 and his Ph.D. in electrical engineering from the University of Minnesota, Twin Cities in 2015. For his Ph.D., he worked on design of PVT-tolerant inverter-based circuits for baseband analog applications. His research interest includes design of low-power filters, ADCs, and amplifiers. In 2011, he was with Broadcom Corporation, Minneapolis, where he worked on design of word line drivers for SRAM. In 2014, he was at Qualcomm, San Diego, where he worked on flicker noise reduction techniques in discrete time delta sigma modulators.

ANINDYA SAHA [S'14] received his B.Sc. (Hons.) degree in electrical and electronic engineering from Bangladesh University of Engineering & Technology (BUET) in 2011 and his M.S degree in electrical engineering from the University of Minnesota, Twin Cities in 2015. Currently, he is working toward his Ph.D. degree at the University of Minnesota. He was a recipient of a University of Minnesota ECE Departmental Fellowship in 2012. His primary research focus is on low-power high dynamic range SAR-pipeline ADC architectures for software-defined radio (SDR).

HUNDO SHIN received his B.S degree in computer science and electrical engineering from Handong University, Pohang, Korea, in 2009, and his M.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, in 2011. Since 2012, he has been working toward his Ph.D. degree at the University of Minnesota, Twin Cities. His research interest includes analog circuits for the channelization of wideband signals.

ERIC REBEIZ [S'06, M'13] received his B.S. degree, summa cum laude, from the University of Massachusetts Amherst in 2008, his M.S. degree from the University of Southern California in 2009, and his Ph.D. degree from UCLA in 2013, all in electrical engineering. He held several positions in the defense industry from 2009 to 2012, working at Exelis Inc. on radar systems as a systems engineer and at the Aerospace Corporation as a member of technical staff. From 2013 to 2015, he was a senior systems engineer at Qualcomm Research developing next-generation 802.11ax WiFi systems in San Diego, California, where his contributions to UL MU-MIMO were awarded a best paper award at the Qualcomm-wide technical conference. He is currently at Tarana Wireless where he designs and develops novel physical-layer signal processing techniques for point-to-multipoint wireless links intended for backhaul applications in non-line-of-sight environments.

SINA BASIR-KAZERUNI [S'09] is a Ph.D. candidate and teaching fellow in the Department of Electrical Engineering at UCLA. He has interned at various companies including NVIDIA, Synopsys, and Honeywell Aerospace. His research interests include design and optimization of energy-efficient communication and biomedical circuits and systems. He is a recipient of a Samueli Fellowship and was awarded the EE Excellence in Teaching Award at UCLA in 2012. He holds an M.S. degree in electrical engineering from UCLA; prior to that, he completed his B.A.Sc. in electrical engineering at the University of Waterloo, Canada.

FANG-LI YUAN [S'10] received his B.S. and M.S. degrees from National Taiwan University, Taipei, in 2006 and 2008, respectively, and his Ph.D. degree from UCLA in 2014, all in electrical engineering. He joined Flex Logix Technologies, a semiconductor IP startup, in 2014. His research interests include flexible DSP architectures and VLSI circuits for communication signal processing, with particular focus on software-defined and cognitive radios. He was awarded the Broadcom Fellowship in 2012 for his research on multicore processors for software-defined radios. In 2014, he received the Distinguished Ph.D. Dissertation Award in Circuits and Systems from UCLA. He won the 2014 ISSCC Lewis Winner Outstanding Paper Award for his work on multi-granularity FPGA for mobile computing.

Flexible Thin-Film NFC Tags

Kris Myny, Ashutosh K. Tripathi, Jan-Laurens van der Steen, and Brian Cobb

ABSTRACT

Thin-film transistor technologies have great potential to become the key technology for leaf-node Internet of Things by utilizing the NFC protocol as a communication medium. The main requirements are manufacturability on flexible substrates at a low cost while maintaining good device performance characteristics, necessary to be compatible with the NFC specifications. Such low-cost flexible NFC tags can be attached to any object with any form factor, connecting this object to the Internet using a smartphone or tablet as an intermediate node. Among all commercial thin-film transistor technologies, metal oxide transistors is a viable technology for this application. The metal-oxide transistors in this work are based on InGaZnO as semiconductor. Since these are unipolar by nature (i.e., they exhibit only *n*-type transistors), different options to make logic circuits are studied from static and dynamic points of view. The different topologies are diode-load logic, dual-gate diode-load logic, and pseudo-CMOS logic. The static parameters lead to a comparison of soft yield between those circuit topologies, while the transient analysis provides insight on the power consumption and circuit speed. This is indicative for selecting the logic style matching the data rate requirements of the NFC standards. Moreover, metal-oxide NFC circuits that combine 12-bit code generators to the analog front-end of RFID tags are integrated on a PCB board to evaluate performance of a matched and optimized system. The measured data rates of these integrated NFC tags are compatible with the ISO 15693 specifications. Finally, a fully integrated, flexible NFC tag is realized, which comprises the tuning capacitor, rectifier, load modulator, and code generator.

INTRODUCTION

Near field communication (NFC) is a short-range wireless communication protocol designed for interactions below 10 cm. Introduced as a secure protocol for mobile payments, data transfer between two devices, smart card interactions, and many other use cases, it has potential to function as a low-level communication protocol for the Internet of Things (IoT). In such an envisioned system, all leaf nodes, or objects with embedded NFC tags, utilize a point-to-point connection with a smartphone, tablet, or other intermediate reader (node), which is subsequently connected to the Internet. NFC tags based on silicon complementary metal oxide semiconductor (CMOS) microchips are only suited for high-end and medium-priced objects. Low-cost products, on the other hand, require embedded NFC tags manufactured at a sufficiently low cost point. One route to low-cost fabrication is the use of thin-film transistor technologies in place of silicon.

An NFC system comprises a reader device and an NFC tag, as depicted in Fig. 1. This tag does not require a battery; it is, instead, powered by the energy of the 13.56 MHz carrier wave produced by the reader device. In the near-field use case, the coupling from the reader to the antenna is inductive. A rectifier is attached to the antenna to generate DC supply voltages for the circuitry from the input AC wave. The output of the code generator is fed back to the antenna with a load modulator to perform amplitude shift keying (ASK) modulation. This will change the load of the antenna as a function of the code, which can be detected at the reader side.

Both the rectifier and load modulator need to be able to operate at the carrier frequency (f_c), 13.56 MHz. The requirements of the digital circuitry and code generator are listed in the NFC standards. The ISO 14443 standard describes data rates from Tag to Reader of 105.9 kb/s, or $f_c/128$. ISO 15693 requires lower data rates, 6.62 kb/s for the slow mode and 26.48 kb/s for the fast operating mode [1]. These performance specifications, combined with cost limitations, set the boundary conditions for the transistor technology to be used.

THIN-FILM TRANSISTOR TECHNOLOGIES

Thin-film transistor technologies are present in a multitude of applications nowadays, ranging from image sensor arrays to TV panels. Figure 2 visualizes three different thin-film transistor technologies that exist today in products. Key parameters for flexible NFC tags are the maximum processing temperature (related to polymer film capability), mobility (thin film transistor, TFT, performance requirement) and uniformity (design and manufacturing robustness). Stability on the other hand is not a key parameter, since tags are idle most of the time.

Kris Myny is with Imec.

Ashutosh K. Tripathi, Jan-Laurens van der Steen, and Brian Cobb are with Holst Centre/TNO.

0163-6804/15/$25.00 © 2015 IEEE

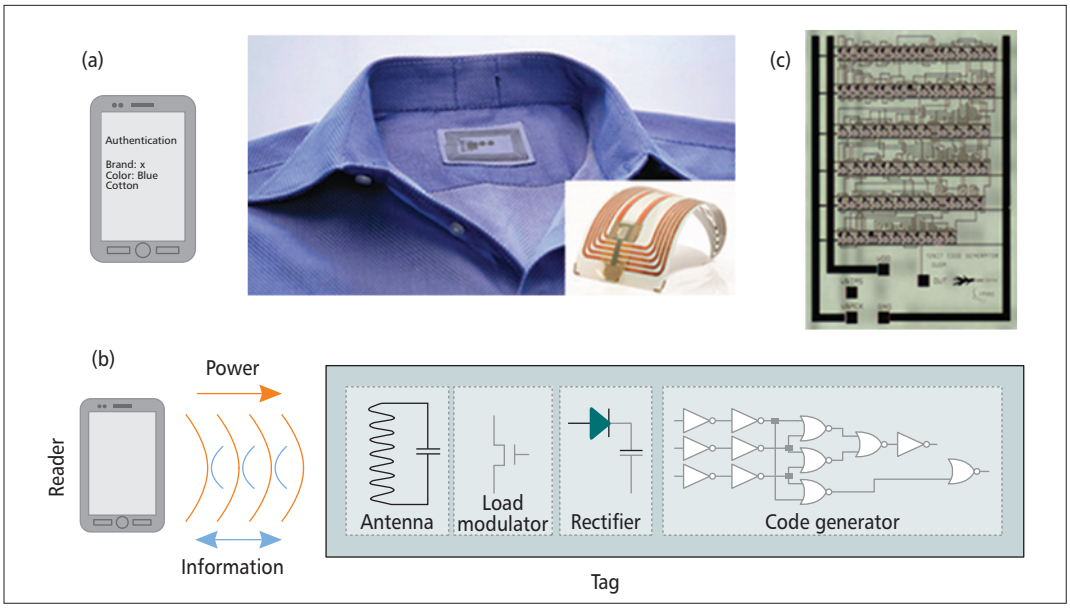


Figure 1. Flexible NFC tag embedded in clothes for leaf-node IoT applications (top), and system blocks and operation of passive NFC tags (bottom).

LTPS and metal oxide TFTs are therefore better suited for NFC tags due to their superior performance characteristics with mobilities one to two orders of magnitude higher.

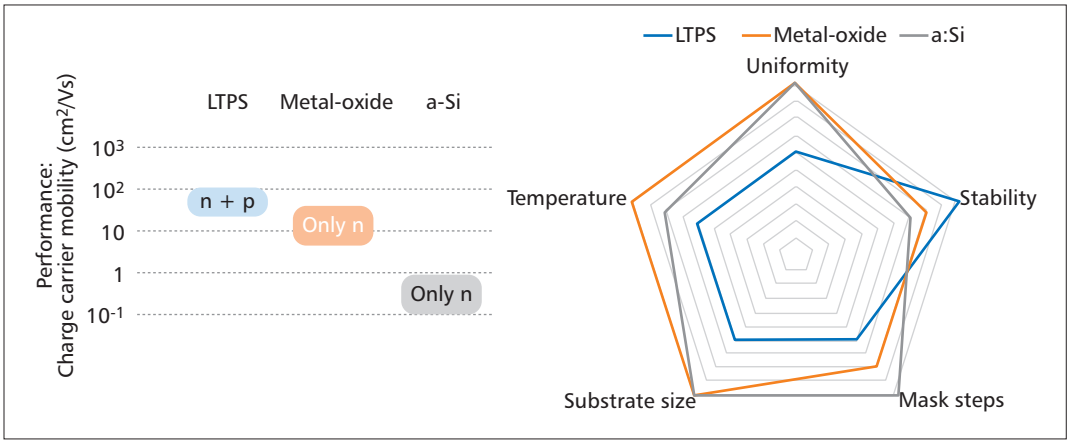


Figure 2. Performance overview of three TFT technologies (left), and comparison of those TFT technologies for uniformity, stability, number of mask steps, substrate size, and process temperature (right).

Amorphous silicon (a:Si) is the most widely used transistor technology whereby the transistors are mainly used as switches in LCD panels. Although cost and large area uniformity are key parameters for the flexible NFC tag, and process temperatures of ~300˚C are compatible with flexible polyimide (PI) substrates, charge carrier mobilities up to 1 cm^2/Vs may be insufficient to comply with all specifications regarding NFC application. Low temperature polycrystalline silicon (LTPS) and metal oxide TFTs are therefore better suited for NFC tags due to their superior performance characteristics with mobilities one to two orders of magnitude higher. The process temperature for LTPS TFTs is relatively high (~450˚C), reducing the possibilities to manufacture the TFTs directly on flexible substrates. An alternative option is to transfer the TFTs after fabrication onto a flexible medium, introducing additional complexity and therefore cost to the manufacturing process [2]. Metal oxide TFTs, on the other hand, can be processed at temperatures from room temperature to ~350˚C [3]. This would enable manufacturing directly on a broad range of flexible substrates with different temperature budgets including polyethylene terephthalate (PET, ~150˚C), polyethylene naphthalate (PEN, ~180˚C) and polyimide (PI, ~300˚C).

The device-to-device uniformity of TFTs plays an important role in the final realization of NFC circuits. The device characteristics of amorphous metal oxide TFTs have proven to be uniform over a large area [4], which is advantageous for integrated circuit yield. Although LTPS TFTs suffer from increased variability due to the intrinsic grain boundaries of the semiconductor, the presence of both n-type and p-type devices for circuit design will be beneficial for the final circuit yield [5]. Metal oxide TFTs, on the contrary, are mostly n-type only devices, requiring novel circuit design techniques to ensure a high

circuit yield. To overcome this hurdle, back in the 1970s, prior to the existence of a silicon CMOS, complex circuits were realized using multiple threshold voltage (V_T) logic, by combining depletion and enhancement transistors into logic gates. In TFT technologies, a similar approach can be followed [6].

LTPS and metal oxide TFT technologies are both good candidates for flexible NFC tags. Recently, Salvatore *et al*. demonstrated IGZO TFTs wrapped around a human hair to showcase extreme flexibility [7]. The stability characteristics of the TFTs (e.g., bias stress) are less important for flexible NFC applications due to the short on-time of the circuit and biased conditions, which is on the order of a few milliseconds to a few seconds. Metal oxide TFTs also require fewer process (photolitho) steps to fabricate, leading to lower projected costs. The total cost of the metal oxide transponder chip may achieve sub-1-cent cost per chip, leveraging the large-scale infrastructure and low-cost model developed for displays upon full scale-up. From there additional savings include simpler integration schemes such as the elimination of the crossover step on the antenna by using the flexible chip instead of the common rigid silicon chip. In this work, we focus on metal oxide TFTs to realize the flexible thin-film NFC tags due to their advantages in process temperature, cost, and uniformity, while still providing sufficient performance.

Figure 3 depicts the cross-section of the metal oxide TFT, where amorphous InGaZnO (a-IGZO) is used as a semiconductor. All layers have been processed directly on a flexible PEN

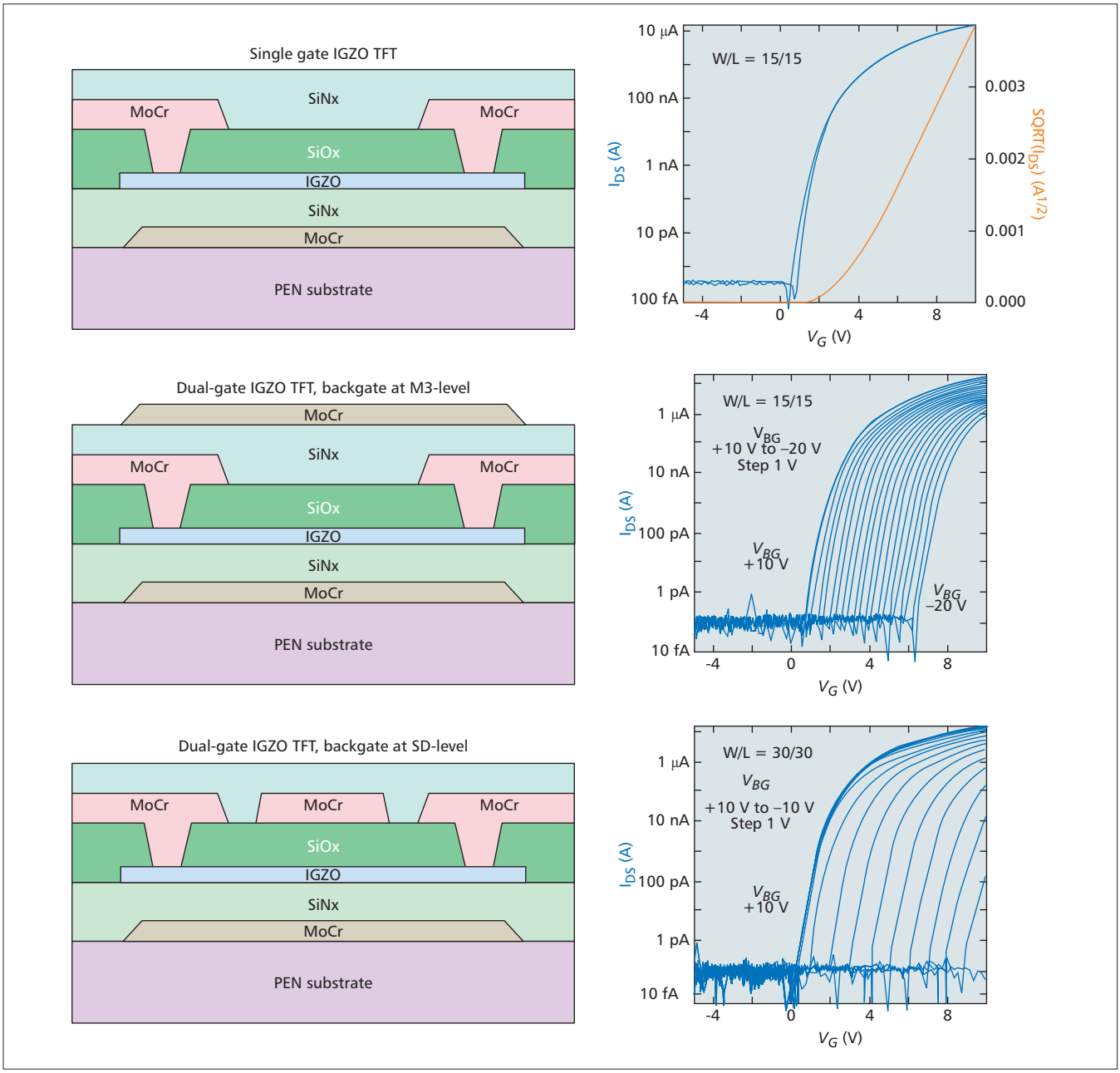


Figure 3. Device cross-section and TFT transfer curves for (top) single gate IGZO TFT; (middle) dual-gate IGZO TFT in metal-3 layer; (bottom) metal-2 layer.

substrate. The cross-section reveals the etch stop layer (ESL) configuration, where the backchannel of the IGZO semiconductor is protected from further process steps by an insulating layer. All layers and processes were optimized to ensure sufficient device stability while keeping the maximum process temperature compatible with the flexible PEN substrate (~180˚C). Furthermore, all materials and processes are compatible with existing flat panel display (FPD) lines enabling mass-scale manufacturability. Figure 3 also depicts the typical measured transfer characteristic of a transistor with both channel length and width equal to 15 µm. The layouts of the transistors have been made using conservative critical dimensions of 5 µm. This sets the minimal channel length of the ESL transistors to 15 µm.

This technology features an individual back gate for each transistor, leading to dual-gate TFTs. The back gate contact can be used to shift the threshold voltage (V_T), leading to multiple V_T logic gates, which is useful to improve the robustness of unipolar logic gates [6]. The back gate metal can be defined in either the third metallization layer or the second (SD) metallization layer. Applying a negative voltage to the back gate with respect to its source shifts the transfer curve toward more enhancement mode devices, as indicated in Fig. 3. The sensitivity of the back gate is 0.3 V for 1 V applied V_{BGS} for the M3-level back gate configuration and approximately 1 V for 1 V applied V_{BGS} for the SD-level implementation. This difference in sensitivity depends on the effective gate-to-channel coupling, which is stronger for the SD level compared to the M3-level back gate. Besides the improved sensitivity of the back gate at the SD level, this option also removes two photo litho steps from the process flow, decreasing the final chip cost. The minimum distance between source, drain, and back gate contacts is slightly increased because of lithographic constraints, resulting in somewhat longer minimum channel length compared to M3-level back gate. The channel length of a transistor with a back gate at the SD level increases to 30 µm, taking into account the critical dimension of the layouts.

Logic Gate Topologies

DC Behavior

We can distinguish between three different options to realize robust logic gates for the flexible NFC tag in a unipolar metal oxide TFT technology. In one, the conventional p-type transistor is replaced by an n-type transistor and used as an active load; in another, a multiple-V_T solution is employed with the same configuration, and finally, a level shifter stage may be added. Because of the unipolar nature of IGZO and the near-zero threshold voltage, diode-load (or enhancement-mode) logic embodies the first implementation. In this inverter configuration, the load transistor has a shorted gate-drain connection. The resulting voltage transfer characteristics (VTC) are shown in Fig. 4, and display asymmetric and non-rail-to-rail curves, leading to limited noise margins. Noise margins are extracted using the maximum equal criteria (MEC) [8–10] and provide a measure of immunity against noise and variability of the circuit. Figure 4e shows that even an increased supply voltage has little impact on noise margins, making this logic style unsuitable for achieving high yields of large-scale integrated circuits.

One way to enhance the noise margin is to realize multiple V_T diode-load inverters by adding a separate back gate contact to the load and drive transistors as shown in Fig. 4. The back gate contact of the drive (or pull down) transistor is biased with –VDD for the dual-gate M3 case and VDD/3 for the dual-gate M2 case. The back gate of the load transistor (M3 case) is connected to the output node, resulting in a 0 V back gate voltage with respect to its source. We have not foreseen a back gate contact at the load transistor for the M2 case. As such, its channel length can be decreased from 30 µm to 15 µm, improving the switching speed of the inverter. The VTC curves of both configurations are shown in Fig. 4. It can be noted that the additional bias voltage (or second V_T) shifts the VTC curve toward the middle of the voltage rail, increasing the corresponding noise margins. The extracted noise margins divided by VDD is the largest for all different logic styles, as depicted in Fig. 4e.

The third option to improve the diode-load inverter is to add a second stage, also shifting the VTC curve toward the middle of the supply voltage range. The schematic is shown in Fig. 4, and was formerly published as pseudo-CMOS logic [11]. The first stage diode-load inverter is supplied with VBIAS (equal to twice VDD), while the input and output range is only between 0V and VDD. The corresponding noise margin divided by VBIAS as a function of the largest supply voltage (VBIAS) is plotted in Fig. 4e, resulting in values larger than diode-load and lower compared to both dual-gate options. Moreover, this logic style requires double the number of TFTs, which is disadvantageous for area consumption and therefore hard yield.

Both the dual gate options and pseudo-CMOS require an additional voltage rail. We have optimized the technology stack and the logic inverter such that these voltage rails can be generated easily in RFID/NFC tags by rectifying the carrier signal using a double-half wave rectifier [12].

Figure 4f depicts the calculated soft yield from the measured noise margins as a function of the number of integrated transistors for a supply voltage of 10 V, assuming a local variability on V_T of 100 mV. The mathematics for the calculations is similar to work published earlier by De Vusser *et al.* [13]. It can be observed that, with this assumed variability and supply voltage, the dual-gate M3 option would enable large-scale integrated circuits that combine more than 100 million TFTs with a high soft yield. Pseudo-CMOS, on the other hand, enables about 1000 TFTs integrated with high soft yield.

Transient Behavior

We have designed a 12-bit code generator as a test vehicle to examine the transient behavior of these logic styles. The block diagram is depicted in Fig. 5; a die picture is shown in Fig. 1. The

Both the dual gate options and pseudo-CMOS require an additional voltage rail. We have optimized the technology stack and the logic inverter such that these voltage rails can be generated easily in RFID/NFC tags by rectifying the carrier signal using a double-half wave rectifier.

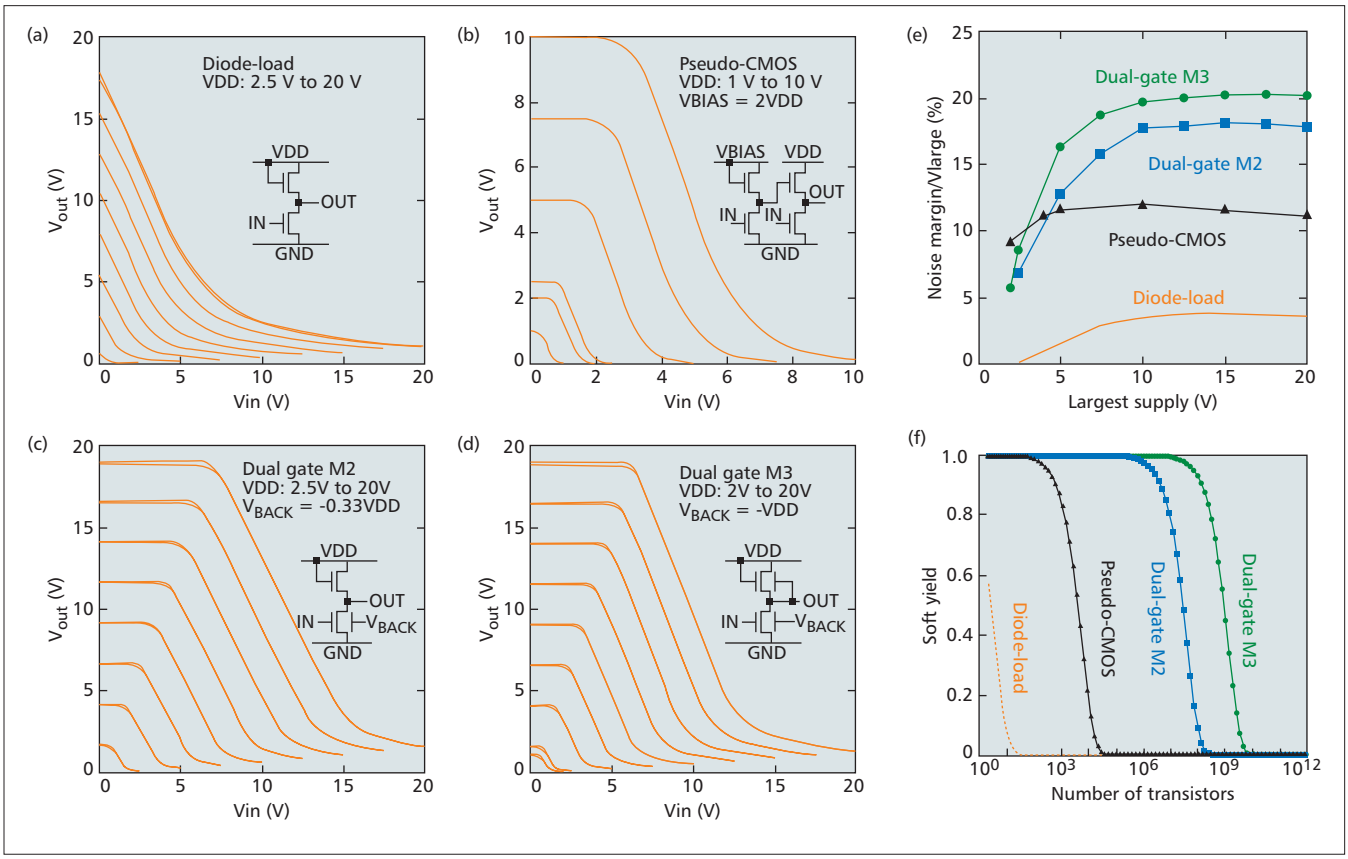


Figure 4. Voltage transfer curve and transistor scheme of a) diode-load; b) pseudo-CMOS; c) dual-gate M2; d) M3 inverters; e) extracted noise margin over the largest supply voltage as a function of its largest supply voltage; f) calculated soft yield as a function of number of integrated transistors for diode-load, pseudo-CMOS, and both dual-gate inverter implementations, assuming a largest supply of 10 V and $\sigma(V_T) = 100$ mV.

clock generator is implemented as a free-running 19-stage ring oscillator. This will clock the 4-bit modulo-12 counter and output register. The 12-bit decoder is a hardcoded synthesized memory with a code sequence 0101 0011 1001. This block diagram has been used to generate all four different implementations. A typical measurement diagram is also shown in Fig. 5. When 5 V VDD and 10 V VBIAS is applied to the 12-bit code generator for the pseudo-CMOS implementation, the above-mentioned code sequence is observed, with a data rate of 9.8 kb/s. The measured data rate of all four implementations as a function of its largest supply voltage (VDD or VBIAS) is plotted in Fig. 5. The diode-load code generator exhibits the fastest data rates, up to 71.6 kb/s at 20 V VDD. The maximum observed data rates for the pseudo-CMOS implementation is 43.9 kb/s at 10 V VDD and 20 V VBIAS. The data rate is smaller because of the two-stage inverter gate. Dual-gate M3 has a data rate of only 25.8 kb/s for 20 V VDD and –20 V VBACK. Compared to single-gate diode-load logic, the drive transistor has less current available due to the negative back gate contact, resulting in slower data rates compared to conventional diode-load logic. The slowest implementation is dual-gate M2, which can be explained by a larger channel length for the drive transistor. In all cases, the design rules for device layout can be projected to scale to smaller sizes in production, yielding large increases in operation speed.

The power consumption of these transponder chips is also plotted as a function of its largest supply voltage. All four implementations exhibit a similar power consumption, whereby the regular diode-load configuration consumes the most. At low supply voltages, power consumption less than 1 μW for these chips has been measured.

Flexible NFC Tags

Figure 6 details the system schematic for three different 12-bit code generators. The rectifier is based on a double half-wave rectifier, which is efficient to realize the three power rails necessary for dual-gate and pseudo-CMOS configurations, as indicated in Fig. 6. Transistors with shorted gate-drain connection are used as diodes in this scheme. The output of the 12-bit code generators is then connected to the gate of the load modulation transistor, which is placed in parallel to the LC-tank. It will modulate the antenna current resulting in ASK modulation of the code sequence on top of the carrier, which can be detected by the reader hardware.

The NFC tags are powered by a commercial USB-connected NFC reader device (SCL011). The first measurements have been performed by connecting on a printed circuit board (PCB) all different modules (rectifier, code generator, load modulator) to the antenna with a tunable capacitor for 13.56 MHz resonance optimization. This PCB is only meant for interconnecting the dif-

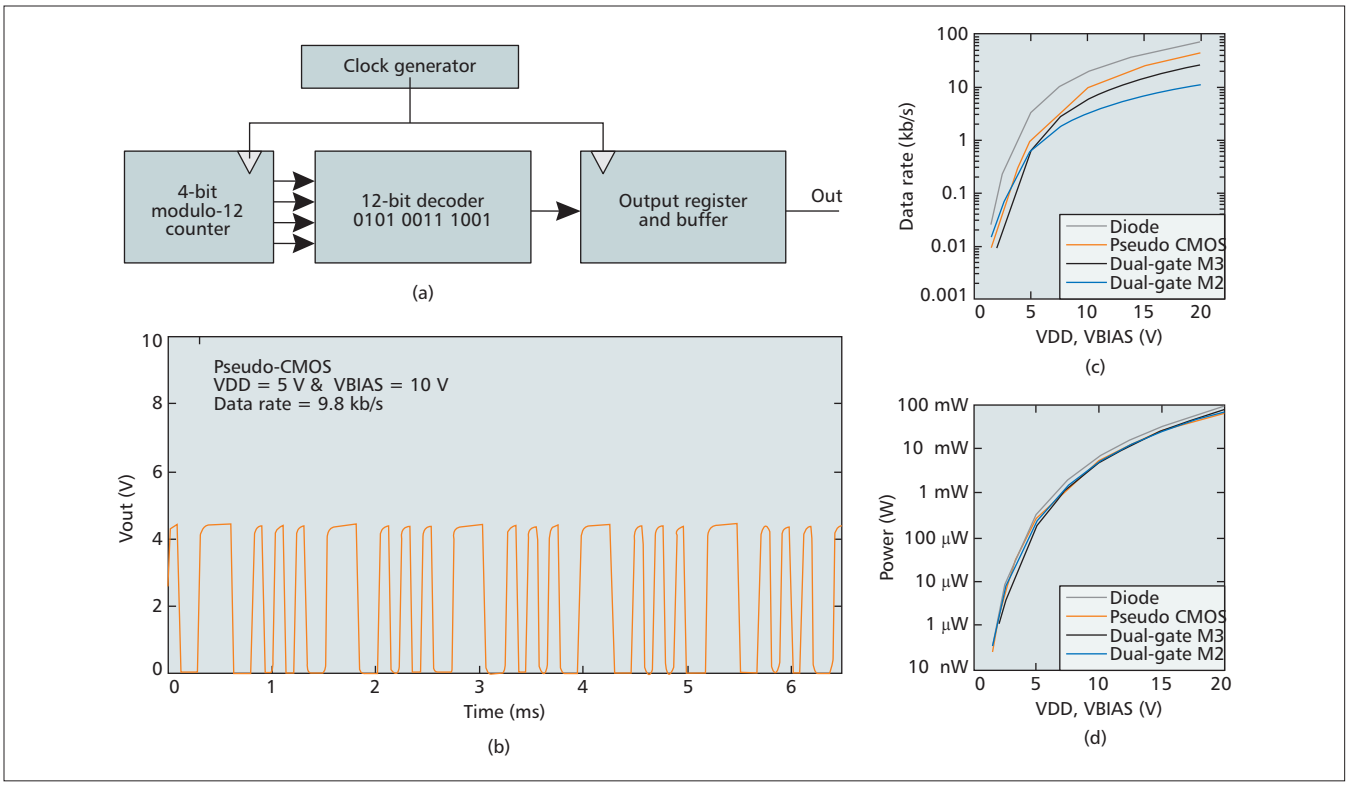


Figure 5. a) Block diagram of the 12-bit digital code generator; b) typical obtained measurement of the output sequence for the pseudo-CMOS case, when VDD = 5 V and VBIAS = 10 V; c) measured data rate; d) power as a function of the largest supply voltage in the circuit.

ferent modules and allows us to measure the internal signals. Figure 6c depicts the rectified large supply voltage (VDD or VBIAS) as a function of the distance from the commercial NFC reader. The tags can be successfully powered within 5 cm distance from the reader, with a maximum supply voltage of 10.3 V for the diode-load tag at 1.1 cm distance. The pseudo-CMOS implementation achieves a maximum VDD of 5.2 V and 9.8 V VBIAS at a distance of 1.9 cm from the reader. The back gate (M3) implementation reaches 9.2 V VDD at 1.1 cm distance from the commercial reader. The measured data rates of the code generator at 1.1 cm from the NFC reader are 23.8 kb/s, 9.4 kb/s, and 1.4 kb/s for the diode-load, pseudo-CMOS, and dual-gate (M3) metal oxide RFID tags, respectively. The data rates of the diode-load and pseudo-CMOS RFID tags are ISO 15693 compatible.

Besides the integration on PCB with a tunable capacitor, the metal oxide RFID/NFC tags have been integrated by laminating the oxide NFC foil directly on the antenna. The oxide NFC foil comprises the matching metal-insulator-metal capacitor, the load modulator, the rectifier, and the code generator. A sense coil or spy antenna is applied in the field to detect the AM modulated signal on the carrier, transmitted from the flexible RFID/NFC tag. Figure 6d plots the different modulated signals obtained from all three metal oxide tags, integrating a diode-load, pseudo-CMOS, or dual-gate M3 code generator. The 12-bit code sequences are characterized to have a data rate of 1.46 kb/s for the diode-load configuration, and only 0.14 and 0.19 kb/s for the pseudo-CMOS and diode-load configuration, respectively. These low data rates indicate a non-perfectly matched capacitor to be resonant at 13.56 MHz for the fully integrated flexible NFC tags. The measured modulation indexes are 0.4, 0.7, and 1.7 percent for the diode-load, pseudo-CMOS, and dual-gate M3 implementation, respectively. There are several possibilities to increase the modulation index to conform the ISO standards (10 percent), among others increasing the size of the load modulator transistor to pull more antenna current, a better matching of the resonance frequency by a better tuned capacitor, and decreasing the channel length in order to have a transistors' cutoff frequency above 13.56 MHz to improve the efficiency of the load modulator.

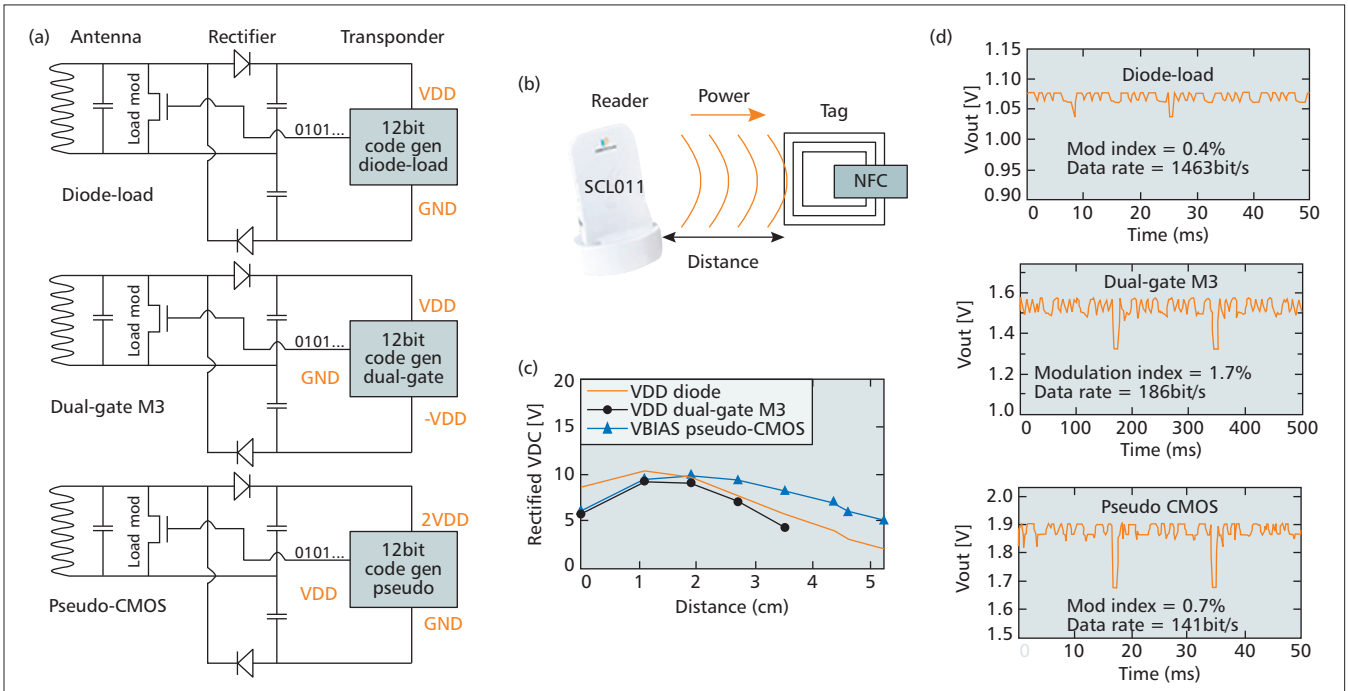
Comparison

The table in Fig. 6 details an overview of all discussed RFID tags and benchmarks to state-of-the-art a-IGZO RFID tags. Ozaki *et al.* implemented an integrated RFID circuit on glass based on zero-V_{GS}-load logic, whereby the load-TFT has shorted gate-source nodes [14]. They achieved a low power consumption below 20 μW at 5 V VDD supply voltage. The clock frequency of the circuit is only 50 Hz, which can be explained by the logic style. This RFID tag has been read out at a distance of 5–8 cm from a commercial 40 mW reader. Yang *et al.*, discussed a transparent a-IGZO RFID tag on glass based on conventional diode-load logic [15]. The data rate of the tag was 3.2 kHz, at 6 V supply voltage consuming 170 μW of power.

Although zero-V_{GS}-load logic yields relatively high values for noise margin considering unipolar technologies, more complex logic styles like dual-gate and pseudo-CMOS exhibit increased noise margins. In terms of data rates, at larger supply voltages, all our transponder chips resulted in data rates compatible with the ISO 15693 RFID standard (> 6.62 kb/s).

Conclusion

Low-cost flexible thin-film NFC tags are a promising route to cost-efficient ubiquitous leaf nodes enabling the Internet of Things and establishing connections between everyday objects and the digital world. Metal oxide TFT technology is a good candidate to enable this market due to its beneficial characteristics in terms of performance and variability, its compatibility with flexible substrates, and the mass production possibilities in existing FPD infrastructure. The main challenge is to deal with the unipolar nature of the semiconductor, which limits the soft yield of the integrated circuits. We have therefore discussed several options to realize robust logic blocks like inverters for a unipolar n-type-only technology. In terms of technology options, adding a back-gate contact could shift the threshold voltage of individual TFTs. Introducing these TFTs in diode-load logic resulted in the most robust implementation, enabling the integration of 100 million n-TFTs with a high soft yield for 100 mV spread on V_T and at 10 V VDD. Pseudo-CMOS logic is an alternative solution to realize robust logic gates, exhibiting a smaller potential of integration density. In terms of dynamic characteristics, the simple diode-load logic yields fastest data rates of 12-bit code generators, up to 71.6 kb/s, while the pseudo-CMOS implementation results in data rates up to 43.9 kb/s. At the expense of reduced speed, dual-gate diode-load logic offers the most robust implementation.



(e)

	Diode-load	Dual-gate M2	Dual-gate M3	Pseudo-CMOS	[14] Zero-V_{GS}-load	[15] Diode-load
# TFTs/inv	2	2	2	24	2	2
Area (mm^2)	2.70 × 2.98 (8.046)	3.91 × 3.87 (15.132)	2.70 × 3.14 (8.478)	4.69 × 3.36 (15.759)	7 × 10 (70)	3.9 × 1.5 (5.85)
# TFTs	218	218	218	436	1026	222
# VDD	2	3	3	3	2	2
Noise margin	< 7.4% VDD/2	~36% VDD/2	~40% VDD/2	~24% VBIAS/2	~15% VDD/2	11.6% VDD/2
Data rate	71.6 kb/s	11.3 kb/s	25.8 kb/s	43.9 kb/s	0.05 kb/s	3.2 kb/s
ISO15693 compatible	Yes	Yes	Yes	Yes	No	No
Carrier	PEN-film	PEN-film	PEN-film	PEN-film	Glass	Glass

Figure 6. a) Power and modulation scheme for the integrated a-IGZO NFC chips; b) schematic representation of the NFC USB reader setup; c) the internally measured DC voltage from the rectifier as a function of the distance between reader and tag; d) measured signals at a sense coil, applied in the field for the fully flexible IGZO NFC tags for three different logic styles.

At the system level, a double half-wave rectifier is used for the NFC tag to generate all necessary voltage rails for the different 12-bit code generator implementations. This rectifier generates 10.3 V VDD for the diode-load configuration, 9.8 V VBIAS for pseudo-CMOS, and 9.2 V VDD for the back gate (M3) implementation, powered by a commercial NFC reader device. The corresponding data rates are ISO 15693 compatible for diode-load and pseudo-CMOS NFC tags. Fully integrated flexible oxide NFC tags have been measured; the ASK modulated signals exhibited a maximum detected modulation index of 1.7 percent for the dual-gate M3 implementation.

Acknowledgment

This work has been supported by the Flemish IWT project ORCA.

References

[1] D. K. Finkenzeller, *RFID Handbook: Fundamentals and Applications in Contactless Smart Cards, Radio Frequency Identification and Near-Field Communication*, 3rd ed., Wiley, 2010.

[2] G. Fortunato, A. Pecora, and L. Maiolo, "Polysilicon Thin-Film Transistors on Polymer Substrates," *Material Sci. Semiconductor Processing*, vol. 15, no. 6, Dec. 2012, pp. 627–41.

[3] E. Fortunato, P. Barquinha, and R. Martins, "Oxide Semiconductor Thin-Film Transistors: A Review of Recent Advances," *Adv. Materials*, vol. 24, no. 22, 2012, pp. 2945–86.

[4] H.-H. Hsieh *et al.*, "Development of IGZO TFTs and their Applications to Next-Generation Flatpanel Displays," *J. Info. Disp.*, vol. 11, no. 4, Dec. 2010, pp. 160–64.

[5] D. Bode *et al.*, "Noise-Margin Analysis for Organic Thin-Film Complementary Technology," *IEEE Trans. Electron Devices*, vol. 57, no. 1, Jan. 2010, pp. 201–8.

[6] K. Myny *et al.*, "Unipolar Organic Transistor Circuits Made Robust by Dual-Gate Technology," *IEEE J. Solid-State Circuits*, vol. 46, no. 5, May 2011, pp. 1223–30.

[7] G. A. Salvatore *et al.*, "Wafer-Scale Design of Lightweight and Transparent Electronics that Wraps Around Hairs," *Nat. Commun.*, vol. 5, Jan. 2014.

[8] J. Lohstroh, E. Seevinck, and J. de Groot, "Worst-Case Static Noise Margin Criteria for Logic Circuits and Their Mathematical Equivalence," *IEEE J. Solid-State Circuits*, vol. 18, no. 6, Dec. 1983, pp. 803–07.

[9] E. Seevinck, F. J. List, and J. Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells," *IEEE J. Solid-State Circuits*, vol. 22, no. 5, Oct. 1987, pp. 748–54.

[10] J. R. Hauser, "Noise Margin Criteria for Digital Logic Circuits," *IEEE Trans. Educ.*, vol. 36, no. 4, Nov. 1993, pp. 363–68.

[11] T.-C. Huang *et al.*, "Pseudo-CMOS: A Design Style for Low-Cost and Robust Flexible Electronics," *IEEE Trans. Electron Devices*, vol. 58, no. 1, Jan. 2011., pp. 141–50

[12] K. Myny *et al.*, "An Integrated Double Half-Wave Organic Schottky Diode Rectifier on Foil Operating at 13.56 MHz," *Appl. Phys. Lett.*, vol. 93, 2008, p. 093305.

[13] S. De Vusser, J. Genoe, and P. Heremans, "Influence of Transistor Parameters on the Noise Margin of Organic Digital Circuits," *IEEE Trans. Electron Devices*, vol. 53, no. 4, Apr. 2006, pp. 601–10.

[14] H. Ozaki *et al.*, "20-μW Operation of an a-IGZO TFT-Based RFID Chip Using Purely NMOS 'Active' Load Logic Gates with Ultra-Low-Consumption Power," *Proc. 2011 Symp. VLSI Circuits*, 2011, pp. 54–55.

[15] B.-D. Yang *et al.*, "A Transparent Logic Circuit for RFID Tag in a-IGZO TFT Technology," *ETRI J.*, vol. 35, no. 4, Aug. 2013, pp. 610–16.

Biographies

Kris Myny received his Master's degree at the Katholieke Hogeschool Limburg in Diepenbeek, Belgium, in 2002. He joined imec in Leuven in 2004 as a member of the Polymer and Molecular Electronics group. In January 2013, he received his Ph.D. degree from KULeuven on the design of thin-film transistor circuits. His main research interests are the design, fabrication, and optimization of digital thin-film circuits for, among others, RFID/NFC tags and AMOLED-backplanes. He received the imec 2010 scientific excellence award and 2011–2012 IEEE SSCS Predoctoral Achievement Award, and was a co-recipient of the 2012 JSID Outstanding Student Paper of the Year Award and the LG Silver Medal at the 14th International Meeting on Information Display in 2014.

Ashutosh Tripathi received his Master's degree in physics from the Indian Institute of Technology, Kanpur, in 2002. He received his Ph.D. in physics from Universität Stuttgart, Germany, in 2008. In his Ph.D. work he did research on growth and characterization of organic single crystals. Since 2008 he has been a researcher at Holst Centre, a joint research initiative of TNO and imec, in the technology program called Organic and Oxide Transistors. He has over 30 publications in peer reviewed journals, conference proceedings, and book chapters. His main research interest is flexible TFT technology for circuits and display applications.

Jan-Laurens P. J. van der Steen received his M.Sc. and Ph.D. degrees in electrical engineering from the University of Twente, The Netherlands, in 2006 and 2011, respectively. Part of his Ph.D. research was carried out at the University of Udine, Italy. Since 2011 he has been with Holst Centre, Eindhoven, The Netherlands. His research interests include device physics, with a focus on modeling and simulation of carrier transport in very thin semiconductor layers, as well as modeling and characterization of emerging technologies such as organic and metal oxide thin film transistors for flexible electronic applications.

Brian Cobb received a B.S. degree in electrical engineering from the University of Virginia in 2004, and M.S. and Ph.D. degrees in solid state electronics from the University of Texas at Austin in 2006 and 2010, respectively. He currently works as a research scientist at the Holst Centre (an open collaboration between TNO and imec) in Eindhoven, the Netherlands.

The corresponding data rates are ISO 15693 compatible for diode-load and pseudo-CMOS NFC tags. Fully integrated, flexible oxide NFC tags have been measured, the ASK modulated signals exhibited a maximum detected modulation index of 1.7 percent for the dual-gate M3 implementation.

ACCEPTED FROM OPEN CALL

Wireless Communications in the Era of Big Data

Suzhi Bi, Rui Zhang, Zhi Ding, and Shuguang Cui

ABSTRACT

The rapidly growing wave of wireless data service is pushing against the boundary of our communication network's processing power. The pervasive and exponentially increasing data traffic present imminent challenges to all aspects of wireless system design, such as spectrum efficiency, computing capabilities, and fronthaul/backhaul link capacity. In this article, we discuss the challenges and opportunities in the design of scalable wireless systems to embrace the big data era. On one hand, we review the state-of-the-art networking architectures and signal processing techniques adaptable for managing big data traffic in wireless networks. On the other hand, instead of viewing mobile big data as an unwanted burden, we introduce methods to capitalize on the vast data traffic, for building a big-data-aware wireless network with better wireless service quality and new mobile applications. We highlight several promising future research directions for wireless communications in the mobile big data era.

Suzhi Bi and Rui Zhang are with the National University of Singapore.

Zhi Ding is with the University of California, Davis.

Shuguang Cui is with Texas A&M University.

This work is supported in part by the National University of Singapore Research Grant R263-000-B46-112. The work of S. Cui is supported in part by DoD with grant HDTRA1-13-1-0029, by NSF with grants ECCS-1508051, CNS-1343155, ECCS-1305979, and CNS-1265227, and by grant NSFC-61328102. The work of Z. Ding is supported in part by the NSF Grant CNS-1443870.

INTRODUCTION

Decades of exponential growth in commercial data services have ushered in the so-called big data era, to which the expansive mobile wireless network is a critical data contributor. As of 2014, the global penetration of mobile subscribers reached 97 percent, producing a staggering 10.7 exabytes (10.7×10^{18}) of mobile data worldwide. The surge of mobile data traffic in recent years is mainly attributed to the popularity of smartphones, phone cameras, mobile tablets, and other smart mobile devices that support mobile broadband applications, such as online music, video, and gaming, as shown in Fig. 1. With a compound annual growth rate of over 40 percent, it is expected that mobile data traffic will increase by 5 times from 2015 to 2020.

In addition to the vast amount of wireless source data, modern wireless signal processing often amplifies the system's pressure from big data in pursuit of higher performance gain. For instance, multiple-input multiple-output (MIMO) antenna technologies are now extensively used to boost throughput and reliability at both mobile terminals (MTs) and base stations (BSs) of high-speed wireless services. However, this also increases the system data traffic to be processed in proportion to the number of antennas in use. Moreover, the fifth generation (5G) wireless network presently under development is likely to migrate the currently hierarchical BS-centric cellular architecture to a cloud-based layered network structure, consisting of a large number of cooperating wireless access points (APs) connected by either wireline or wireless fronthaul links to a big-data-capable central processing unit (CPU). New wireless access structures, such as coordinated multipoint (CoMP or networked MIMO) [1], heterogeneous network (HetNet) [2], and cloud-based radio access network (C-RAN) [3], are under development to achieve multi-standard, interference-aware, and energy-friendly (green) wireless communications. In practice, the use of cooperating wireless APs could easily generate multiple-gigabits-per-second data from a single user's fronthaul links due to the need for baseband joint processing such that the high traffic load may overwhelm the fronthaul link or the system CPU for signal processing and coordination. Such intensely high system traffic volume, together with the rapidly growing mobile data source volume, surpasses both the processing power improvement speed of our current computing capabilities and the fronthaul/backhaul link rate increase pace of our networking systems. It necessitates a new wireless architecture along with efficient signal processing methods to make wireless systems *scalable* to continued growth of data traffic.

On the other hand, timely and cost-efficient information processing is made possible by the fact that the vast volume of mobile data traffic is not completely chaotic and hopelessly beyond management. Rather, it often exhibits strong *insightful features*, such as user mobility patterns, and spatial, temporal, and social correlations of data contents. These special characteristics of mobile traffic present us with opportunities to harness and exploit big data for potential performance gains in various wireless services. To effectively utilize and exploit these characteristics, they should be identified, extracted, and efficiently stored. For instance, caching popular contents at wireless hot spots could effectively reduce the real-time traffic in the fronthaul links. Additionally, network control decisions,

0163-6804/15/$25.00 © 2015 IEEE

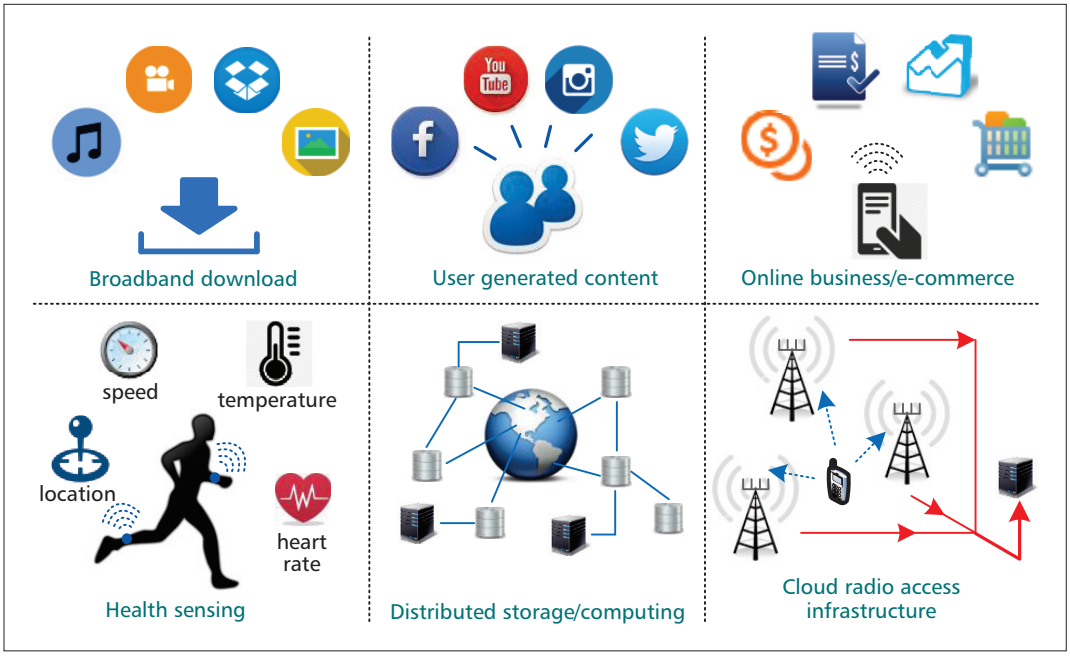


Figure 1. Some example sources of wireless big data traffic.

A hybrid structure could take advantage of the benefits from the two design paradigms, that is, a wireless system that could adaptively choose only local processing at the BS level, only central processing at the CPU level, or parallel processing at both levels, based on, for instance, physical channel conditions and correlations in the data contents.

such as routing, resource allocation, and status reporting, instead of being rigidly programmed, could be made data-driven to fully capture the interplay between big data and network structure. Presently, however, these advanced data-aware features cannot be efficiently implemented in current wireless systems, which are mainly designed for content delivery instead of analyzing and making use of the data traffic.

Bearing in mind the aforementioned challenges and opportunities brought by big data traffic, in this article we address two important problems of wireless communication system design in the big data era:

Q1: What may constitute a *scalable* wireless network architecture for efficient handling of big data traffic?

Q2: How can *big data awareness* be effectively incorporated and utilized to improve wireless system performance?

Specifically, to answer Q1, we introduce a hybrid signal processing paradigm to enable flexible data processing at both the BS/AP and CPU levels, and correspondingly a number of scalable data traffic management techniques to serve the conflicting needs between overall system performance and data processing complexity. For Q2, we first discuss typical big data features and efficient data analytics to extract these features. Next, we introduce a number of big-data-aware signal processing methods and wireless networking structures to capitalize on big data interplay, including mobile cloud processing, crowd computing, software-defined networking, and so on. We also suggest several future research directions for wireless communications in the big data era. Finally, we conclude this article.

Before proceeding to detailed discussions, it is worth mentioning that the scalable network structure and big data awareness considered are both important mechanisms for accommodating mobile big data in future wireless networks, although their focus is different: on the physical and network/application layers, respectively. Nonetheless, the two solutions can also be complementary to each other. For instance, as we discuss later, we can optimize the overall caching strategy by combining long-term cache provisioning (network/application layer) and real-time cache-assisted signal processing (physical layer) techniques. In addition, although this article focuses on the design aspects of cellular networks in the big data era, most of the key enabling mechanisms for mobile big data processing are also applicable to other wireless networking structures, such as WLANs and HetNets. Some representative system designs are also discussed in this article.

Scalable Wireless Big Data Traffic Management

A Hybrid Network Structure

Neither the current cellular systems nor the next-generation cloud-based radio access network (C-RAN) [3] under development were designed to provide a scalable solution for the arrival of the big data era. The current 3G and 4G cellular systems exemplify a BS-centric design, in which a BS bears much of the responsibility for radio access, baseband processing, and radio resource control execution to serve the mobile users in its vicinity. To meet the fast growing mobile data service demand, a smaller cell size is commonly used to improve frequency reuse, which may generate complex and severe inter-cell interference. Furthermore, small cells can also be costly due to densely deployed BSs. The cloud-centric network proposed for 5G mitigates inter-cell interference by centralized signal processing and reduces the unit cell deployment cost by moving computations to the cloud. At the same time, only inexpensive relay-like remote

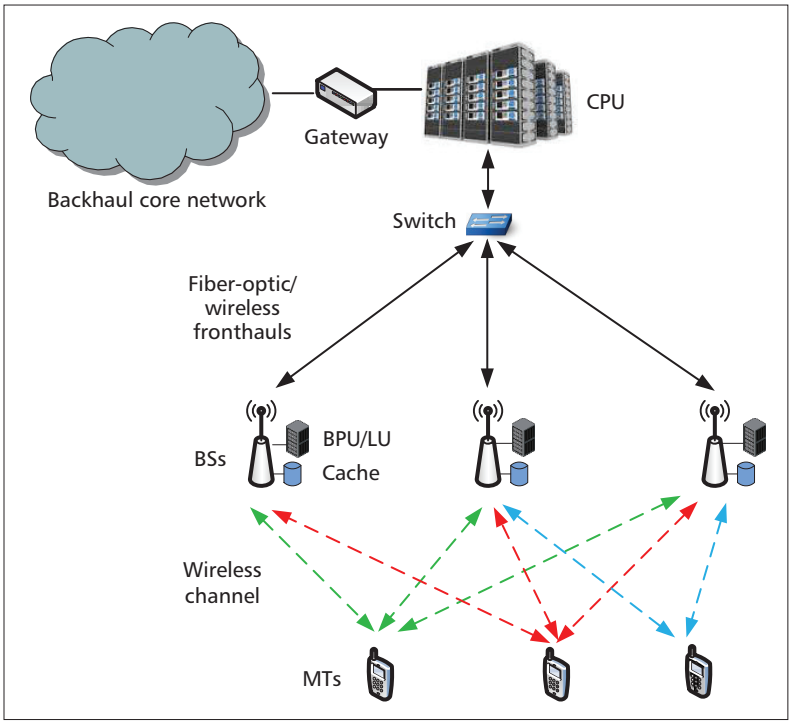


Figure 2. A hybrid CPU-BS processing network structure.

radio heads (RRHs) are used for RF-level wireless access. However, such a fully centralized scheme may be overwhelmed by the huge wave of data traffic beyond its fronthaul link capacities and computational power.

Alternatively, a *hybrid* structure could take advantage of the benefits from the two design paradigms, that is, a wireless system that could adaptively choose only local processing at the BS level, only central processing at the CPU level, or parallel processing at both levels, based on, for instance, physical channel conditions and correlations in the data contents. We thus consider such a generic network structure shown in Fig. 2, which mainly inherits the skeleton of C-RAN, but has integrated several programmable modules to carry out intelligent signal processing at the BS level. In the RAN, mobile users could be served simultaneously by multiple BSs, where each BS is equipped with multiple antennas and is linked to the CPU via high-speed fiber/wireless fronthauls for exchanging user data and control signals. The CPU is further connected to the backhaul core network for external content access. In the proposed hybrid network structure, baseband processing units (BPUs) are available at both the BSs and CPU, which enable user message encoding/decoding at both levels. In addition, learning units (LUs) are installed for data traffic analytics, the functions of which are detailed later. Caches are also installed at the BSs and CPU to save the fronthaul bandwidth consumed for frequent retransmissions of popular contents.

Before entering the discussion of hybrid signal processing models, it is worth mentioning that applicable fronthaul data management methods are directly constrained by the fronthaul technologies in use. Specifically, the system could choose between optical analog and optical/wireless digital fronthaul technologies. Optical analog modulation using RF signal as input is commonly referred to as radio over fiber (RoF). Alternatively, analog RF input signal could also be quantized and encoded into binary codewords for digital wireless or fiber optic communication (DFC). In practice, RoF is simpler and less expensive than DFC. Furthermore, it also exhibits lower processing delays and better interoperability with multiple wireless standards, for example, 3G, Long Term Evolution (LTE), and WiFi, as it is oblivious to user codebooks and wireless modulation schemes. However, its limitations are also evident, for example, susceptibility to noise and signal distortion, and difficulty of synchronization. More importantly, available signal processing techniques for fronthaul traffic management using RoF are less sophisticated, generally limited to simply transforming or removing certain parts of the received RF signals, for example, sub-channel and antenna selection methods. In contrast, DFC could be combined with data compression, opportunistic decoding, and many other advanced digital signal processing techniques. In the following, we mainly focus on data traffic management methods using digital fronthaul.

Hybrid Signal Processing Models

The wireless/fiber-optic link has its own throughput limit. For instance, a commercial fiber optic link normally operates at a link rate in the order of 10 Gb/s for digital communication over a single optical carrier. Transmission rates beyond the link rate capacity may lead to severe signal distortions, and consequently poor decoding performance. Therefore, the system performance must be optimized under the fronthaul link capacity constraints. With respect to the hybrid network structure in Fig. 2, we now introduce some scalable fronthaul data management techniques in three major categories:

Data Compression — The uplink direction would require unlimited fronthaul capacity to transmit an analog RF signal perfectly without any distortion from a BS to the CPU. An analog signal could be more efficiently transmitted through the fronthaul if it is quantized and compressed into binary codewords. From an information theoretic perspective, the effect of data compression could be modeled as a test channel (often Gaussian for simplicity of analysis) for which uncompressed signals as the input and compressed signals are the output. The compression design is equivalent to setting the variance of the additive compression noise [4]. To achieve successful compression, the encoder needs to transmit to the decoder at a rate at least equal to the mutual information between the input and the output over the Gaussian test channel. Intuitively, a tighter fronthaul capacity constraint would therefore require coarser compression with more compression noise. Existing compression designs in general take the following approaches.

Joint compression across different BSs: When multiple BSs compress and forward their received signals to the CPU in the uplink, the

compression design requires setting the covariance of the compression noises across different BSs. A common objective is to maximize the information rate under the fronthaul capacity constraints. In this setting, distributed Wyner-Ziv lossy compression can be used at the BSs, exploiting signal correlation across the multiple BSs [4]. The distributed Wyner-Ziv compression scheme is shown to yield significant capacity gains over independent quantization methods, especially in the low backhaul capacity region [4]. Similar data compression methods could also be applied in the downlink. Interestingly, it has been shown in [3] that downlink compression and multi-user precoding design (for interference mitigation) could be designed separately without compromising maximum system throughput, which is achieved by an optimal but much more complicated joint compression-precoding design.

Independent BS-level compression: The practical implementation of distributed Wyner-Ziv compression is difficult mainly because of the high complexity in determining the optimal joint compression codebook and the joint decompressing/decoding at the CPU. Accordingly, independent compression methods, where the quantization codebook at a BS is only determined by its local signal-to-noise ratio (SNR), can be used to reduce the computational complexity and the signaling exchange overhead in the fronthaul.

Uniform scalar quantization: Even when using independent BS-level compression, real-time computation and exchange of quantization codebooks using the information-theoretical source coding approaches are often difficult to realize in practice. Instead, simple uniform scalar quantization methods compatible with analog-to-digital (A/D) modules are proposed to reduce the implementation cost [5]. Interestingly, it is shown in [5] that the achievable rate using simple uniform scalar quantization in fact performs closely to that of the Gaussian test channel model. This indicates that efficient fronthaul capacity usage is achievable in practical systems with simple quantization methods.

BS-level encoding/decoding: Besides acting as relays to compress/decompress and forward the user signals, BSs with advanced baseband processing capabilities could also encode/decode the received messages to further improve the system performance under stringent fronthaul capacity constraints.

Partial cooperation: In the uplink, one direct method to reduce fronthaul traffic is to limit the number of cooperating elements when serving mobile users. Many sparsity inducing optimization methods could be applied to satisfy a certain quality of service level using minimum numbers of sub-channels, antennas, or cooperating BSs. In the downlink, similar sparse precoding methods could be studied to optimize precoders by jointly maximizing the user utilities (e.g., data rate) and minimizing the total number of data streams in the fronthaul [6].

Distributed encoding/decoding: Distributed decoding allows the BSs to decode user messages locally without forwarding quantized signals to the CPU. For instance, [4] considers a rate-splitting approach to divide an MT's message into two parts, where one part is decoded locally by the serving BS, and the remainder is compressed and jointly decoded by the CPU. In another case, [7] proposes an opportunistic hybrid decoding method, where a user's message is either decoded locally at a BS when its SNR is sufficiently high, or jointly decoded by the CPU based on signals forwarded from a subset of cooperating BSs when the SNR at each individual BS is too low. Note that the locally decoded user messages can be used to cancel their interference to the received RF signals at the BSs, which can effectively reduce the amount of data transmitted to the CPU over the fronthaul links. In the downlink case, BSs could encode and modulate the baseband symbols to RF signals before transmitting them to the MTs. Therefore, instead of transmitting complete signal waveforms (or waveform samples) to the BSs, the CPU could save fronthaul bandwidth by separately transmitting the information symbols and the beamforming vectors, while leaving RF modulation to the BSs.

To show the performance advantage of the hybrid signal processing model, we present a numerical example in Fig. 3 to compare the throughput performance among the BS-centric, cloud-centric, and hybrid processing networks. Let us consider a cellular uplink, where three MTs transmit over orthogonal sub-channels, each with 100 MHz bandwidth. Besides, each fronthaul link has 1.2 Gb/s capacity. The decoding methods of the three networks are described as follows.

BS-centric network: BS_1 decodes the messages from MT_1 and MT_2, and BS_2 decodes the message from MT_3. Then both the BSs send the decoded messages to the CPU.

Cloud-centric network: Both BSs compress the received signals using the scalar quantization method considered in [5]. They then forward the compressed signals to the CPU for joint decoding. In particular, each user is equally allocated 400 Mb/s fronthaul bandwidth at a BS to transmit its compressed signal;

Hybrid processing network: BS_1 and BS_2 first decode the messages from MT_1 and MT_3, respectively, before transmitting the decoded messages to the CPU. Meanwhile, each BS uses the remaining fronthaul bandwidth to compress and transmit the signal from MT_2 to the CPU for the joint decoding of MT_2's message.

From the aforementioned network setups, we calculate in Fig. 3 the achievable user data rates under a random channel realization, and compare the common throughput performance (the minimum data rate among the three users) in different cases. We can see that the BS-centric network achieves the lowest common throughput, due to the low data rate of the cell edge user MT_2, which is only 210 Mb/s. The cloud-centric network slightly improves the data rate of MT_2 and hence the common-throughput to 230 Mb/s, thanks to its joint processing gain. However, the data rates of MT_1 and MT_3 are severely degraded, since the limited fronthaul capacity introduces high compression noises to the useful signals. The hybrid processing network achieves the highest common throughput (301 Mb/s)

Besides acting as relays to compress/decompress and forward the user signals, BSs with advanced baseband processing capabilities could also encode/decode the received messages locally to further improve the system performance under stringent fronthaul capacity constraints.

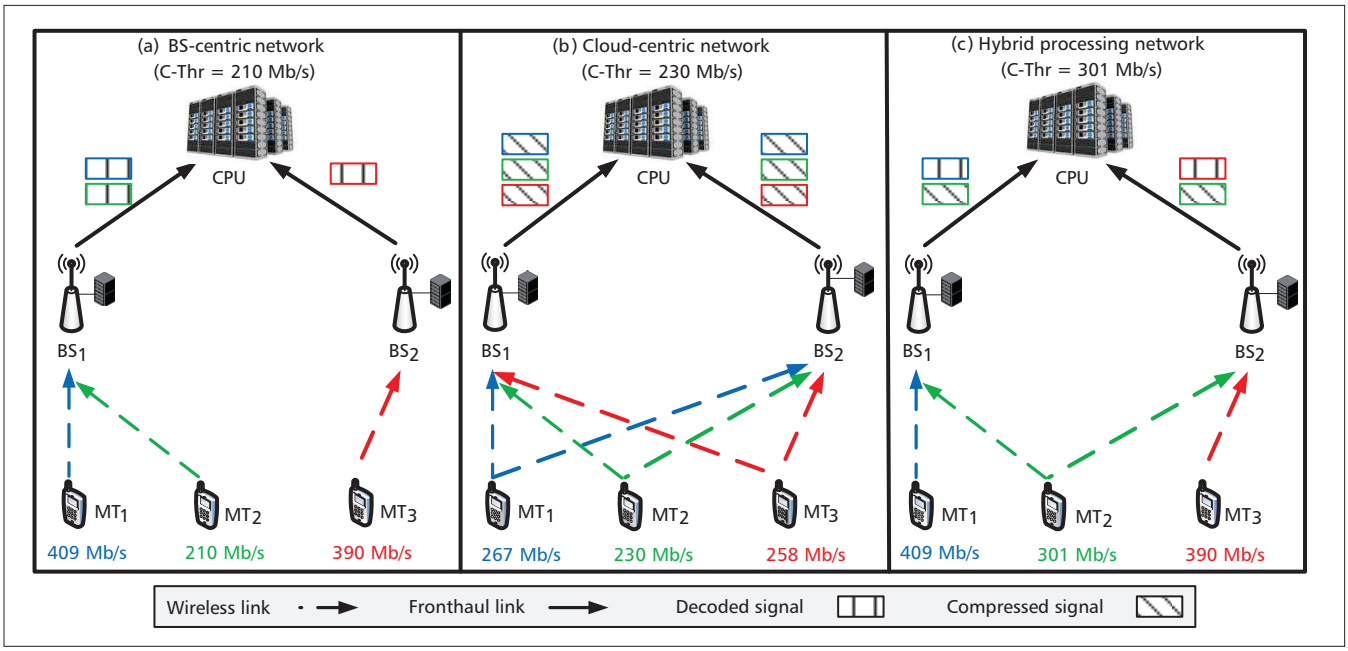


Figure 3. Throughput performance comparison of three network structures: a) BS-centric; b) cloud-centric; c) hybrid processing networks. The common-throughput (C-Thr) achieved by the three networks are 210, 230, and 301 Mb/s, respectively.

among the three schemes that we considered, which is 43 and 31 percent higher than those of the BS-centric and cloud-centric networks, respectively. Compared to the cloud-centric networks, by decoding the messages from MT_1 and MT_3 at the BS level, the hybrid processing network has a larger fronthaul bandwidth to spare for transmitting MT_2's signals to the CPU with more refined compression, thus achieving a higher joint processing gain.

Cache-Assisted Processing — In downlink transmission, caching at the BSs is cost effective to reduce real-time traffic on the fronthaul, thereby enabling significant improvement in overall C-RAN performance. Cache-assisted wireless resource allocation is a cross-layer approach that incorporates the status of application-layer data flow in wireless physical-layer design. As an illustrative example in Fig. 4, BS_2 serves two requests from the two MTs, whereas caches of the other two BSs are empty. Although MT_1 is closer to BS_1 with a better wireless channel condition, the maximum downlink data rate is only 1 unit/s if BS_1 is selected to transmit directly, due to the constraint of link congestion between BS_1 and the CPU. Instead, the CPU could select BS_2 to send the cached contents to MT_1 at a rate of 2 units/s, which has an end-to-end data rate not constrained by the congestion level of the CPU-to-BS_2 link. On the other hand, MT_2 could be served by two cooperating BSs (BS_2 and BS_3) with an improved wireless channel gain from coordinated beamforming. In particular, the CPU only needs to transmit the content requested by MT_2 to BS_3 before the cooperative transmissions of the two BSs. Thanks to such wireless cooperation, MT_2 could achieve a higher data rate at 3 units/s.

In a more general setting, caches could be located not only at the BSs, but also at the routers and the CPU. Furthermore, distributed caching could also be adopted at MTs to allow mobile users to serve popular contents requested by nearby peer users in a device-to-device (D2D) manner. We could foresee cache-assisted resource allocation becoming a key enabling factor of significant bandwidth saving, since frequent overlapping of requested objects will occur as the volume of mobile traffic increases. However, it also becomes a more challenging problem to optimize system-wide resource allocation due to the interleaving among cache placement, wireless interference, routing, and the combinatorial nature of node selection in the wireless network. A more comprehensive understanding of the design trade-off remains open for future study.

Another interesting topic in cache-assisted resource allocation is cache provisioning for popular contents to reduce the real-time backhaul traffic. In particular, cache provisioning addresses the questions of what, where, and when to cache in the wireless infrastructure. In this case, accurate knowledge of the mobile user demand profiles is key to efficient cache provisioning. The extraction of user demand profiles from mobile data traffic is performed by wireless big data analytics, discussed in the next section.

Developing a Big-Data-Aware Wireless Network

Instead of viewing mobile big data as a pure burden, in this section we investigate the potential performance gain from developing a big-data-aware intelligent wireless network. However, its efficient operation relies on in-depth knowledge of wireless big data traffic characteristics. As most such characteristics are implicit, we first introduce data-analytical methods necessary to

extract these big data features. We then discuss how to leverage these big data characteristics in designing wireless networks to capitalize from the mobile big data traffic.

Useful Mobile Big Data Features and Applications

There is clearly a strong connection between wireless service usage and human behavioral patterns in the physical world. For this reason, wireless data traffic contains strong correlative and statistical features in various dimensions, including time, location, the underlying social relationship, and so on. On one hand, mobile traffic has strong *aggregate features*. For instance, there are severe load imbalances spatially and temporally, such that presently, 10 percent of "popular" BSs carry about 50~60 percent traffic load. The peak traffic volume at a given location is much higher than the regular average. These aggregate features could be exploited to reduce real-time fronthaul/backhaul traffic and improve wireless network efficiency. Example applications include cell planning according to geographical data usage distribution, peak load shifting via load-dependent pricing, and cache provisioning based on aggregate demand profile, among others.

On the other hand, each mobile user's data usage profile also exhibits a unique set of *individual features*, such as mobility pattern, preference of various data applications, and service quality requirements. For instance, a mobile user's trajectories often consist of a very limited number of frequent positions and quasi-repetitive patterns. Besides, the recent popularity of mobile social networking interconnects seemingly uncorrelated individual data usages into a unified social profile, thereby presenting a novel perspective to analyze the mobile traffic pattern. These individual and social features are useful for system operators to personalize and improve wireless service quality. Many intelligent data-aware services could be provided according to user profiles. Examples include resource reservation in handoff using location prediction, context-aware personal wireless service adaptation, and mobility-based routing and paging control.

Big Data Analytical Tools

The ability to acquire, analyze, and exploit mobile traffic characteristics can be accomplished by specially designed learning units (LUs) installed at both the BSs and CPUs (Fig. 2). Their core enabling factors are embedded data analytical algorithms. Some commonly used algorithms for wireless traffic analysis and their main applications to wireless communications are classified as follows and summarized in Table 1.

Stochastic Modeling — Stochastic modeling methods use probabilistic models to capture the *explicit* features and dynamics of the data traffic. Commonly used stochastic models include order-the K Markov model, hidden Markov model, geometric model, time series, linear/nonlinear random dynamic systems, and so on. For example, Markov models and Kalman filters are widely used to predict user mobility and service requirements [8]. The collected user data are often used for parameter estimation of stochastic models, such as estimating the transition probability matrix of a Markov chain.

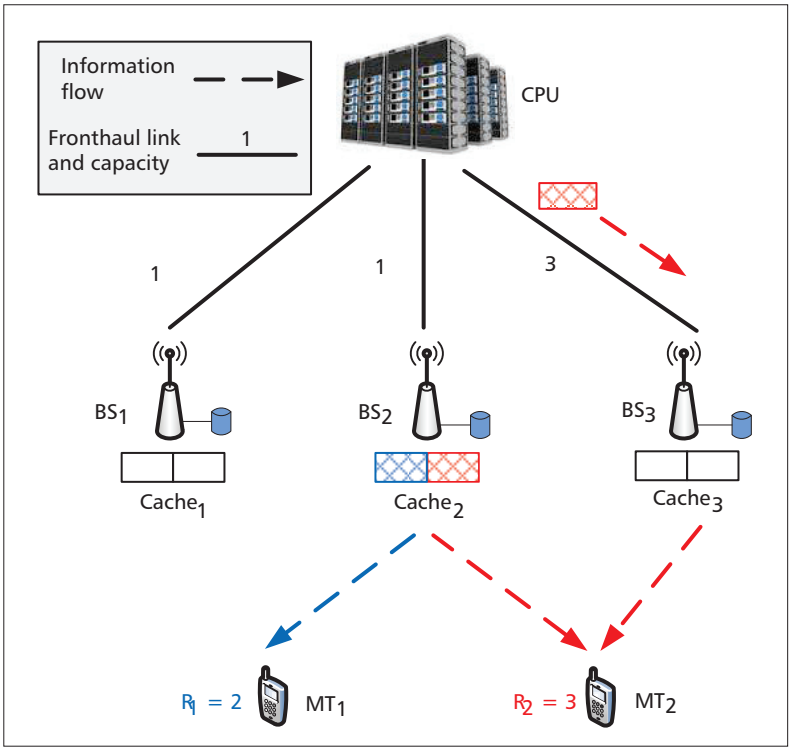


Figure 4. Downlink cache-assisted wireless signal processing.

Data Mining — Data mining focuses on exploiting the *implicit* structures in the mobile data sets. Also taking the mobility prediction problem as an example, an individual user's mobility pattern could be extracted and discovered by finding the most frequent trajectory segments in the mobility log. Prediction could be made accordingly by matching the current trajectory to the mobility profile. Clustering is another useful technique to identify the different patterns in the data sets. It is widely used in context-aware mobile computing, where a mobile user's context and behavioral information, such as sleeping and working, are identified from wireless sensing data for providing context-related services [9].

Machine Learning — The main objective of machine learning is to establish a functional relationship between input data and output actions, thus achieving *auto-processing* capability for unseen patterns of data inputs. Among the many useful techniques in machine learning applied to wireless communications, classification (determining the type of input data), and regression analysis (data fitting) are two common methods, with applications including context identification of mobile usage and prediction of traffic levels (classification), or fitting the distributions of trajectory length, mobile user location, and channel holding times (regression). Also, reinforcement learning, such as Q-learning [10], is useful for taking proper real-time actions

Subjects	Models/algorithms	Example wireless applications
Statistical modeling	Markov models, time series, geometric models, Kalman filters	Mobility prediction, resource provision, device association/handoff prediction
Data mining	Pattern matching, text compression, clustering, dimension reduction	Mobility prediction, social group clustering, context-aware processing, cache management, user profile management
Machine learning	Classification algorithms, neural network, regression analysis	Context identification, traffic prediction, fitting trajectory length, user location and the channel holding time
	Dimension reduction algorithms: PCA, PARAFAC, Tucker3	User data compression/storage, traffic feature extraction, blind multiuser detection
	Q-learning	Handoff and admission controls
	Primal/dual decomposition, ADMM	Distributed routing/rate control and wireless resource allocation
	Online convex optimization, stochastic learning	Online mobility predictions, handoffs, and resource provisioning
	Active learning, deep learning	Incomplete/complex mobile data processing

Table 1. Summary of common wireless big data analytic tools and example applications.

to maximize certain long-term rewards. A typical example is making the handoff and admission control decision (action), given the current traffic load (state) and incoming new requests (event), in which the reward could be evaluated against the reduction of dropped calls or failed connections.

Large-Scale Data Analytics: Wireless big data poses many challenges to the aforementioned conventional data analytical methods due to its high volume, large dimensionality, uneven data qualities, and the complex features therein. To improve signal processing efficiency, one can combine the following complexity reduction techniques with the conventional data analytical tools for large-scale data processing.

Distributed optimization algorithms: such as primal/dual decomposition and alternating direction method of multipliers (ADMM), are very useful to decouple large-scale statistical learning problems into small subproblems for parallel computations so as to relieve both the computational burden at the CPU and the bandwidth pressures to the fronthaul/backhaul links.

Dimension reduction: methods are useful to reduce the data volume to be processed while capturing the key features of big data. Among various methods, principal component analysis (PCA), along with its many variants, is the mostly used method today. In addition, tensor decomposition methods are also popular in mobile data processing, and seek to approximately represent a high-order multi-way array (tensor) as a linear combination of outer products of low-order tensors. By doing so, the hardware requirement and cost for storing the high-order arrays of mobile data could be reduced.

Other advanced learning methods: could be used to handle incomplete or complex data sets. Interesting examples include active learning, which deals with partially labeled data sets; online learning for responding in real time to sequentially received data; stochastic learning, which makes a decision periodically in each time interval; and deep learning for modeling complex behaviors contained in a data set.

Big-Data-Aware Wireless Network

Once identified and extracted, data characteristics could be used to improve wireless service quality and generate new mobile applications. For simplicity of illustration, we have postulated in Fig. 5 a structure of a big-data-aware wireless network, consisting of several mutually complementary components that enable data-driven mobile services the functionalities of which are described below.

Data-aware cache management: For quick access under high traffic volume, cached contents need to be carefully categorized, compactly organized, and quickly updated. Many types of content objects, such as music and video files, are embedded with metadata labels that describe the properties of the contents, from which the data contents could be well classified. By classifying data into a number of sub-classes based on contents, such as sports videos and news pictures, LUs could achieve more accurate evaluation of the content popularity by jointly considering its own access count and the total access count of its type, which reflects the average frequency of potential future access. Accordingly, popular contents are continuously cached, while unpopular contents are removed regularly to maximize the effective system bandwidth given limited cache size.

Crowd computing: Mobile users of similar interests could share their resources with peers in their vicinity, either with or without taking advantage of the wireless infrastructure. For instance, a complete 3D street view could be generated by a BS from relevant photos contributed by users from different angles. Meanwhile, when MT-to-BS connection is unavailable, an MT could ask for assistance from its neighboring MTs to share available contents and applications, or to even act as relays to the cellular network, and so on. Such an idea is explored in [11], where a crowd-enabled data transmission mechanism is proposed to let mobile users assist with the data dissemination of other users. In particular, it makes use of personal social information and market incentives to enhance the "willingness" of mobile users to act as data brokers of others such that a higher chance of successful data delivery could be achieved. Essentially, this peer-to-peer nature of crowd computing exploits user mobility and spatial correlation of data traffic, which also helps reduce the conventional cellular traffic to and from the wireless infrastructure.

Mobile cloud processing: Multiple interconnected C-RANs constitute a mobile cloud, which

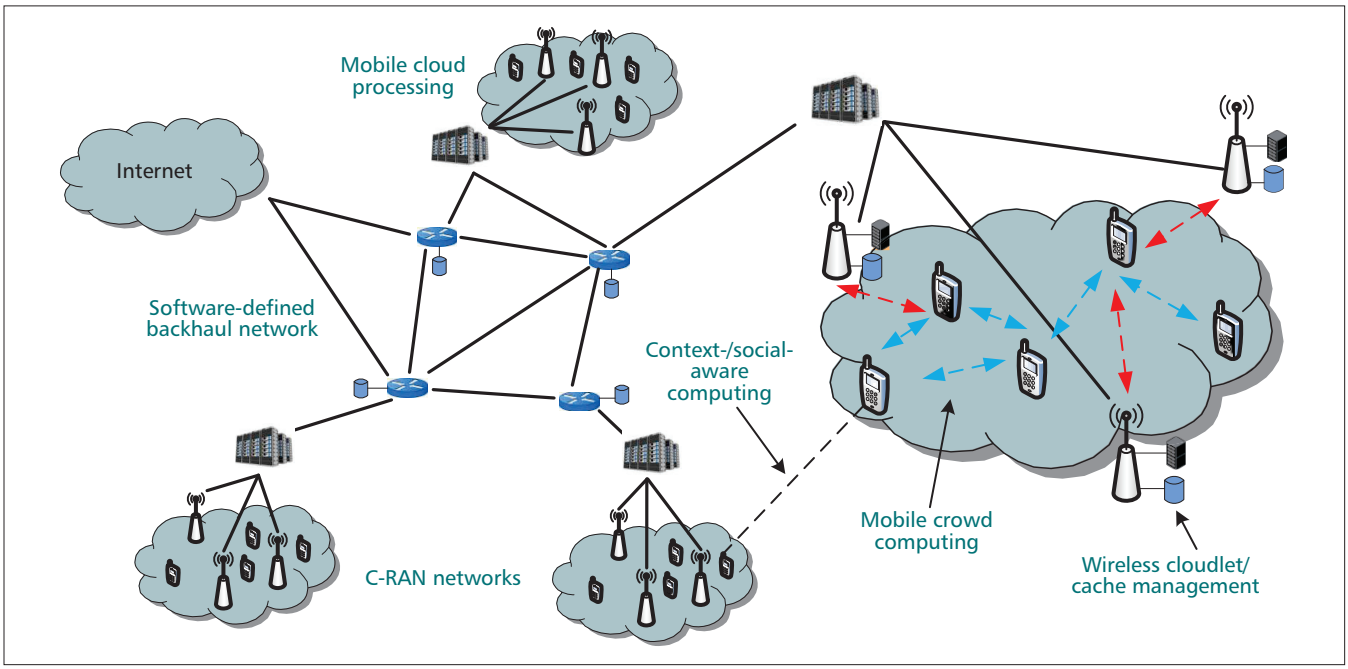


Figure 5. An illustrative structure of a big-data-aware wireless network.

could optimize the wireless services based on knowledge with respect to mobile traffic patterns, especially when user mobility spans across different C-RAN clusters. For instance, based on the mobility pattern of an MT, a CPU could reserve channel resource in advance and prefeed the contents to the BSs along the anticipated MT's route. As such, chunks of contents could be sent from different BSs to achieve seamless handoffs. Similarly, aggregate characteristic behavior of data traffic could also be used to allocate resources such as bandwidth and cache space to some popular locations ahead of some real-time events. This approach could evidently reduce connection time, delay jitter, and burden of real-time traffic bursts on both cellular fronthaul and backhaul.

Wireless Cloudlet: The concept of cloudlets introduced in [12] defines a self-organized light cloud with limited storage and computing power installed at the BSs to enhance their local data processing capability. The deployment of cloudlets could effectively reduce the packet round-time delay by an order of magnitude. A cloudlet may be owned by the network operator but leased to commercial clients for improving performance of delay-sensitive applications, such as online gaming. Besides, a cloudlet could also allow commercial clients to access local cache to provide better location-based services. For instance, an advertising company could send to its subscribers in the vicinity the latest deals based on the information posted by local stores and queries made by prospective customers. With cloudlet, real-time traffic in the backhaul network could also be largely reduced, since many services could be provided locally instead of burdening the core network.

Context/Social-Aware Processing: Context/social-aware computing is an emerging paradigm for exploiting complex data characteristics besides conventional user profiles such as mobility pattern and demand distribution [13]. The idea of context-aware computing is to provide personalized services adaptive to the MT's real-time "context," such as traveling, working, and recreation, either directly reported by the MT or inferred from various available data. Social computing, on the other hand, calls for wireless resource allocation to follow closely the interaction within and among social groups [13]. Conceptually, a social group is a subset of users that share some similar interests, professions, hobbies, life experiences, and so on. In general, a social group has unique "eigenbehaviors" such that the group members require and generate similar data contents. The knowledge of a social community's composition, activities, and interests could be used to improve the wireless services for the targeted social group members.

Software-defined network (SDN): An SDN replaces conventional hardware-configured routing and forwarding devices with software programmable units. In particular, it decouples the user's data plane (U-plane) from the control and management plane (C-plane) such that the network is managed by a central controller while the underlying devices are only responsible for simple functions such as packet forwarding. This decoupling provides unprecedented flexibility to network traffic management, where packet forwarding decisions may now be programmed based on many new considerations such as quality of service (QoS) requirement, application type, and payload length, in addition to the conventional destination-oriented and distance-based metrics. For SDN-enabled wireless networks, [14] proposes a flow-based resource management framework in C-RAN, where the packet routing in the backhaul network and beamforming design in the wireless access network are jointly optimized based on an individual data flow's

BS-level caching is expected to play an important role in future wireless big-data processing, due to its simplicity, low cost, and natural integration with big-data analytical tools. However, research on cache-assisted wireless resource allocation is still in its infancy.

source-destination pair, wireless channel condition, backhaul link capacity, user QoS requirements, and so on. In the case of WLAN networks, [15] introduces an SDN-based enterprise WLAN framework named Odin, which is built with programable functions, global knowledge of network status, and direct control of network devices. The SDN-based system makes many difficult or costly tasks in conventional WLANs easier and less expensive, including seamless user handoffs, global load balancing, and hidden terminal problem mitigation.

Future Research Directions

In the mobile big data era, wireless system designs contain rich research problems of important applications and impact that are yet to be studied. Beyond the many research issues that arise among the number of topics we have discussed so far, in this section we highlight several interesting research topics that we find particularly exciting.

Reduced-Complexity Fronthaul Processing

In many data compression proposals, real-time calculation of the optimal compression noise covariance matrix is often impeded by the large number of fronthaul capacity constraints and the non-convex nature of many fronthaul-constrained problems. The problem is further exacerbated by the difficulty in generating practical joint compression codebooks based on the obtained covariance matrix. Therefore, sub-optimal but practical compression schemes, such as scalar quantization, should be given more consideration in future study of fronthaul-constrained compression design. Similarly, CPU-level encoding and decoding also suffers from high computational complexity on large-scale multi-user detection and the combinatorial nature of many limited cooperation schemes, such as optimal antenna, relay, modulation and coding combinations, as well as BS selection. It therefore calls for practical complexity reduction algorithms that are truly scalable to the number of mobile users and network entities.

Cache-Assisted Wireless Resource Allocation

BS-level caching is expected to play an important role in future wireless big data processing due to its simplicity, low cost, and natural integration with big data analytical tools. However, research on cache-assisted wireless resource allocation is still in its infancy. For cache-assisted cellular networks with BS-level caching, currently there is a shortage of both concrete theoretical analysis on the capacity gain of cache-assisted processing and practical optimization frameworks for cache-assisted resource allocation. Furthermore, effective and optimized integration of various identified big data characteristics in cache-assisted network design is an interesting problem that awaits future investigation.

Distributed Network Traffic Control

In large-scale wireless networks, distributed control/computing algorithms could be integrated to alleviate computational complexity of the CPU, reduce backhaul traffic volume, and mitigate the risk of single-point failures without compromising overall system performance. Due to the programmability of SDN-enabled system infrastructure, distributed control mechanisms could be implemented with much better flexibility and lower cost. However, the feasibility and complexity reduction of distributed algorithms are often constrained by underlying structure problems, including coupling constraints in the backhaul, partial knowledge of data traffic, and so on. Distributed control, or a mixed centralized and decentralized control framework, is a promising work direction toward a future wireless networking design supporting mobile big data. Additionally, SDN-based design may also incorporate distributed caching (at BSs and routers) to enhance the efficiency of the routing decision.

Mobile Data Security and Privacy

Harvesting over large mobile data sets and data analytics naturally give rise to concerns with respect to data security and privacy. In a cloud-based wireless network, a large amount of data is stored in the fronthaul/backhaul network either for customers' personal use or as a commercial database for future analytical purposes. The system operators or commercial entities that collect the user data should be responsible for data security and privacy. For example, personal data should be only available for legitimate and authenticated users. Similarly, data integrity should be guaranteed such that no data is lost or modified by unauthorized entities. Furthermore, it is also important to maintain confidentiality of user data either when they are in storage or during processing. It is therefore important to develop secure but efficient data processing and storage methods. Promising security measures may include privacy-aware distributed data storage and decentralized processing, which aim to maintain local data confidentiality.

Conclusions

This article addresses challenges and opportunities that we face in the era of wireless big data. We review state-of-the-art signal processing methods and networking structures that may allow us to effectively manage and in fact take advantage of wireless big data traffic. We outline the major obstacles of big data signal processing and network design with respect to the scale of problem size and the complex problem structures. Nevertheless, research on big data for wireless communications and networking is not only promising but also inevitable in light of the continuing data volume explosion. We also suggest several interesting research problems aimed at stimulating future wireless research innovations in the big data era.

References

[1] R. Irmer *et al.*, "Coordinated Multipoint: Concepts, Performance, and Field Trial Results," *IEEE Commun. Mag.*, vol. 49, no. 2, Feb. 2011, pp. 102–11.

[2] J. G. Andrews, "Seven Ways that Hetnets Are a Cellular Paradigm Shift," *IEEE Commun. Mag.*, vol. 51, no. 3, Mar. 2013, pp. 136–44.
[3] S. H. Park *et al.*, "Joint Precoding and Multivariate Backhaul Compression for the Downlink of Cloud Radio Access Networks," *IEEE Trans. Signal Processing*, vol. 61, no. 22, Nov. 2013, pp. 5646–58.
[4] A. Sanderovich *et al.*, "Uplink Macro Diversity of Limited Backhaul Cellular Network," *IEEE Trans. Info. Theory*, vol. 55, no. 8, Aug. 2009, pp. 3457–78.
[5] L. Liu, S. Bi, and R. Zhang, "Joint Power Control and Fronthaul Rate Allocation for Throughput Maximization in OFDMA-Based Cloud Radio Access Network," to appear, *IEEE Trans. Commun.*, http://arxiv.org/abs/1407.3855
[6] M. Hong *et al.*, "Joint Base Station Clustering and Beamformer Design for Partial Coordinated Transmission in Heterogeneous Networks," *IEEE JSAC*, vol. 31, no. 2, Feb. 2013, pp. 226–40.
[7] R. Zhang and J. M. Cioffi, "Exploiting Opportunistic Multiuser Detection in Decentralized Multiuser MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, Aug. 2011, pp. 2474–85.
[8] I. F. Akyildiz and W. Wang, "The Predictive User Mobility Profile Framework for Wireless Multimedia Networks," *IEEE/ACM Trans. Networks*, vol. 12, no. 6, Dec. 2004, pp. 1021–35.
[9] A. Krause, A. Smailagic, and D. P. Siewiorek, "Context-Aware Mobile Computing: Learning Context-Dependent Personal Preferences from a Wearable Sensor Array," *IEEE Trans. Mobile Computing*, vol. 5, no.2, Feb. 2006, pp. 113–27.
[10] K. L. A. Yau, P. Komisarczuk, and P. D. Teal, "Reinforcement Learning for Context Awareness and Intelligence in Wireless Networks: Review, New Features and Open Issues," *J. Network and Computer Applications*, vol. 35, no. 1, Jan. 2012, pp. 253–67.
[11] B. Guo *et al.*, "Mobile Crowd Sensing and Computing: When Participatory Sensing Meets Participatory Social Media," to appear, *IEEE Commun. Mag.*, 2015.
[12] M. Satyanarayanan *et al.*, "The Case For VM-Based Cloudlets in Mobile Computing," *IEEE Pervasive Computing*, vol. 8, no. 4, Oct.–Dec. 2009, pp. 14–23.
[13] P. Lukowicz, A. Pentland, and A. Frescha, "From Context Awareness to Socially Aware Computing," *IEEE Pervasive Computing*, vol. 11, no. 1, Jan.-Mar. 2012, pp. 32–41.
[14] W. C. Liao *et al.*, "Min Flow Rate Maximization for Software Defined Radio Access Networks," *IEEE JSAC*, vol.32, no.6, June 2014, pp. 1282–94.
[15] L. Suresh *et al.*, "Towards Programmable Enterprise WLANs with Odin," *Proc. ACM HotSDN*, Helsinki, Finland, Aug. 2012, pp. 115–20.

BIOGRAPHIES

SUZHI BI [S'10, M'14] (bsz@nus.edu.sg) received his B.Eng. degree in communications engineering from Zhejiang University, Hangzhou, China, in 2009. He received his Ph.D degree in information engineering from the Chinese University of Hong Kong in 2013. He is currently a research fellow in the Department of Electrical and Computer Engineering at the National University of Singapore. His current research interests include wireless information and power transfer, medium access control in wireless networks, and smart power grid communications.

RUI ZHANG [S'00, M'07, SM'15] (elezhang@nus.edu.sg) received his Ph.D. degree from Stanford University in 2007. He is now an assistant professor with the ECE Department of the National University of Singapore. His current research interests include energy-efficient and energy-harvesting-enabled wireless communications, wireless information and power transfer, and data-aware wireless networking. He was the recipient of the 6th IEEE ComSoc Asia-Pacific Best Young Researcher Award in 2011, and the co-recipient of the IEEE Marconi Prize Paper Award in Wireless Communications in 2015. He is now an Editor for *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Signal Processing*, and *IEEE Journal on Selected Areas in Communications*. He was selected as a Thomson Reuters Highly Cited Researcher in 2015.

ZHI DING [S'88, M'90, SM'95, F'03] (zding@ucdavis.edu) is a professor of electrical and computer engineering at the University of California, Davis. He received his Ph.D. degree from Cornell University in 1990. He was the Technical Program Chair of IEEE GLOBECOM 2006. He was also an IEEE Distinguished Lecturer (Circuits and Systems Society, 2004–2006; Communications Society, 2008–2009). He served as an *IEEE Transactions on Wireless Communications* Steering Committee Member (2007–2009) and its Chair (2009–2011). He received the 2012 IEEE Wireless Communication Recognition Award from the IEEE Communications Society and is a coauthor of *Modern Digital and Analog Communication Systems*, 4th ed. (Oxford University Press, 2009).

SHUGUANG CUI [S'99, M'05, SM'12, F'14] (cui@ece.tamu.edu) received his Ph.D from Stanford in 2005. He is now an associate professor at TAMU. His current research interests are data-oriented large-scale information analysis and system design. He was selected as a Thomson Reuters Highly Cited Researcher and listed in the World's Most Influential Scientific Minds by Sciencewatch in 2014. He was the recipient of the IEEE Signal Processing Society 2012 Best Paper Award. He is a ComSoc Distinguished Lecturer.

It is important to maintain confidentiality of user data when they are either in storage or during processing. It is therefore important to develop secure yet efficient data processing and storage methods. Promising security measures may include privacy aware distributed data storage and decentralized processing, which aim to maintain local data confidentiality.

ACCEPTED FROM OPEN CALL

Privacy and Incentive Mechanisms in People-Centric Sensing Networks

Daojing He, Sammy Chan, and Mohsen Guizani

ABSTRACT

Leveraging on the ubiquity and increasing number of smartphone users, people-centric sensing is a new computing paradigm that enables distributed data collection by voluntary participants, using the rich sensing capabilities of smartphones. In this article we identify the challenges on two important issues for a successful people-centric sensing system, i.e., privacy and incentives, because if either there is no incentive to participate or participants' privacy is invaded, smartphone owners will be reluctant to participate. Then we review some recent works that address these two issues. Finally, we suggest directions for future work on people-centric sensing by describing several open issues.

INTRODUCTION

In recent years, smartphones have been quickly replacing traditional mobile phones. In 2013 worldwide smartphone shipments surpassed one billion units and accounted for 55.1 percent of total mobile phone shipments in the year [1]. Effectively, a smartphone is a handheld computing device with various networking capabilities. It enables its user to access the Internet via WiFi or GSM/3G/4G networks. Moreover, a myriad of third-party applications are available and they can be downloaded to smartphones in a convenient manner. More importantly, smartphones are generally embedded with a set of powerful sensors such as a gyroscope, accelerometer, compass, camera, GPS, and microphone (Fig. 1). These features together enable smartphones to participate in a new sensing paradigm that is often referred to as people-centric sensing or participatory sensing (PS).

In a PS system, smartphone users help to collect data from their surroundings and upload the data to a central processor, where data contributed by different people are processed and delivered to interested third parties. With the great potential of PS, numerous applications and systems have been developed. Some examples include applications that use collected data to generate noise maps [2] and WiFi network coverage, as well as applications for monitoring air pollution and parking availability [3].

A PS system is essentially a wireless sensor network (WSN) formed by ubiquitous sensors. However, compared with traditional WSNs, PS offers a number of advantages. First, there is no need to pre-distribute sensors into the field; smartphones already in the field are recruited to provide services, so the setup cost is lower. Second, smartphones have much greater resources than motes, so sensing tasks become less constrained. Third, due to the inherent mobility of mobile phones, better spatial and temporal coverage can be achieved.

Although PS applications offer a great deal of potential to revolutionize different sectors of our lives, such as social networks, healthcare, environment monitoring, and transportation, there are two issues that might critically hinder their successful deployment. The first issue is the participants' privacy. Different from traditional WSNs, sensing devices are not owned by the entities that will make queries to the collected data. Instead, they are personal devices that are carried by participants all the time. When participants generate reports of their sensed data, inevitably the reports would expose some of their personal and sensitive information, such as physical location, behavior, political views, health-related information, or even income. Participants need to be assured that their privacy is properly protected before they will participate in a PS system. The second issue is incentive. A PS system requires the participants to carry out sensing tasks specified by individual applications. However, carrying out the sensing activities will consume resources of the participants' smartphones, such as energy and computing power. In other words, there is a cost incurred by participants to fulfill the sensing tasks. Clearly, a smartphone owner would be reluctant to participate unless sufficient incentive is provided.

Until now, these two issues have not received sufficient research attention, and thus they need to be addressed urgently for PS applications to gain widespread acceptance. In this article we summarize some recently proposed schemes that address various aspects of these issues. We also identify new challenges and suggest several future research directions.

Daojing He (corresponding author) is with East China Normal University.

Sammy Chan is with City University of Hong Kong.

Mohsen Guizani is with the University of Idaho, ID, USA.

0163-6804/15/$25.00 © 2015 IEEE

People-Centric Sensing Services and Their Privacy and Incentive Requirements

Architecture of PS systems

As shown in Fig. 2, the basic infrastructure of a typical PS system involves the following entities.

- Mobile nodes are the smartphones carried by the participants. They are used to sense the parameters specified by the service provider, compile data reports, and submit them to the server operated by the service provider through WiFi or mobile telecommunication networks.
- Queriers subscribe to specific information collected by PS applications and receive reports from the service provider. Being customers of PS services, they could indicate which types of data they are interested to obtain.
- Service providers manage the PS service to facilitate effective sharing of data between mobile nodes and queriers. Usually, the service provider has abundant storage capacity and computation power, thus it can process the acquired data before distributing them to the queriers.
- Trusted third party (TTP) is an optional entity. It is needed in some privacy-preserving schemes.

The flow of a PS service is described as follows. A service provider learns from queriers about the types of data that they are interested in. It then announces the tasks of collecting these data. Those mobile nodes who are willing to participate will carry out the task. Having completed a task, a mobile node submits a report to the server of the service provider. The service provider aggregates all the returned reports, extracts the appropriate information, and forwards to queriers.

Privacy and Incentive Requirements

In general, service providers should not be considered as trustworthy for protecting participants' privacy. Information collected by a service provider might possibly be exploited by malicious administration personnel belonging to the service provider. Related to this, privacy requirements often include the following.

Node Privacy: Data reports should not reveal the private information of mobile nodes, such as identities and locations, to the service provider, other mobile nodes, or queriers. Moreover, only authorized queriers can access reports to obtain information about parameters being measured and the measured values.

Querier Privacy: Personal information may be inferred from query interests. No entity in the PS system, including the service provider, mobile nodes, or other querier, can learn any information about the interests of a querier.

Anonymity, Report Unlinkability, Location Privacy: A PS system with privacy preservation should provide anonymity, report unlinkability, and location privacy. Here report unlinkability means no entity can link two or more reports associated with the same mobile node.

Figure 1. Sensors commonly available on smartphones.

No Trusted Third Party: Trust between different entities is limited. Since there is no trusted third party, a single point of failure is eliminated.

The following are basic requirements for the design of incentive mechanisms. First, a mechanism should motivate users to participate. That is, each participant should achieve a non-negative utility. At the same time, total payments to participants should be minimized such that the service provider does not incur any loss. Second, a mechanism should encourage participants to reveal their true costs. For example, a mechanism can be designed in such a way that no participant can increase its utility by reporting a cost that deviates from the true value, irrespective of the costs reported by other participants. This prevents participants from over charging the service provider, as it generally does not know the actual participation cost. Third, a mechanism should be computationally efficient, that is, the outcome can be computed in polynomial time.

Privacy Mechanisms

Privacy-Aware Data Collection

Privacy in PS systems is first addressed by AnonySense [4], which is a general framework to provide anonymous tasking and reporting based on mix network techniques. Specifically, it makes use of multiple relays and onion routing to anonymize the connections between participants and the service provider. As a result, the IP addresses and information about current locations of the participants are hidden. AnonySense also prevents leakage of participants' private information through submitted reports by providing k-anonymity. Later, this property is strengthened in [5] by achieving l-diversity. However, AnonySense suffers from several drawbacks. First, mix networks do not provide

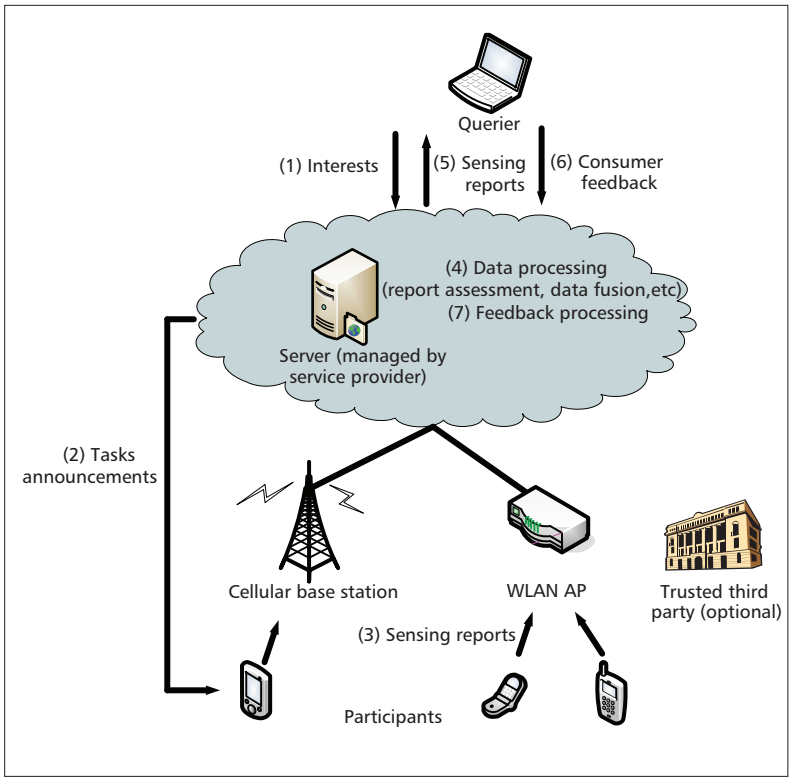


Figure 2. Architecture of a people-centric sensing system.

provable privacy guarantees. For example, anonymity in a mix network providing *k*-anonymity depends on the number of received reports. Second, AnonySense assumes WiFi as the network infrastructure of PS systems so that user unlinkability between reports with respect to access points can be guaranteed by MAC-IP address recycling techniques. Unfortunately, a ubiquitous presence of non-restricted WiFi networks is not yet a practical assumption. Third, reports are sent to the so called report server (RS), which then distributes them to queriers. The degree of report confidentiality depends on whether the RS is trustworthy.

Caceres *et al.* [6] suggest that privacy can be protected if each participant has access to a private server (e.g. a virtual machine hosted by a cloud service) and uses it as a proxy between its sensors and the service providers. However, the cost and availability of per-user proxies would severely limit the feasibility of this approach in large scale PS systems.

In [7] De Cristofaro and Soriente propose the Privacy-Enhanced Participatory Sensing Infrastructure (PEPSI) to provide node privacy and query privacy. When a mobile node submits a report to the service provider, the report is encrypted such that only queriers subscribing to the parameters contained in the report have the key to decrypt it. Since the service provider cannot access the contents of reports and queries, a tag is associated with an individual report or query. Upon receiving a report, the service provider matches its tag with queries carrying the same tag and forwards the report to the corresponding queriers.

This approach of comparing tags is a very efficient way to forward reports to the corresponding queriers. Otherwise, if a report does not carry such a tag, the service provider simply has to forward all encrypted reports to all queriers. Each querier has no clue about the type of each received report, but can only try to decrypt each one using each of its decryption keys. Only those reports that can be decrypted provide the measurements of its interest. Unfortunately, given the large number of reports generated by mobile nodes, this means that a lot of processing effort is wasted in order to obtain the desired measurements.

The cryptographic building block of PEPSI is the identity-based encryption (IBE) approach, which is an asymmetric cryptographic primitive based on bilinear map pairings. The distinguishing feature of IBE is that some unique information of the recipient's identity can be used to derive its public key, and the corresponding private key is issued by a trusted authority. In PEPSI, each report or subscription is identified by a set of labels, which naturally are parameters contained in the report. Mobile nodes encrypt their reports using the reports' labels as the public encryption keys. The privacy functionalities of the above schemes are compared in Table 1.

Privacy-Aware Data Collection and Reward Claiming

The schemes in [4–7] only focus on how participants upload the collected data to the server without revealing their identities, but they do not consider the privacy aspect in the reward claiming process. To address this issue, some approaches have been proposed. In [8] Zhang *et al.* propose a method to protect the privacy of a mobile node when submitting a report and claiming the corresponding reward. Their method requires the existence of a TTP, which is called a certification center. In the system initialization phase, the service provider first chooses a suitable one-way hash function, a master key, and an asymmetric encryption function. The one-way hash function is distributed to all mobile nodes and the certification center, while the master key and the encryption function are sent to the certification center via a secure channel.

In the data collection and upload phase, data is collected by mobile nodes according to the specifications sent from the service provider. Then a report containing the data, hash value of a randomly generated pseudonym, and a timestamp is sent to the service provider via an ordinary channel. Upon receiving the report, the service provider generates a certificate using the encryption function based on the master key, received hash value, and timestamp. This certificate is contained in a confirmation message that is returned to the mobile node via an ordinary channel. Later, when the mobile node wants to claim its reward, it prepares a message containing the pseudonym, certificate, and the timestamp, and sends the message to the certification center via a secure channel. Upon receiving the claim message, the certification center only needs to generate a certificate from the received parameters and then compares it with the received certificate. If they match, the mobile node can obtain a reward for its uploaded report.

In this scheme, since only pseudonyms are used in reports and claims, the real identities of mobile nodes are not revealed and thus their privacy is protected. In addition, an adversary cannot steal any reward by eavesdropping the confirmation message, because when it claims the reward, it needs to provide the pseudonym. However, the confirmation message only contains the hash value of the pseudonym. Due to the irreversible property of hash functions, it is infeasible for the adversary to recover the pseudonym from its hash value. Thus, it cannot generate a legitimate claim. Moreover, since the master key is only kept by the service provider and the certification center, no adversary can generate any fake certificate to claim a reward.

In [9] Li and Gao propose another scheme to protect the privacy of a mobile node when submitting a report and claiming the corresponding reward. The scheme assumes that there exists a certificate authority, which is responsible for issuing different sets of public and private keys to the service provider and mobile nodes, and that anonymous communication between mobile nodes and the service provider are maintained. Similar to [8], pseudonyms are used to protect the privacy of mobile nodes. For a given report, the service provider cannot tell which mobile node has submitted the report. In addition, reports submitted by the same mobile node are unlinkable, which means that the service provider cannot tell if those reports are submitted by the same mobile node. When a mobile node claims the reward, its claim is also identified by a pseudonym rather than the real identity. Most importantly, a comprehensive mechanism is included in the scheme to ensure that the reward claimed by each mobile node is no more than that allowed by the service provider. In other words, if the service provider is only willing to pay c credits for one report of a task, each mobile node can claim at most c credits even though it submits more than one report.

To achieve the above goal, the mechanism enforces the following three conditions:

- Each mobile node can only accept a task at most once.
- The mobile node can only submit at most one report for each accepted task.
- The mobile node can earn c credits from a report.

For the first condition, the basic idea is to associate each task with a request token. When a mobile node is interested to accept a task, it needs to send the *request token* to the service provider to obtain approval. If only one request token for each task is issued to a mobile node and the token is consumed when the mobile node accepts the task, then a mobile node cannot accept a task more than once. Similarly, for the second condition, each report is associated with a *report token* and each mobile node is only given one report token for each task. When a mobile node submits a report, it consumes the report token and hence cannot submit more reports. To ensure that the appropriate tokens are consumed by mobile nodes, each token is also associated with a *commitment*, which is used by the service provider to verify the validity of a token.

In its simplest version, the scheme in [9] assumes the existence of a TTP, which is responsible for generating tokens for each mobile node and their commitments. Since most computations are carried out by the TTP, the computation and storage cost incurred at each mobile node is low.

As shown in Fig 3, there are three major phases: setup, task assignment, and report submission. In the setup phase, the TTP pre-computes a set of request tokens and their commitments that each mobile node and the service provider will use for the next M tasks. Then it distributes the request tokens and commitments to mobile nodes and the service provider, respectively. Moreover, the TTP also sends a secret key to the service provider to generate report tokens and commitments. In the task assignment phase, the service provider broadcasts a list of tasks to the mobile nodes. If a mobile node decides to accept task i, it sends a request message containing the request token to the service provider. The service provider verifies the validity of the request token by comparing it with the commitment obtained from the TTP. If the result is positive, it sends the mobile node an approval message, which contains the necessary parameters for the mobile node to generate the report token. Clearly, this steps ensures that a mobile node cannot generate the report token for a task without the approval from the service provider. Subsequently, the service provider generates the corresponding commitment for the report token. In the report submission phase, after the mobile node has generated a report for task i, it submits both the report and its corresponding token. If the service provider verifies that the report token is indeed committed to task i, it sends pseudo-credits to the mobile node. From the pseudo-credits, the mobile node computes c credit tokens. Later, a mobile node can deposit its earned credits, but using its real identity.

	AnonySense [4] and its extended version [5]	Caceres's method [6]	PEPSI [7]
Node privacy	Not guaranteed	Yes	Yes
Querier privacy	Not guaranteed	Yes	Yes
Report unlinkability and location privacy	Not guaranteed	No	No
Trusted third party	Needed	Needed	Needed

Table 1. Privacy functionality comparison among the existing methods.

Incentive Mechanisms

Another important issue in the design of a PS system is how to design incentive mechanisms to motivate mobile nodes to participate in a PS system. For example, it is assumed in [8] and [9] that c credits are awarded to a mobile node for a submitted report. How should c be determined such that the requirements in the previous section are satisfied.

In [10] the problem is modeled as a reverse auction in which mobile nodes bid for participa-

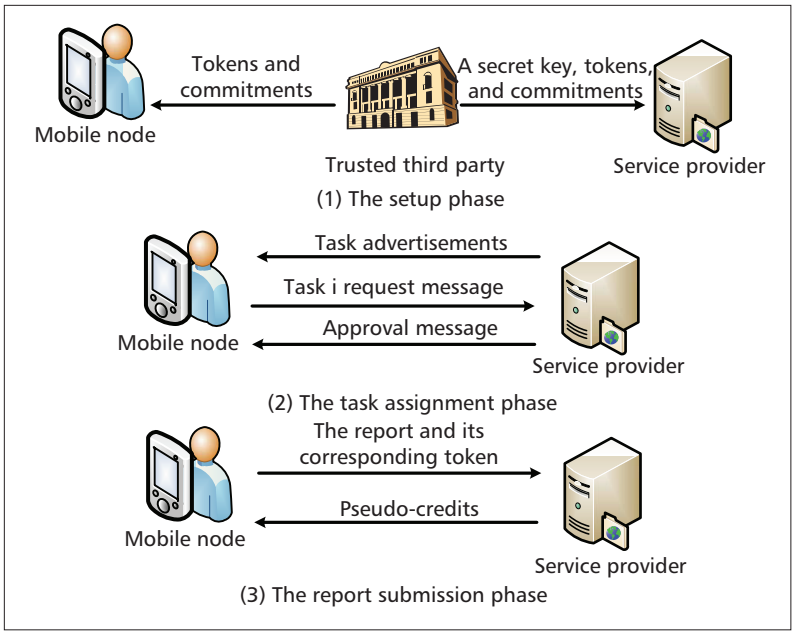


Figure 3. Information processing flow in the TTP-based scheme of [9].

tion, and the service provider then selects a predefined number of participants with the lowest bid prices. Those selected participants receive their bid prices as rewards for their participation. Each mobile node has its true cost of participation, which is private information. In order to gain positive utility from participation, each mobile node is expected to submit a bid price which is higher than its cost. This simple reverse auction suffers from the problem of *incentive cost explosion*, in which mobile nodes with high costs eventually drop out of the auction. Such dropping out of mobile nodes reduce price competition among the remaining nodes, who can afford to increase their bids in future auction rounds and yet are still selected by the service provider. As a result, the total reward paid by the service provider explodes.

To address this problem, the authors of [10] design a mechanism which, in each auction round, awards some virtual credits to those mobile nodes who are not selected. Then, when selecting participants in each auction round, the selection criterion used by the service provider is the competition bid prices, which is the actual bid price minus the accumulated virtual credits. The selected participants are awarded according to their bid prices. This mechanism increases the probability of previously rejected nodes being chosen and hence encourages their continuous participation. Simulation results have shown that the total reward made by the service provider is reduced. Note that enforcing truth-telling is not an objective of this work.

In [11] Yang *et al.* consider two types of incentive mechanisms: *platform-centric* and *user-centric*. In the platform-centric incentive mechanism, the service provider offers a total reward R for a single task to a set of participants. The level of participation of mobile node i is represented by the number of time units t_i that it is willing to provide the sensing service. It is assumed that the sensing cost per unit time of each participant is known by all other participants. For a given R, the participants are playing a non-cooperative game in which the strategy of participant i is t_i. The problem of assigning tasks is modeled as a Stackelberg game, with the solution specifying the optimal value of R which maximizes the utility of the service provider, and the corresponding unique Nash equilibrium strategy profile of mobile nodes. The incentive to each participant is that it will earn a payment that is not less than its cost of sensing.

Alternatively, in the user-centric incentive mechanism, the sensing cost of each mobile node is kept confidential and only known to itself. When a mobile node is interested in participating in a task, it announces a reserve price, the lowest price at which it is willing to participate. Then, based on the bids submitted by mobile nodes, the service provider needs to select a subset of mobile nodes and determine the payment to each participant. This problem is modeled as a reverse auction with two objectives. First, in order to provide incentive to participants, the payment to an individual participant cannot be less than its reserve price. Second, the utility of the service provider (due to task completion) minus the total payment is maximized. This objective is to ensure that the service provider has incentive to operate the service. The authors of [11] solve this problem by a heuristic algorithm. Very importantly, the resultant payment mechanism has the desirable property that each mobile node truthfully sets its bid equal to its sensing cost because it cannot obtain a higher utility by doing otherwise. As a result, over-payment is implicitly controlled.

In [12] the above problem is significantly generalized. First, rather than selecting a subset of mobile nodes as participants, the service provider allocates tasks in a finer level. Mobile nodes can be allocated with different participation levels in terms of, for example, the number of sensing samples per unit time. Second, the service provider needs to fulfill a constraint of the quality of service (QoS) delivered to users. Since this QoS depends on the quality of data contributed by participants, when the service provider allocates tasks to participants, it takes into account the quality of their past contributed data. The author of [12] derives a mechanism that explicitly computes the payment and participation level for each mobile node, while minimizing the total payment. Moreover, the mechanism also ensures that each mobile node truthfully sets its bid equal to its sensing cost, and the payment to an individual participant is not less than its cost.

The works in [11, 12] require the concurrent presence of bidding mobile nodes for the service provider to execute the mechanism in each round of task assignment. They are offline approaches that deal with a static population of mobile nodes. In a practical environment, mobile nodes may join or leave the auction dynamically. In [13] Zhang *et al.* propose an online mechanism for the problem in [11] to distribute tasks to participants on their arrivals such that immediate decisions are made according to bids at those instants. The characteristics of these incentive schemes are summarized in Table 2.

	[10]	[11]-Platform-centric	[11]-User centric	[12]	[13]
Enforce truth-telling	No	Not relevant	Yes	Yes	Yes
Minimize over payment	Implicitly	Implicitly	Implicitly	Explicitly	Implicitly
Online/offline	Offline	Offline	Offline	Offline	Online
Task assignment	Coarse	Coarse	Coarse	Fine	Coarse

Table 2. Characteristics comparison among the reviewed incentive schemes.

Research effort can be channeled to define some general privacy metrics which can capture the level of privacy protection regardless of the application scenarios. Such metrics are useful to compare the performance of different privacy-preserving schemes.

Prospects

Despite the proposed solutions to privacy-preservation and incentive mechanisms summarized in the previous sections, there are still challenging issues that require more research. Here we discuss some of the important ones.

Developing General Privacy Metrics: Currently, different metrics, criteria, or methods are used to evaluate the performance of different proposed privacy-preservation schemes. Ideally, a set of universal metrics should be available to quantify privacy. However, it would be very difficult, if not impossible, to define such metrics. Alternatively, research efforts can be channeled to define some general privacy metrics that can capture the level of privacy protection regardless of the application scenarios. Such metrics are useful to compare the performance of different privacy-preserving schemes.

Resisting Corrupted Data While Preserving Privacy: While it is important for PS systems to preserve participants' privacy, such systems are vulnerable to the corrupted data purposely submitted by malicious participants. As a possible way to address this problem, the service provider should have methods to evaluate the reliability of the data, e.g. using the reputation of individual participants. Moreover, for incentive mechanisms it is common to select participants based on their bids. It would be interesting to take into account the reputation of individual mobile nodes in the selection process.

Maintaining a Balance Between Privacy Preservation and Accountability: Alternatively, in order to ensure data trustworthiness, participants should be made accountable for their submitted data. Clearly, this requires the identities of participants to be disclosed and counteracts privacy-preserving schemes. Therefore, how to maintain an acceptable balance between privacy preservation and accountability is an important issue that needs to be addressed.

Designing Efficient Incentive Mechanisms: For incentive mechanisms, further research can be directed to improve existing mechanisms to be more efficient. For example, similarity of measurements from devices near each other can be considered when allocating tasks and deciding payments. This could reduce the payment made by the service provider and eliminate possible redundancy in measured data [12].

Developing Efficient Privacy-Preservation Methods: More work should be done to improve the efficiency of privacy-preservation methods. For example, although PEPSI [7] can offer node privacy and query privacy, it suffers from heavy communication and computation costs because its underlying pairing-based IBE method involves bilinear pairing, which is known as a costly operation against the operations executed in elliptic curve groups.

Privacy and Reward Claiming: From the above reviewed works, we learn that privacy and payment determination are two orthogonal problems and their solutions can be designed independently. However, we do need to consider how privacy is preserved when participants claim their entitled rewards, and how rewards can be correctly distributed when privacy is preserved. For example, with the protection of anonymity, credits may be claimed by a greedy user who uses different anonymous identifiers to submit many duplicated reports for the same sensing task, or a malicious user who steals and uses other users' credentials. Although the methods described above address these two interplaying issues, they suffer from some security weaknesses and efficiency problems. For example, the method in [8] relies on the existence of a TTP. Also, the method in [9] is designed for a special scenario in which each sensing task only requires one report from each user. Thus, more novel solutions are required.

Privacy in Incentive Mechanisms: Privacy is not only important in report submission and reward claiming, but also in incentive mechanisms. For example, users may not want their bid information to be known by others as this reflects their true valuations on the sensing tasks. Another example is that the service provider wants to protect the threshold payment. Thus, privacy preservation for both users and the service provider should be taken into account in the design of incentive mechanisms. Recently, researchers started addressing this issue [14].

Conclusion

Privacy preservation and incentives are two factors that can hinder the large-scale deployment of people-centric sensing. In this article we have reviewed a number of proposed solutions to these two issues, and suggested some future research directions. We hope this article will stimulate more research from the community.

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

Privacy is not only important in report submission and reward claiming, but also in incentive mechanisms. For example, users may not want their bid information to be known by others as this reflects their true valuations on the sensing tasks.

ACKNOWLEDGMENT

This research is supported by a strategic research grant from City University of Hong Kong (project no. 7004429), the Pearl River Nova Program of Guangzhou (no. 2014J2200051), the National Science Foundation of China (grant no. 51477056), the Shanghai Knowledge Service Platform for Trustworthy Internet of Things (no. ZF1213), the Shanghai Rising-Star Program (no. 15QA1401700), and a visiting scholar project of the State Key Laboratory of Power Transmission Equipment & System Security and New Technology (no. 2007DA10512713406).

REFERENCES

[1] IDC, "Worldwide Smartphone Shipments Top One Billion Units for the First Time, According to IDC," press release, 27 Jan 2014.

[2] N. Maisonneuve, M. Stevens, and B. Ochab, "Participatory Noise Pollution Monitoring using Mobile Phones," *Information Policy*, vol. 15, no. 1–2, 2010, pp. 51–71.

[3] S. Mathur *et al*., "Parknet: Drive-by Sensing of Road-Side Parking Statistics," *Proc. ACM MobiSys*, 2010, pp. 123–36.

[4] C. Cornelius *et al.*, "AnonySense: Privacy-Aware People-Centric Sensing," *Proc. ACM MobiSys.*, 2008, pp. 211–24.

[5] K. L. Huang, S. S. Kanhere, and W. Hu, "Preserving Privacy in Participatory Sensing Systems," *Comput. Commun.*, vol. 33, no. 11, Jul. 2010, pp. 1266–80.

[6] R. Caceres, L. P. Cox, H. Lim, A. Shakimov, and A. Varshavsky, "Virtual Individual Servers as Privacy-Preserving Proxies for Mobile Devices," *Proc. MobiHeld Workshop*, 2009., pp. 37–42

[7] E. De Cristofaro and C. Soriente, "Extended Capabilities for a Privacy-Enhanced Participatory Sensing Infrastructure (PEPSI)," *IEEE Trans. Inf. Forens. Security*, vol. 8, no. 12, Dec. 2013, pp. 2021–33.

[8] J. Zhang *et al.*, "A Novel Privacy Protection Scheme for Participatory Sensing with Incentives," *Proc. IEEE CCIS*, 2012, pp. 1017–21.

[9] Q. Li and G. Cao, "Providing Privacy-Aware Incentives for Mobile Sensing," *Proc. IEEE PerCom*, 2013, pp. 76–84.

[10] J.-S. Lee and B. Hoh, "Dynamic Pricing Incentive for Participatory Sensing," *Pervasive and Mobile Computing*, vol. 6, no. 6, Dec. 2010, pp. 693–708.

[11] D. Yang *et al.*, "Crowdsourcing to Smartphones: Incentive Mechanism Design for Mobile Phone Sensing," *Proc. ACM MobiCom*, 2012, pp. 173–84.

[12] I. Koutsopoulos, "Optimal Incentive-driven Design of Participatory Sensing Systems," *Proc. IEEE INFOCOM*, 2013, pp. 1402–10.

[13]X. Zhang *et al*., "Free Market of Crowdsourcing: Incentive Mechanism Design for Mobile Sensing," *IEEE Trans. Parallel Distrib. Syst*., vol. 25, no. 12, Dec. 2014, pp. 3190–3200.

[14] J. Sun and H. Ma, "Privacy-Preserving Verifiable Incentive Mechanism for Online Crowdsourcing Markets," *Proc. IEEE ICCCN*, 2014, pp. 1–8.

BIOGRAPHIES

DAOJING HE (S'07, M'13) received his B.Eng. (2007) and M. Eng. (2009) degrees from Harbin Institute of Technology, P.R. China, and his Ph.D. degree (2012) from Zhejiang University, P.R. China, all in computer science. He is currently a professor at the Software Engineering Institute, East China Normal University, P.R. China. His research interests include network and systems security. He is an associate editor or on the editorial boards of several international journals such as *IEEE Communications Magazine* and *IEEE/KICS Journal of Communications and Networks*.

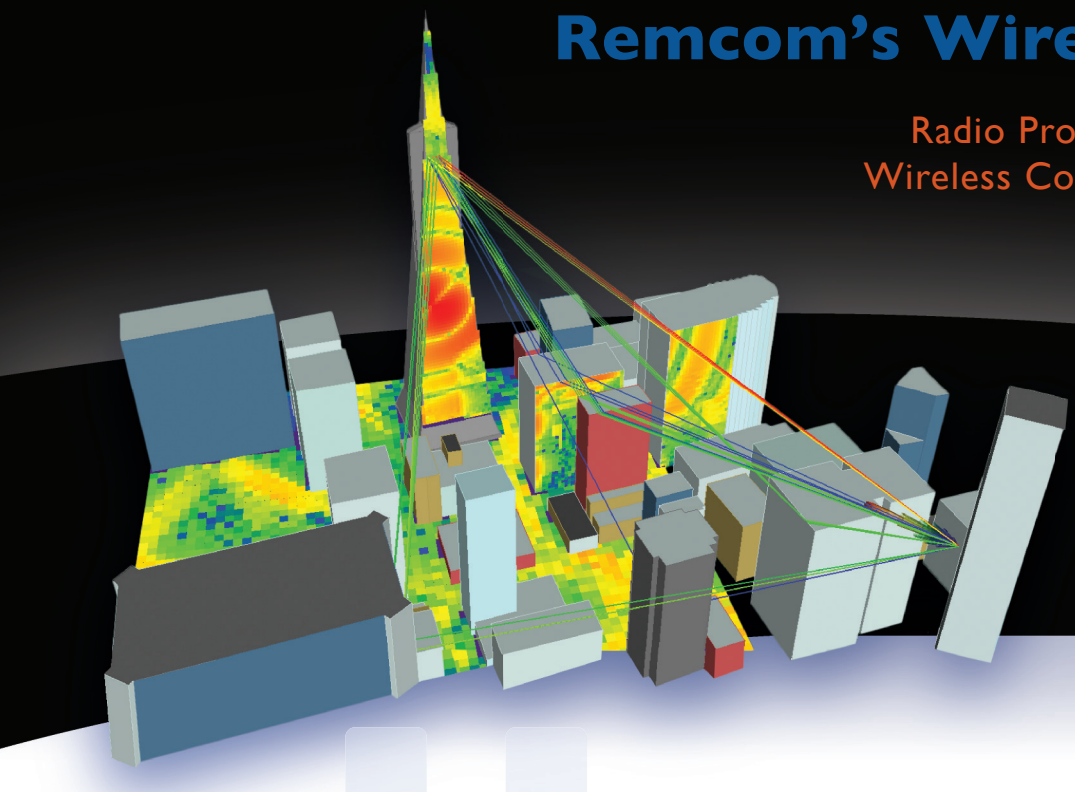
SAMMY CHAN (S'87-M'89) received his B.E. and M.Eng.Sc. degrees in electrical engineering from the University of Melbourne, Australia, in 1988 and 1990, respectively, and a Ph.D. degree in communication engineering from the Royal Melbourne Institute of Technology, Australia, in 1995. From 1989 to 1994 he was with Telecom Australia Research Laboratories, first as a research engineer, and between 1992 and 1994 as a senior research engineer and project leader. Since December 1994 he has been with the Department of Electronic Engineering, City University of Hong Kong, where he is currently an associate professor.

MOHSEN GUIZANI (S'85, M'89, SM'99, F'09) received the B.S. (with distinction) and M.S. degrees in electrical engineering, and the M.S. and Ph.D. degrees in computer engineering from Syracuse University, Syracuse, NY, USA, in 1984, 1986, 1987, and 1990, respectively. He is currently a professor and Chair of the Electrical and Computer Engineering Department at the University of Idaho, USA. He was a professor and the Associate Vice President for Graduate Studies at Qatar University, Doha, Qatar. He was the Chair of the Computer Science Department at Western Michigan University, Kalamazoo, MI, USA, from 2002 to 2006, and the Chair of the Computer Science Department at the University of West Florida, Pensacola, FL, USA, from 1999 to 2002. He has held academic positions at the University of Missouri-Kansas City, MO, USA; the University of Colorado-Boulder, CO, USA; Syracuse University, Syracuse, NY, USA; and Kuwait University, Kuwait City, Kuwait. His research interests include computer networks, wireless communications and mobile computing, and optical networking. He currently serves on the editorial board of six technical journals and is the founder and the editor-in-chief of the *Wireless Communications and Mobile Computing Journal* published by John Wiley. He is the author of eight books and more than 300 publications in refereed journals and conferences. He has guest edited a number of special issues in IEEE journals and magazines. He has also served as a member, Chair, and General Chair of a number of conferences. He has served as the Chair of the IEEE Communications Society Wireless Technical Committee and the Chair of TAOS Technical Committee. He was an IEEE Computer Society Distinguished Lecturer from 2003 to 2005. He is a Senior Member of the ACM.

Advertisers' Index

Remcom's Wireless InSite®

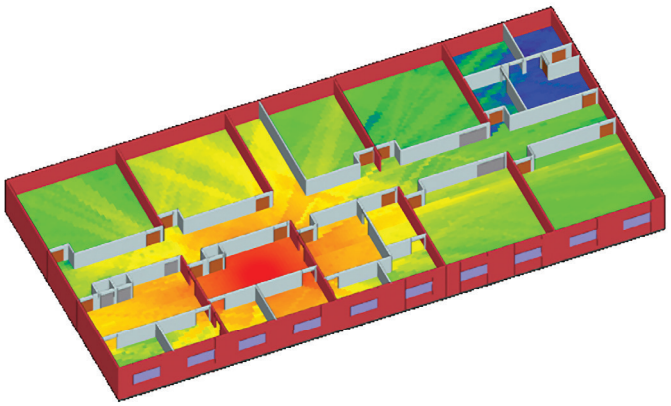
Radio Propagation Software for Wireless Communication Planning



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

Wireless EM Propagation Capabilities for a Variety of Applications

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



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

REMCOM®

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

Visit Us at MILCOM 2015
Booth #426