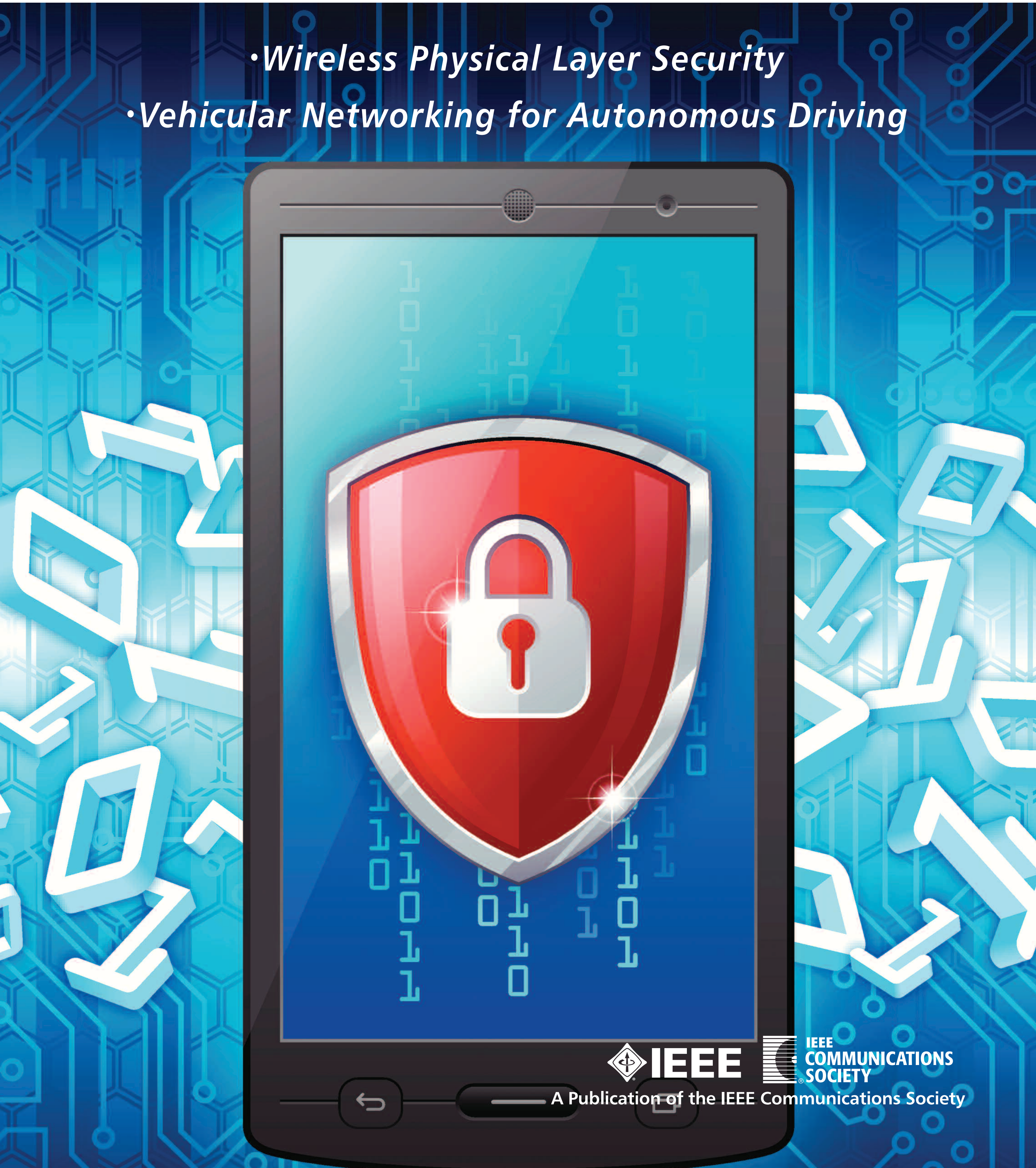
December 2015, Vol. 53, No. 12

IEEE Communications MAGAZINE

www.comsoc.org



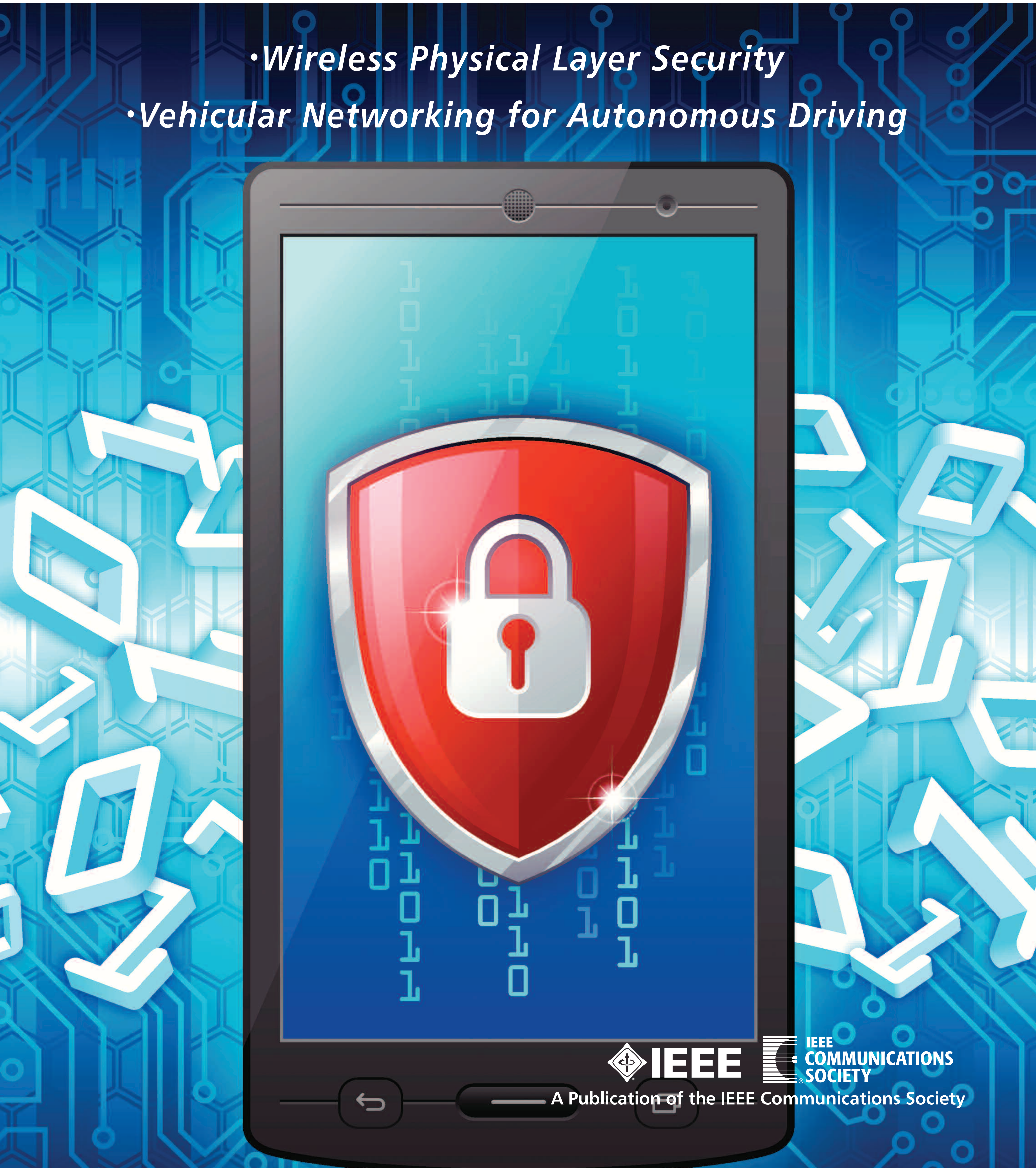
- Wireless Physical Layer Security
- Vehicular Networking for Autonomous Driving

IEEE
IEEE COMMUNICATIONS SOCIETY
A Publication of the IEEE Communications Society

December 2015, Vol. 53, No. 12

IEEE Communications MAGAZINE

www.comsoc.org

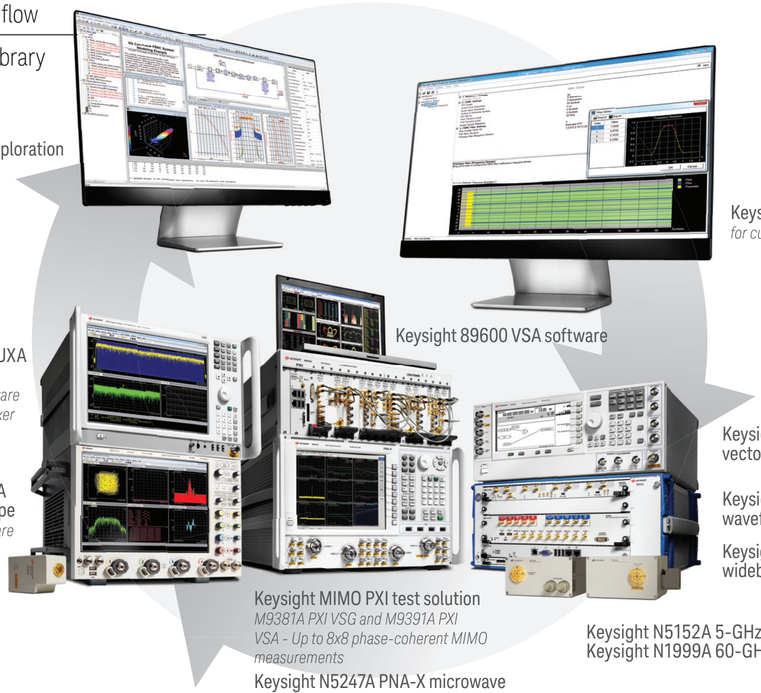


·Wireless Physical Layer Security

·Vehicular Networking for Autonomous Driving

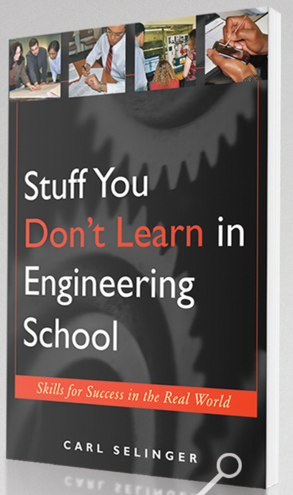
IEEE
IEEE COMMUNICATIONS SOCIETY
A Publication of the IEEE Communications Society

C = BW* log2 {1 + S/N}
mmWave
Massive MIMO
PRX = PTX GTX GRX (λ/4πR)2
Need more antenna gain!
a
3dB BW

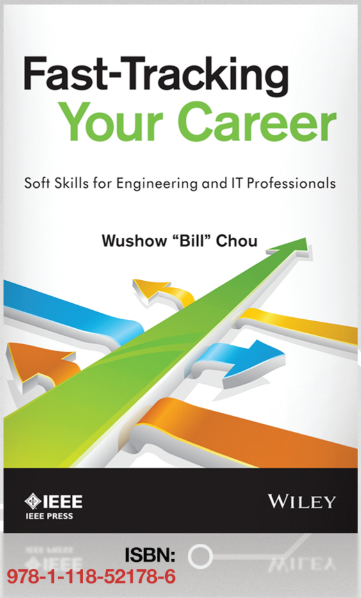


Advance Your Career

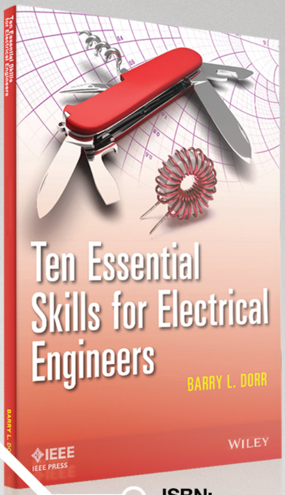
With the Experts at Wiley-IEEE Press


Stuff You Don't Learn in Engineering School
Skills for Success in the Real World
CARL SELINGER

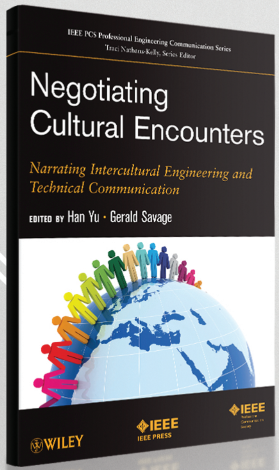
ISBN: 978-0-471-65576-3


Fast-Tracking Your Career
Soft Skills for Engineering and IT Professionals
Wushow "Bill" Chou
IEEE
IEEE PRESS
WILEY

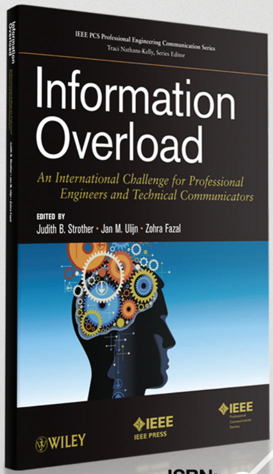
ISBN: 978-1-118-52178-6



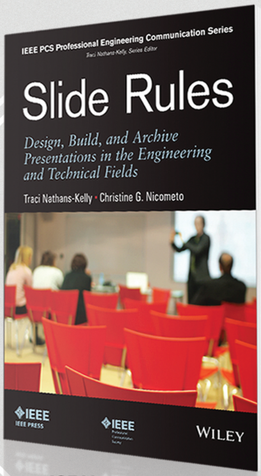
ISBN: 978-1-118-52742-9


IEEE PCS Professional Engineering Communication Series
Traci Nathans-Kelly, Series Editor
Negotiating Cultural Encounters
Narrating Intercultural Engineering and Technical Communication
EDITED BY Han Yu · Gerald Savage
WILEY
IEEE
IEEE PRESS
IEEE

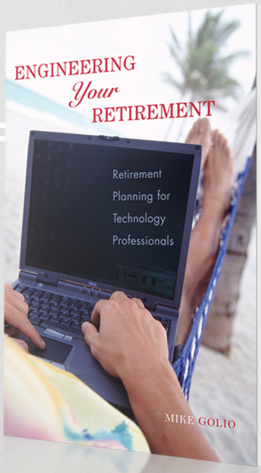
ISBN: 978-1-118-06161-9


IEEE PCS Professional Engineering Communication Series
Traci Nathans-Kelly, Series Editor
Information Overload
An International Challenge for Professional Engineers and Technical Communicators
EDITED BY
Judith B. Strother · Jan M. Ulijn · Zohra Fazal
WILEY
IEEE
IEEE PRESS
IEEE

ISBN: 978-1-118-23013-8


IEEE PCS Professional Engineering Communication Series
Traci Nathans-Kelly, Series Editor
Slide Rules
Design, Build, and Archive Presentations in the Engineering and Technical Fields
Traci Nathans-Kelly · Christine G. Nicometo
IEEE
IEEE PRESS
IEEE
WILEY

ISBN: 978-1-118-00296-4



ISBN: 978-0-471-77616-1

ARE YOU AN EXPERT IN YOUR FIELD?
TO LEARN MORE ABOUT PUBLISHING WITH WILEY-IEEE VISIT:

HTTP://WWW.IEEE.ORG/PUBLICATIONS_STANDARDS/PUBLICATIONS/PRESS/AUTHORS/BOOKS.HTML

IEEE

VISIT: WWW.WILEY/IEEE.COM

WILEY

IEEE Communications MAGAZINE

DECEMBER 2015, Vol. 53, No. 12

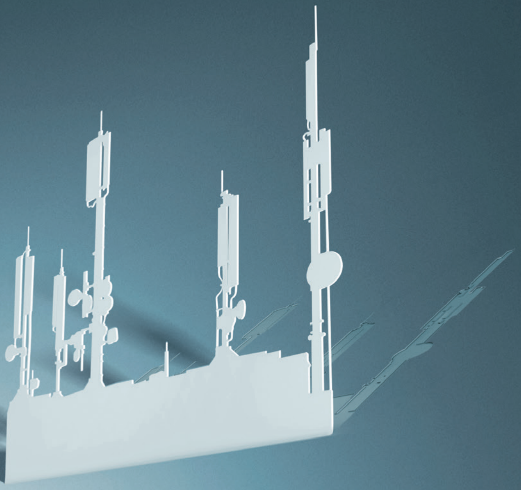
www.comsoc.org/commag

Wireless Physical Layer Security: Part II

Guest Editors: Walid Saad, Xiangyun Zhou, Mérouane Debbah, and H. Vincent Poor

Vehicular Networking for Autonomous Driving

Guest Editors: Alexy Vinel, Henrik Pettersson, Lan Lin, Onur Altintas, and Oleg Gusikhin



ROHDE&SCHWARZ
SMW200A · VECTOR SIGNAL GENERATOR
ROHDE&SCHWARZ
FSW · SIGNAL & SPECTRUM ANALYZER · 2 Hz ... 26.5 GHz

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.

In the Current Issue of the Communications Standards Supplement

IoT/M2M from Research to Standards: The Next Steps (Part II)

Guest Editors: Omar Elloumi, JaeSeung Song, Yacine Ghamri-Doudane, and Victor C.M. Leung

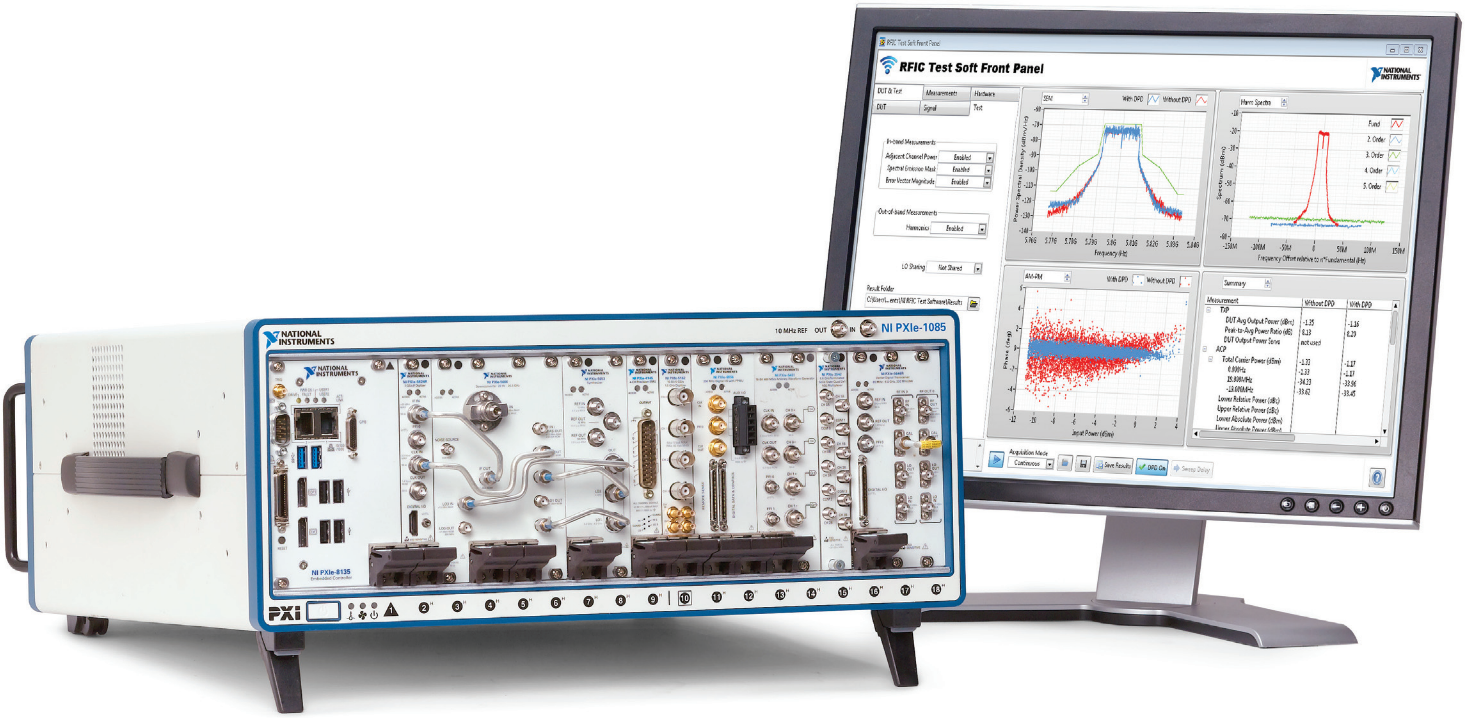
Research & Standards: Advanced Cloud & Virtualization Techniques for 5G Networks (Part II)

Guest Editors: Kan Zheng, Tarik Taleb, Adlen Ksentini, Chih-Lin I, Thomas Magedanz, and Mehmet Ulema

Currently Scheduled Topics

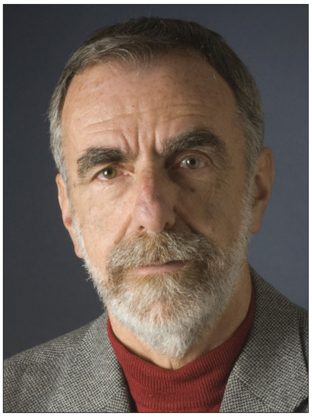
	Publication Date	Manuscript Due Date
Recent Advances in Green Industrial Networking	October 2016	December 15, 2015
Communications, Caching, and Computing for Content-Centric Mobile Networks	August 2016	January 1, 2016
Social and Mobile Solutions in Ad Hoc and Sensor Networking	July 2016	January 11, 2016
SDN Use Cases for Service Provider Networks	October 2016	January 31, 2016

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



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

LOOKING BACK ON TWO BUSY YEARS



SERGIO BENEDETTO

My term as President of the IEEE Communications Society will end on December 31, 2015 — two very busy years, exciting and pleasant for the close, friendly collaboration with all ComSoc leaders and staff.

In the previous September, October, and November pages, the five Vice-Presidents (Technical Activities, Conferences, Publications, Member Relations, and Standards) have described the main achievements in their respective areas of competence. I owe them (Khaled Ben Letaief, Hikmet Sari, Katie Wilson, Stefano Bregni, and Rob Fish) a debt of gratitude for their dedication to serve the ComSoc community, and for their friendly and loyal collaboration in every moment of the past two years.

I will describe in this, my last President Page, what I consider the most important facts of my term and the future actions that I suggest Harvey Freeman, the next President, to consider. Before starting, though, I want to clarify that the successes are the result of an intense, synergistic teamwork, whereas I am fully responsible for those areas where we did not succeed.

COMSOC BUDGET

In Fig. 1 the last five years of ComSoc's budget are shown. The years 2012 and 2013 were deeply in the red, and in 2014 we had to present to the IEEE Finance Committee a firm commitment and plan to reduce the deficit, in order to avoid being put on the IEEE watch list. Though our action has been helped by a new algorithm to compute the overhead imposed by IEEE on societies, we have succeeded in repositioning our budget in the positive zone for two consecutive years, and the budget forecast for 2016 is in line with the previous two.

The budget recovery has been achieved by cutting expenses, without reducing the services to members, so that from now on all future incremental revenues can be allocated to improve the members' experience. We have identified, in agreement with the suggestion of the Strategic Planning Committee led by Byeong G. Lee, the area of education and training as the most promising in terms of potential revenues, and have already started to invest volunteer resources to follow that path (see the August 2015 President Page on the subject).

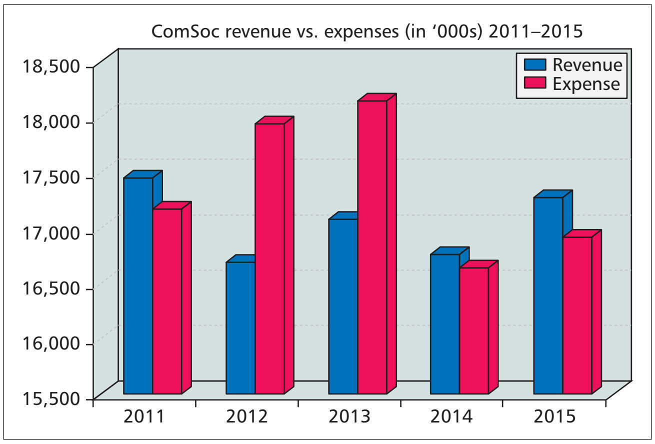


Figure 1. ComSoc budget, 2011–2015.

COMSOC GOVERNANCE

Based on the plan prepared by the Strategic Planning Committee, the Board of Governors approved in June 2015 a restructuring of ComSoc governance more in line with the strategic actions to pursue. The new structure is shown in Fig. 2. The changes are highlighted in red. They consist in renaming the Vice President of Standards Activities to Vice President of Standards & Industry Activities; the Vice President of Member Relations to Vice President of Member & Global Activities; and the Vice President of Technical Activities to Vice President of Technical & Educational Activities. Moreover, the Director of Education & Training has been renamed the Director of Educational Services; the Director of Marketing & Industry Relation has been renamed the Director of Industry Outreach and transferred from under the VP–Member Relations to under the VP–Standards & Industry Activities; the Director of Conference Publications has been eliminated and the new position Director of Technical Services was created and placed under the VP–Technical & Educational Activities; and the Director of Membership Service Development has been renamed the Director of Member Services.

The two main purposes of the restructuring are to concentrate the activities aimed at regaining industry participation in ComSoc under one VP, dealing with both Standards (of great interest to industry) and Industry, and to underline the enhanced importance of Education Activities as the third pillar of ComSoc.

The governing documents of ComSoc, i.e. Constitution, Bylaws, and Policies & Procedures, were suffering from misalignments, lack of updating, and inconsistencies between them and with IEEE prescriptions. In the last two years I asked the Governance Committee to revise them all with the interested leaders and bodies. It has been a huge effort, which has achieved a first important result with the approval at ICC 2015 in June of a first batch of revised documents. In the second meeting of the Board of Governors, at GLOBECOM 2015, I plan to approve the remaining documents, so as to pass to the next President a truly consistent set of governing documents.

STAFF REORGANIZATION

To better serve the needs of the new ComSoc Governance, ComSoc has approved, upon the proposal of the ComSoc Executive Director, a new organizational structure, which is shown in Fig. 3. The three vacant positions will be covered soon, dealing with technical activities, marketing, and meeting management.

IEEE RELATIONSHIPS

In the past two years ComSoc has been very active in coordinating actions with other society presidents in order to improve the relations of societies with IEEE and engage in interdisciplinary projects.

On the first task, some society presidents started an infor-

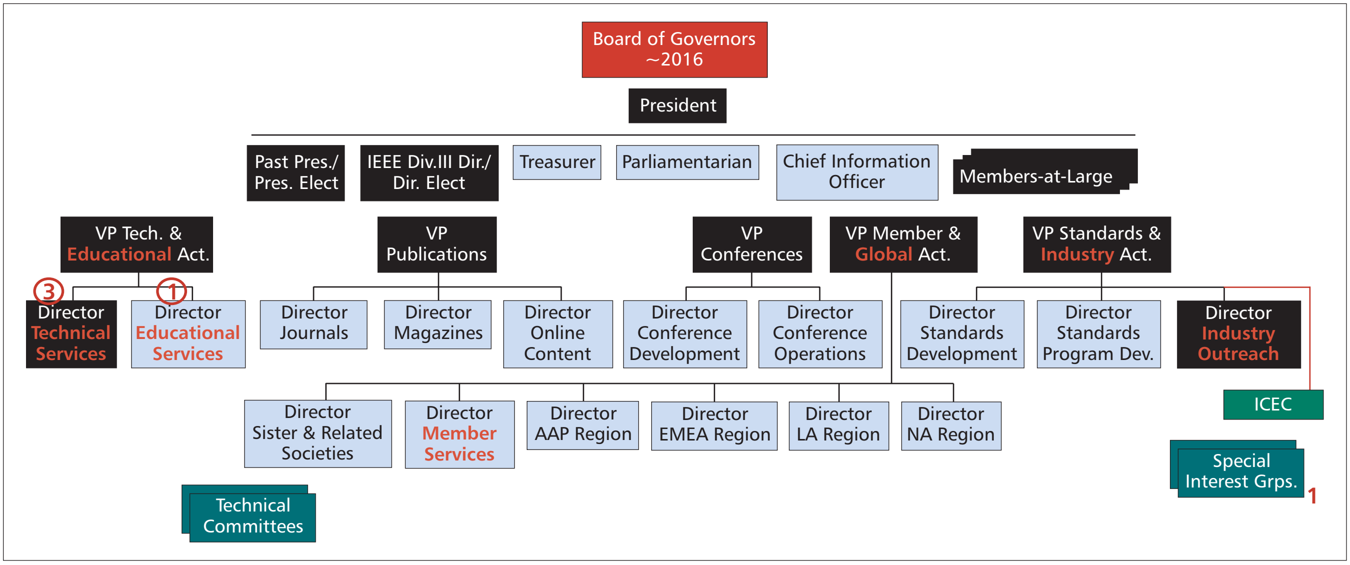


Figure 2. New structure of ComSoc governance.

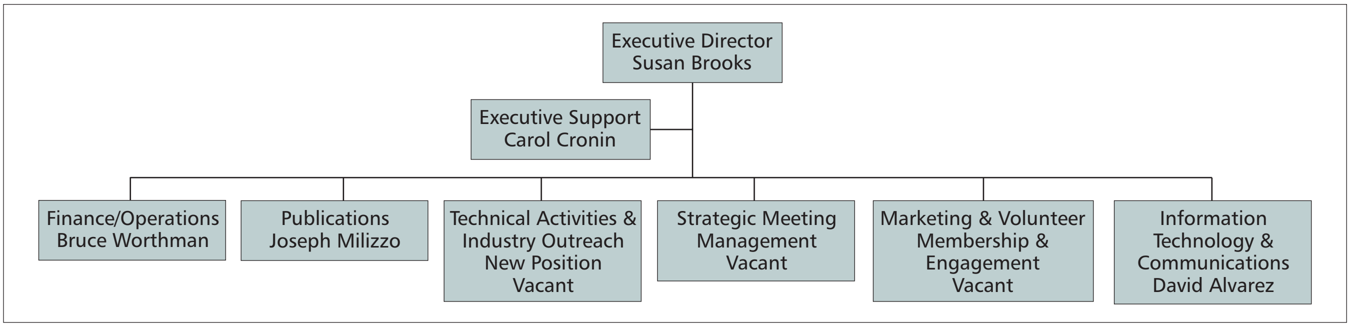

Executive Director Susan Brooks
Executive Support Carol Cronin
Finance/Operations Bruce Worthman
Publications Joseph Milizzo
Technical Activities & Industry Outreach New Position Vacant
Strategic Meeting Management Vacant
Marketing & Volunteer Membership & Engagement Vacant
Information Technology & Communications David Alvarez

Figure 3. New organizational structure of ComSoc.

mal financial transparency group that led the IEEE Technical Activity Board to vote motions aimed at obtaining a clearer and more transparent way of presenting financial data to the societies. Significant steps toward the goal have been covered, and now the financial roll-ups allow the society leaders to better understand the societies' revenue/expenditure flow and to run the societies with a higher degree of awareness. Moreover, a motion approved by TAB under presentation by several societies presidents was successful in stopping the steady increase in the general and administrative overhead applied to societies until a deeper analysis on the overhead costs will allow a more focused attribution of costs to services.

On the technical side, ComSoc has been active in the Future Directions Committee of the IEEE, and is now a member of several new initiatives, such as Internet, Cloud Computing, Smart Grid, Internet of Things, Software Defined Networks, Big Data, and Cybersecurity, and is the leader in the Green ICT initiative, which has received funding for the first two years (2015 and 2016). All the initiatives tackle crucial topics in the areas of interest of IEEE, which are by nature interdisciplinary and benefit from the wide range of competencies that IEEE societies can leverage. The initiatives aim at creating large communities around the main topics of each initiative, and to create new IPs in the form of journals, conferences, educational tools, and standards. In doing so, they interact with all ComSoc organizational areas, and there is a need for a better cooperation between the initiative leadership and societies to avoid overlaps and favor synergistic efforts. To be an active partner in this wide range of initiatives, ComSoc needs a deeper involvement of its appropriate Technical Committees, something that has been tackled by the VP–Technical Activities by appointing a Task Force to conduct a review to evaluate the technical committees' mission and structure.

MEMBERSHIP

The graph in Fig. 4 shows the evolution of ComSoc's membership in the last five years. It shows slowly varying numbers from 2011 to 2014, and then a drop taking place in 2015. The discontinuity is due to the fact that in 2015 ComSoc stopped offering free membership for the first year. In fact, sterilizing the data from this effect (represented by the blue portion of the bars in the figure), the number appears to be in continuity with the previous years. Overall, looking at the magenta bars in the figure, we notice a small decrease from 2012 to 2015.

Althoug mild, the membership decline is an issue, since it testifies that ComSoc's value proposition is not well presented and/or understood by potential members. ComSoc is under-represented in Latin America and in the BRICS countries, where the rapid development of telecommunications should bring with it a parallel growth in ComSoc membership. In November, a ComSoc delegation formed by myself, the VP–Standard Rob Fish, and the Executive Director Susan Brooks, made an outreach trip to India, meeting in Delhi and Bangalore with several leaders from industry, academia, research centers, and government, trying to understand the local needs and to design a plan to substantially enhance Com-

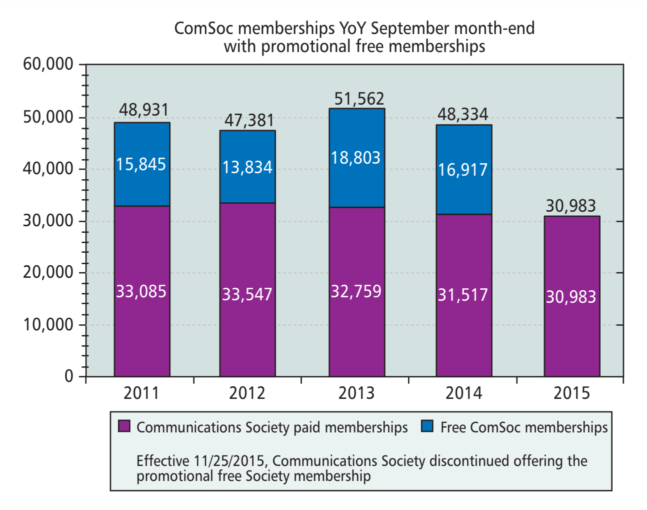


Figure 4. Membership trend in the last five years.

Soc's footprint. We are presently working on the plan.

I wrote in my first President's Page (January 2014) that "ComSoc needs to tighten the bonds between chapters and central leadership, by surveying chapter activities, offering targeted incentives to the most active chapters, and increasing the in-person (e.g. distinguished lecturers tours) and web-based (webinars, conference session broadcasts, etc.) presentation services." Active chapters well rooted in the local environment and well connected with ComSoc central offices are crucial to new member recruitement and, even more, for member retainment. We have followed the above indications, and must keep doing so, but this is a long process and its effects in terms of membership will be seen in the coming years.

INDUSTRY PARTICIPATION

One of the main goals of my term as ComSoc President has been to regain industry participation in ComSoc, after the decline that followed the storm of changes in the economy, society, and professions that took place in the early years of the new century. We put in place several initiatives toward this goal, including recruiting members of editorial boards and conference technical program committees from industry; focusing the form and content of magazine papers on topics of interest to industry and practitioners; starting a supplement on Standards in *IEEE Communications Magazine* (which has recently been approved as an independent magazine by the IEEE Periodicals Committee); strengthening ComSoc's activities in standards through concerted actions with the IEEE Standards Association; and through the "invention" of one-day meetings called Rapid Reaction Standardization events, in which researchers from academy and industry discuss "hot" technologies amenable to standardization efforts; and revamping the Committee named ICEC (Industry Content and Exhibition Committee), responsible for developing and promoting a strategic vision and a management approach for conference programs attractive to attendees from industry, administrations, or other non academic sectors.

Again, the results in terms of renewed industry interest and participation in ComSoc will be evident only in a few years. I feel, however, that the path we started moving on is the correct one, and would recommend the next leadership to insist on it.

CONFERENCES

ComSoc started in INFOCOM an optimized procedure for paper assignment and double-blind reviewing, which has been received extremely well and will be shared and tested by the GITC for our flagship conferences.

GIMS has defined and implemented a new rotation plan for our two major flagship conferences, different from the previous one prescribing GLOBECOM to take place only in the U.S. and ICC outside the U.S., ensuring that an ICC or a GLOBECOM will be located in each region (the Americas, EMEA, and Asia-Pacific) every 18 months.

The next step will be to hold for the first time a major ComSoc conference in India, and extend the model of "local" conferences to other areas beyond Latin America (with LatinCom) and BlackSeaComm.

With implementation starting January 1, 2016, IEEE changed the expense charging model for technically co-sponsored conferences, requesting a flat fee of $1,000 plus $15 per paper uploaded in IEL for each co-sponsored conference. The society co-spnsoring the conference can pass the charge completely to the conference organizers, or to cover costs in part or in total. Temporarily, ComSoc has decided to charge the $1,000 to all co-sponsored conferences, but I think this occasion could and should be used to analyze each technically co-sponsored conference in terms of quality and impact (to be measured through the number of IEL downloads), in order to rethink and possibly resize ComSoc's conference portfolio and to decide on a two-tier or three-tier charging model matching the society's interest in the conference.

PERSPECTIVES

The IEEE Communications Society's strength lies in the thousands of active volunteers spanning all topical areas and constituencies of telecommunications. In thanking again all those (volunteers and staff) who collaborated with me in running the society and those who interacted with me with suggestions and criticisms, I am happy to pass the leadership to the next President, Harvey Feeeman, entrusting him a healthy society and wishing him success in running ComSoc in the interest of all members, the telecom community, and, as the IEEE motto says, humanity.

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
17th Floor
New York, NY 10017, USA

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

STATE OF THE COMMUNICATIONS SOCIETY'S MAGAZINES

The three Communications Society (ComSoc) magazines remain very healthy, especially in terms of the number of submissions, their impact factor, and the number of downloads from IEEE Xplore. *IEEE Communications Magazine* has become the publication of record for 5G wireless topics. I want to take this opportunity to update our ComSoc members and magazine subscribers on some of the new initiatives and activities that are in progress to further strengthen our magazines and their value to our readers. As my term as Director of Magazines draws to a close, I also want to take this opportunity to thank several people involved with the magazines for their hard work. It has been a privilege and joy working with them.

PRAC Review: As mentioned in the President's Page of *Communications Magazine*'s November issue, the three ComSoc magazines went through their regular review by the IEEE Periodical Review and Advisory Committee (PRAC). This review, conducted every five years, involves an extensive written report and a face-to-face review of the report to address the comments by PRAC reviewers. I worked closely with the three Editors-in-Chief on the initial and final versions of the report submitted earlier this year, and had the privilege of representing all three magazines at last November's PRAC meeting. Not only did all three pass, there were several areas that the PRAC reviewers regarded as exemplary and worthwhile documenting as best practices. One such area is that we require one more reviewer for each paper than the general IEEE requirement. The PRAC reviewers were amazed that we were able to consistently find so many volunteer paper reviewers. So I want to extend my great 'Thank You' to all our wonderful reviewers, and the many editors who recruit them, for volunteering your precious time to make this possible!!

The PRAC preparation and feedback provided an excellent opportunity for us to recognize areas for improvement in the magazines. Even more, it helped us to share and harmonize the best practices from the magazines with each other. My thanks to Editors-in-Chief (EiCs) Sean Moore (*Communications Magazine*), H.H. Chen (*Wireless Communications Magazine*), and Sherman Shen (*Network Magazine*) for their efforts with the PRAC reports and also their willingness to both share their own best practices and adopt ones from the other EiCs.

As I had reported in my Director of Magazines editorial last year, I had worked with the EiCs to harmonize the web pages across the magazines and incorporate many best practices at that time. In addition, with the help of Sean Moore, I have compiled a series of best practice advice "from one EiC to another" and other helpful documents. These documents are now available to each new EiC on a private site that I set up within the Communications Society Community pages. I know from my own *Communications Magazine* EiC term that there are many things that an EiC learns the hard way. My goal was to pass these along to new EiCs so they can learn them the 'easy' way and consequently have more time and energy for other matters that enhance the quality of their magazines.

Potential New Magazines: In my last Director of Magazines report, I explained the new process of using *Communications Magazine* supplements as a vehicle for testing potential new magazines. As mentioned there and in last month's President's Page, we are now taking the formal steps to move the IEEE Communications Standards Supplement to become a new *IEEE Communications Standards Magazine*. The Phase 1 proposal was presented to the IEEE TAB Periodicals Committee on November 19, and approved that same day. We anticipate that the Phase 2 proposal will be approved in early 2016. The tentative plan is to continue to make the new magazine available to all ComSoc members in 2016 and begin subscriptions in 2017. The *Communications Standards Magazine* will then be the first new ComSoc magazine in many years. Many thanks to Alex Gelman and Rob Fish for championing this idea, to Glenn Parsons for his efforts as EiC, and to staff member Joe Milizzo for his great efforts with logistics and paperwork. We are currently evaluating two other potential candidate magazines.

New Editors-in-Chief and Associate Editors-in-Chief: Osman Gebizlioglu became *Communications Magazine* EiC at the beginning of 2015, and Zoran Zvonar has been appointed as Associate Editor-in-Chief (AEiC). I am confident that Osman and Zoran will make a very strong team. I want to thank outgoing EiC Sean Moore for his many efforts with the magazine.

Due to bringing in new AEiCs late last year for *IEEE Wireless Communications* and *IEEE Network*, we asked their respective EiCs, H.H. Chen and Sherman Shen, to continue in their roles through the end of June. My thanks to H.H. and Sherman, under whom both magazines thrived and grew in their impact in our communications field. Beginning in July, their AEiCs were promoted to EiCs and two new AEiCs were named. Nei Kato is now serving as EiC for *Network Magazine*, with David Soldani serving as AEiC. Hamid Gharavi is the new EiC for *Wireless Communications Magazine*, with Yi Qian appointed as AEiC. I am very excited about the quality of our new EiCs and AEiCs, and look forward to seeing how they continue to improve their respective publications.

Administrative Support: As you know, all of the magazine editors, including the EiCs, are volunteers with separate full-time jobs. As I explained in my last editorial, the primary tasks of a magazine EiC revolve around ensuring the highest quality timely content in the magazine. ComSoc has established a Managing Editor position in order to help the EiC with the many logistical tasks that are required to successfully bring this content to publication. The Managing Editor is a part-time ComSoc staff position with tasks that include sending reminder emails to editors, checking paper submissions against the list of prohibited authors, processing the open call editor applicants, and corresponding with authors. Charis Scoggins had been the *Communications Magazine* Managing Editor, and last year we expanded her role to support all three ComSoc magazines. Charis has done a fantastic job, including her great efforts in helping the EiCs and me prepare for the PRAC reviews. I want to thank her for all her hard work and good-natured patience with us volunteers, and wish her well as she leaves this position to pursue her graduate studies. At the same I want to welcome Peggy Kang as the new magazine Managing Editor.

In conclusion, there are several others I would especially like to thank. First, my thanks to Vincent Chan, who appointed me to fill the remaining year of Sergio Benedetto's Director of Magazines term three years ago when Sergio was elected to be President of the Communications Society, and to Katie Wilson for appointing me to continue to serve in this capacity during her term as VP–Publications. There were many things I learned during my experience as *Communications Magazine* EiC. I wanted the opportunity to put structures and tools in place that would help pass along best practice experiences from EiCs across all three magazines and make it easier for future EiCs to transition into their roles. My term as Director of Magazine has allowed me the privilege of doing this. My work has been enjoyable due to the fantastic support from Katie, our great EiCs and AEiCs named above, and consistently wonderful staff support, especially from Charis Scoggins and Joe Milizzo.

Emerging Technologies Committee: Letter from the Chair

Accelerating the Emergence of Emerging Technologies

Zhisheng Niu
Chair, Emerging Technologies Committee

Following the first column, "Promoting Emerging Technologies in ComSoc" in October 2013, and the second column, "Nurturing Emerging Technologies in Clouds" in November 2014, this column is aimed at further accelerating the emergence of emerging technologies. Emerging technologies, as the name shows, should still be emerging from, or only recently emerged from, the research base. In other words, emerging technologies, in particular in our information and communication field, should be up-to-date by nature and therefore should not remain in that category for a long time. However, the reality is that the number of ComSoc subcommittees on emerging technologies is growing quite fast, and some of them have been there for quite a long time (e.g., nine years since establishment). On one hand, every year there are new and interesting technical areas being proposed, which leads to increasing numbers of subcommittees. On the other hand, it also leads to more and more overlap with existing committees and/or subcommittees; hence, management of them becomes a challenge. A key question then arises: what is a good number of emerging subcommittees in our field, and how can we better accelerate the emergence of emerging technologies?

In this regard, the Emerging Techology Committee (ETC) has had thorough discussions through the years, and came up with a so-called 2+2 UP-or-DOWN mechanism, which has been approved by the Technical Activities Council in London, United Kingdom. Here, the key points of the mechanism are:

1) ETC opens its door throughout the years to welcome more proposals on any emerging technology areas.
2) In addition to annual reviews, a rigorous review will be made after two years of operation since formation with the recommendation of:
   a) Elevated/merged to a full TC
   b) Continue as a sub-TC
   c) Put on probation
   d) Disband

If it is rated as b) or c), a critical evaluation and final judgment of either "UP" (elevated to or merged with an existing Technical Committee) or "DOWN" (outdated and therefore disbanded) will be made two years later (i.e., four years since formation). In such a way, the subcommittees can be rotated in a healthy fashion, and more emerging technologies can be expected to emerge in a timely manner. The mechanism will be implemented from 2016; that is, the first rigorous review will be executed at the end of 2016. All the subcommittees that have been operating for more than two years have to go through this review.

By the end of 2014, there were 13 Emerging Technologies subcommittees in total, which were highlighted in the previous two columns. Here, I first outline the scope of the four newly established Emerging Technologies subcommittees. Members with a common interest in these technology areas are strongly encouraged to join the subcommittees.

Technical Subcommittee on Big Data, Established in 2014

The goal of the Technical Subcommittee on Big Data (TSC-BD) is to provide a premier platform for its members, and the research, development, services, applications, and standardization communities of big data processing, analysis, analytics, integration, retrieval, and networking, to interact and exchange technical ideas, identify relevant challenges, and collaborate on and investigate solutions in the development of methodology, and for the science and technologies of big data processing, analysis, analytics, integration, retrieval, and networking.

The technical issues addressed by the subcommittee include all aspects of big data processing, analysis, analytics, integration, retrieval, and related research issues, such as theories, algorithms, solutions, practices, applications, and challenges for big data processing, analysis, analytics, integration, retrieval of information, and communications technologies; machine learning, data mining, web mining, graph mining, and processing; computational intelligence for big data; knowledge discovery for big data; big data for cloud computing and networking; big data for network design and architectures; big data for network protocols; big data for green information and communication technologies; big data for security, privacy, and trust; crowdsourcing and crowd intelligence using big data; big data maintenance; data science; big data platform design; data-intensive workflows; big data benchmarks; reliability for systems with big data; big data for wireless access and mobility; big data for the Internet of Things; big data for software-defined networking; big data for cognitive communications and computing; big data for smart homes; big data for smart sensing; big data for smart grids; big data for relevant signal processing techniques; big data for biomedical and health technologies; big data for social networks; and so on.

Technical Subcommittee on Tactile Internet, Established in 2015

The Tactile Internet (TI) subcommittee will focus on exploring and elucidating all facets of the next generation of "tactile Internet" technology, and business and societal gaps and challenges. The objectives of the TI subcommittee are to facilitate the worldwide harmonization of research, pre-standardization, and best practices for deployment user scenarios of the global TI ecosystem; design built-in security and privacy; and explore ways in which tactile technology can be realized in different segments such as in engineering, automobile, transport and logistics, health service, and public service. The TI subcommittee will target understanding the tactile requirements, specification and identification of tactile use cases, defining system specifications to meet these requirements, developing breakthrough technology to the identified challenges of the tactile Internet, and enabling Internet protocols over the next generation of empowered devices in order to reach convergence and end-to-end transparency through IPv6.

Technical Subcommittee on Quantum Communications and Information Technology, Established in 2015

This sub-TC is aimed at fostering engineering in the newly upcoming quantum technology by applying our (ComSoc's) technical knowledge in areas like RF technology, coding theory, communications and information theory, photonic communications technology, interconnection and complexity theory, error correction, control instrumentation, modeling and simulation, communication systems architecture and hardware, optimized algorithms, and applications, which are all highly required to drive quantum technology forward and get it ready for applications.

TECHNICAL SUBCOMMITTEE ON BACKHAUL NETWORKING AND COMMUNICATIONS, ESTABLISHED IN 2015

There is considerable market interest in the development of small cell backhaul/fronthaul solutions that are an evolution of the existing backhaul/fronthaul technologies (i.e., SDH, ATM, MPLS, and Ethernet). One of the main considerations operators are faced with today is how to migrate existing backhaul/fronthaul infrastructure toward adaptive and smart backhauling/fronthauling solutions that optimize their operations jointly with the access network for the next generation of cellular technology. The deployment availability, cross-layer convergence, and economics of smart backhauling/fronthauling systems are the most important factors in selecting the appropriate backhaul/fronthaul technologies for multiple networks (cellular, WiFi, WiMax, WiGig, etc.); a variety of cell sizes (macro, micro, pico, femto); and multiple technologies (visible light communications, D2D, distributed antennas, etc.).

The aim of this sub-TC is to put forward IEEE's agenda and contribution to the research and standardization activities on future backhaul/fronthaul communications and networking. This sub-TC will create a forum for researchers, developers, and practitioners from both academia and industry to identify and discuss the backhaul/fronthaul requirements, challenges, recent developments, and smart end-to-end solutions pertaining to fifth generation (5G) mobile communication networks. The sub-TC will serve as a prolific opportunity to educate about, promote, and accelerate the evolution of next generation backhaul/fronthaul networking and communications by fostering technical activities in the related area.

Another way the ETC promotes emerging technologies is the *IEEE Journal on Selected Areas in Communications* bonus issue on Emerging Technologies; the first issue was published in May 2015. The key idea here is to showcase the emerging technologies that are cutting-edge and/or interdisciplinary but have no obvious home in other journals by invitation. It is published once a year; all the papers are by invitation and limited to a relatively small number of cutting edge papers on emerging technologies. ETC is the Editorial Board of each issue and responsible for selecting three to four subcommittees to showcase. The Chairs of the selected subcommittees are then responsible for inviting four to six paper submissions in the corresponding fields. Regardless of their invited nature, all papers go through a normal review process based on the standards of *JSAC*, and final acceptance is made by the ETC. In the May 2015 issue, 10 papers out of 23 submissions from the 4 subcommittees (Internet of Things, Social Networks, Green Communications and Computing, and Innovation and Standards in Communication and Information Technologies) were published. The second issue, which is targeted to be published in the second quarter of 2016, is now in the paper review phase under the Guest Editorship of Shuguang Cui (Texas A&M University, ETC member, lead), Tomohiko Taniguchi (Fujitsu Labs, Japan, ETC member), John Thompson (University of Edinburgh, U.K., ETC member), Andrew Eckford (Chair, Nano-Scale, Molecular & Quantum Networking Subcommittee), Latif Ladid (Chair, 5G Mobile Wireless Internet Subcommittee), Vincent Wong (Chair, Smart Grid Subcommittee), and Jie Li (Chair, Big Data Subcommittee). Their hard work and great contributions are highly appreciated. Thus, the selected subcommittees to be showcased are Nano-Scale, Molecular & Quantum Networking, Smart Grid Communications, 5G Mobile Wireless Internet, and Big Data.

Last but not least, I would like to encourage all ComSoc members to participate in ETC subcommittees that intersect with your interests, and to propose new ones associated with emerging technologies in the field of communications and related disciplines. I also encourage all readers to contact me if you have other ideas about how ComSoc can promote and participate in emerging technologies, which will help to maintain its leadership and vision in the field of communications.

Book Reviews

Edited by Piotr Cholda

Nature-Inspired Optimization Algorithms

By Xin-She Yang, Elsevier, 2014, ISBN 978-0-12-416743-8, hardcover, 263 pages

Reviewer: Roza Goscien

Science and technology development brings many benefits to our daily lives. However, alongside these advantages, new challenges arise at the same time. In the case of telecommunication networks, new and very often complex optimization problems emerge. As a result, dedicated efficient algorithms are necessary to provide good quality solutions in a reasonable time. Here, an indispensable help is a group of metaheuristic approaches that were proved to be very efficient even for very complex problems. Among these methods, interesting and effective approaches are based on nature-inspired mechanisms. These methods are the main focus of the Yang's new work. The book is a kind of a compendium, where the reader can find the most important information about a number of optimization nature-inspired algorithms. The presented material combines both verbal and pseudo-code description, as well as mathematical basics. The list of the covered methods is large and the given approaches are up-to-date.

The book consists of 15 chapters and two appendices. In the first three chapters, the author introduces the readers to the basics of algorithms and optimization theory. Here, the definitions of an algorithm and a metaheuristic method are presented. Moreover, the evolutionary operators are discussed with a special focus on their usage motivation and application in different nature-inspired algorithms. The importance of randomization in algorithms is also covered.

Next, the author presents the most important nature-inspired approaches that are widely used to solve different optimization problems. Among the discussed methods the reader can find the following: simulated annealing, genetic algorithms, differential evolution, particle swarm optimization, firefly algorithms, cuckoo search, or bat and flower pollination algorithms. The author shows the base idea of each method and a description of a natural process that provides an inspiration to invent the algorithm. The pseudo-codes and input parameters of the algorithms are also given. The discussion is extended with issues such as different algorithm variants, mathematical description fundamentals, convergence and efficiency analysis, and examples of a method application and implementation.

Subsequently, the author discusses several important issues that can be met when solving different optimization problems. First, a framework for self-tuning algorithms is presented. Second, the approaches that can be applied for constrained optimization problems are discussed (e.g. the Langrangean method, KKT conditions, the penalty method). Third, multi-objective optimization and related solving methods are considered. The discussion is supported by practical examples. At the end of the book, additional nature-inspired algorithms are briefly presented: ant optimization, bee-inspired algorithms, harmony search, and hybrid algorithms. Also appreciated is that in the two appendices, the test function benchmarks for global optimization and Matlab codes of some of the discussed methods are included.

Summarizing, the book is well written and easy to follow, even for algorithmic and mathematical laymen. Since the book focuses on optimization algorithms, it covers a very important and actual topic. It should be noted that the work does not present practical examples related to communication network problems. Rather, the book contains an overview of nature-inspired metaheuristic approaches, thus it is especially recommended for people who start solving problems with such approaches. It will also be useful for those looking for an appropriate nature-inspired algorithm suitable for a defined optimization problem.

Secure Messaging on the Internet

Rolf Oppliger, Artech House, 2014, ISBN 978-1-60807-717-5, hardcover, 265 pages

Reviewer: Marcin Niemiec

Electronic mail service, known as e-mail or mail, has been designed without security needs in mind. However, the importance of private and business communication requires the implementation of security mechanisms. Nowadays, there are a few well established technologies that can secure messages delivered on the Internet. Fortunately, Rolf Oppliger does not focus in his book on a selected solution, such as OpenPGP or S/MIME, but presents a much broader view. His book is a comprehensive summary of technologies that provide security for Internet messaging. The work is well organized and easy to read.

The book contains 13 chapters, as well as appendices including a description of popular character sets, widely used transfer encoding schemes, ASN.1 syntax, encoding rules, and public key cryptography standards. Each part of the work can be read independently, but only learning all of them will allow a reader to obtain a broad view of all the aspects related to secure messaging on the Internet.

After the short introductory Chapter 1, the author devotes Chapter 2 to a brief overview of the core technologies used for Internet messaging. The Internet mail architecture and its main components are introduced first. Next, the Internet message format (including the header section and message body) is described. The crucial protocols used for message store access, transfer, and delivery (e.g. SMTP, POP, IMAP) are also presented. Chapter 3 introduces the basics of cryptography: random number generators, hash functions, symmetric ciphers, public key cryptosystems, etc. The next chapter continues with this subject and describes the most popular public key certificates: X.509 and OpenPGP. The formats and trust models are described in detail. Chapter 5 offers a brief introduction to the secure messaging concept. Various approaches to secure messaging are discussed and general threats as well as attacks are elaborated. The next two chapters present the most popular secure messaging technologies/standards, i.e. OpenPGP and S/MIME. After presenting their history and the rationale for them, the author presents technology details and discusses the trust model employed by OpenPGP. The chapter related to S/MIME introduces the technology and usage of certificates in the hierarchical trust model. Both chapters include short security analysis sections. Chapter 8 is devoted to web-based messaging and its security. In addition to a brief introduction of this concept and the reference architecture, the author presents example service providers. Chapters 9 to 11 are devoted to gateway solutions, a certified mail, and instant messaging, respectively. In all these chapters the readers will find example products and solutions as well as security considerations. In Chapter 12 research challenges and open questions are considered, namely spam protection, P2P technologies, new approaches, and protections. The last chapter summarizes the book.

I strongly recommend this book. The greatest advantage of this book is the fact that it collects a comprehensive and complete summary of secure Internet messaging in one place. It is worth noting that the book is written with no unnecessary complexity and is not overloaded with any theoretical data. The simplicity is an unquestionable advantage of this work. In my opinion, this book is worth recommending to network and security professionals who are interested in Internet messaging services. University researchers as well as graduate students in communications and computer science will find interesting and useful topics here. Also, system architects and developers will find this book valuable. Some parts include basic principles and ideas, so beginners will be able to learn the basics. Professionals will appreciate that each chapter includes a discussion section ('Final Remarks') and extensive bibliographic notes.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

December 2015
ISSN 2374-1082

CHAPTER REPORT

Recent Activities in the ComSoc Southwestern USA Region 5

By T. Scott Atkinson, North America Region Board Member, USA

This is a summary highlighting some of the ComSoc Chapter activities in IEEE ComSoc Southwestern USA Region 5, which includes the states of Texas, Louisiana, Arkansas, Missouri, Kansas, Colorado, Oklahoma, Texas, and parts of New Mexico and South Dakota.

IEEE Central Texas Section COMSOC/SP Austin Joint Chapter

The Austin Chapter hosts technical meetings regularly during the year. Following are the topics of the most recent meetings.

September: "Multicore Processor Solutions for Smart Factories, Smart Cities and Smart Energy." Presenter: Mr. Altaf Hussain, Business Development Manager, Freescale's Digital Networking group.

October: "Machine Learning and Signal Processing Methods in Live Business Intelligence Operations." Presenter: Dr. Choundur Lakshminarayan, Principal Research Scientist, HP.

November: "Trustworthy Hardware." Presenter: Distinguished Prof. Ramesh Karri, Prof. ECE, Ploytechnic Institute, NYU.

In November, chapter members celebrated their two ComSoc awards:

•2015 ComSoc North America Region Chapter Achievement Award.

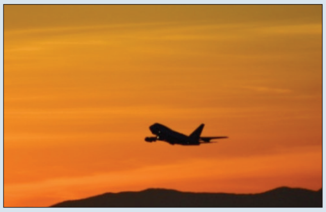
•2015 ComSoc Chapter of the Year Award.

IEEE Central Texas Section COMSOC/SP San Antonio Joint Chapter

In February, the chapter hosted Distinguished Lecturer Prof. Koichi Asatani at a joint meeting with the San Antonio Life Members Group and the Student Branch at the University of Texas at San Antonio.

In June, at a meeting held jointly with the San Antonio Life Members Group, Scott Atkinson spoke on the topic of NASA's Stratospheric Observatory for Infrared Astronomy (SOFIA) Project.

In October the ComSoc/SP Joint Chapter participated in the Engineering in Medicine and Biology Society Chapter Distinguished Lecturer meeting, along with Computer Society members. Dr. Michael Jirjis spoke on "Diffusion Tensor Imaging of the Central Nervous System Following Spinal Cord Injury and Stem Cell Transplant."



NASA's Stratospheric Observatory for Infrared Astronomy (SOFIA) takes off from Palmdale, California at sunset. SOFIA is a partnership of NASA and the German Aerospace Center (DLR); NASA and DLR have collaborated on a range of activities and signed agreements on June 16 to work together to reduce aircraft noise and advance research into rotorcraft.

IEEE ComSoc Galveston Bay Chapter

The August Section/Chapter meeting focused on a seminar presentation on the topic "A Glimpse of the Future World: When Integrated Unmanned Systems Meet Artificial Intelligence," presented by Hao Xu, Ph.D., Director of the Unmanned Systems Laboratory (TAMUCC-USL).

Unmanned systems (US) (unmanned aircraft systems (UAS), unmanned ground vehicles (UGV), and unmanned underwater robots) are smart mission-based agents equipped with well-developed on-board sensors, microprocessors and communication devices (e.g. lasers, GPS, HD-cameras, sonars, optical/RF transceivers, etc.). USs can successfully complete numerous tasks such as disaster search/rescue, emergency response in extreme environments, etc., which human beings would not be able to finish easily. Recently, different types of USs have been integrated to cover broader operations. However, due to a lack of effective integration methods, emerging integrated unmanned systems (IUSs) are still very difficult to be implemented into practical systems. To overcome this challenge, the cutting-edge artificial intelligence (AI) technique is being adopted to work with IUSs. This talk covered the background of USs and AI and their potential IUS applications, enumerating their applications and challenges. Effective AI-based algorithms will be developed to improve the performance of IUSs, leading to their future applications.

In October, the Chapter supported the Section in the coordination of the Section's annual DUAL Conference of Innovation and Automation 2015.

In November, a unique seminar was organized by the Chapter for the Section meeting. The topic of the seminar was "Engineering Contributions to Medicine: What Else Can Be Done?" The presenters were Paul Frenger, MD, SM-IEEE, SM-ISA, LM-ACM (practicing physician); Reese Terry, Life Fellow-IEEE, retired founder and CEO of Cyberonics; John Criscione, MD, Ph.D., associate professor, Texas A&M; and Ershad Sharifahmadian, Ph.D., visiting professor, University of Houston Clear Lake.

IEEE ComSoc Tulsa Chapter

The IEEE ComSoc Tulsa Chapter offers a weekly seminar (Tuesday 11:45 AM-1:00 PM) for all IEEE ComSoc Tulsa Chapter members. These weekly seminars are presented by The University of Oklahoma–Tulsa's MS or Ph.D. students, or occasionally by a guest lecturer. These seminars are related to telecommunication technologies such as wireless communication, quantum communication, image processing, etc. Following are titles of the most recent seminars:

•October 6, 2015: "Research and Analysis of Medication Adherence in an Adult Type 2 Diabetes Cohort."

•October 13, 2015: "Big Data!"

•October 20, 2015: "Secure AES Key Transmission Using Polarization Encoding Over Optical Fibers."

•October 27, 2015: "Development and Implementation of a Versatile Real-Time Traffic Monitoring System Using Wireless Smart Sensors Networks."

•November 3, 2015: "Multi-Photon Tolerant Approach for Satellite Communications."

(Continued on Newsletter page 4)

CONFERENCE REPORT

First IEEE Convention of Electrical & Electronics Engineering Students in Israel

An Overview of the Activities of the IEEE Comsoc Israel Chapter

By Yiftach Richter, ComSoc Student Member, Israel, and Itsik Bergel, ComSoc Senior Member and Head of Student Activities of the IEEE Israel Section

First IEEE Convention of Electrical & Electronics Engineering Students

Israel is well known for its label as the startup nation, and engineering at Israeli universities is flourishing as well. However, so far IEEE student members have had very few organized activities. This dramatically changed this year at the First Convention of Electrical & Electronics Engineering Students in Israel (IEEEIs 2015). IEEEIs 2015 was held at Bar-Ilan University, Ramat-Gan, and featured keynotes, oral presentations, poster sessions, and a job fair.

The convention targeted mostly graduate students, and gave them an opportunity to present their research in a professional yet casual environment. The convention was also a platform for meetings and exchanges of ideas between students and between students and senior researchers. The job fair also focused on graduate students, and featured leading high-tech companies. As an example, one of the participating companies, Mellanox LTD, offered more than 50 positions.

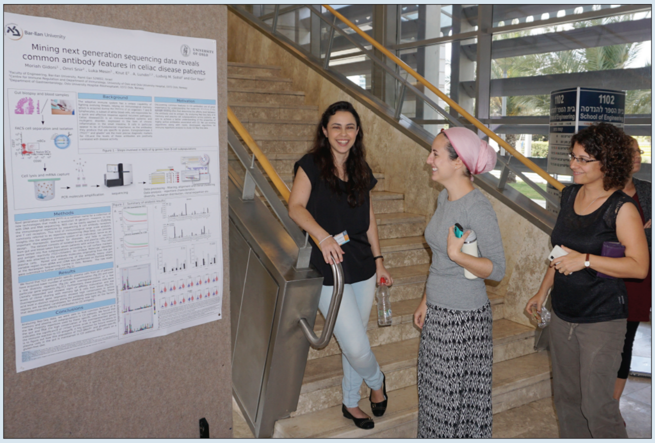
As appropriate to an Israeli student conference, all keynotes were startup oriented (which also attracted B.Sc. students). Rafi Nave, a leading Israeli hi-tech entrepreneur and manager of the Entrepreneurship Center at the Technion, presented a talk titled "Hi-Tech Nation." In his talk Mr. Nave surveyed that main strengths that made Israel a hi-tech nation, and pointed out that more than 50 percent of the GDP of Israel comes from the hi-tech industry.

The second keynote, titled "Photonic Ear for Remote Detection and Diagnostic of Diseases," was presented by Prof. Zeev Zalevsky, the vice dean of the Faculty of Engineering at Bar Ilan University. Prof. Zalevsky presented a few of his cutting-edge innovations involving the development of a laser based technology allowing remote and continuous sensing of nano vibrations of different tissues. The translation and analysis of the nano vibrations can be used for the detection of various diseases, as well as for establishing a very directional and noise immune sound-communication channel. The realization of the above mentioned capabilities is obtained by proper image processing of the time-space varying secondary speckle patterns generated by the back reflection of the laser light from the inspected tissue.

The convention was closed by a keynote by Dr. Akihiko Sugiyama, who is an IEEE distinguished lecturer on signal processing. The talk, titled "What I Wish I Knew When I was an Entry-Level Engineer," compiled Dr. Sugiyama career experience into a useful guide to the starting engineer.

Most importantly, between the keynotes, IEEE student members had an opportunity to present their novel research in eight poster and oral sessions. These sessions covered all topics of electrical and electronics engineering, including communication, signal processing, information theory, electro-optics bio-engineering, and more. Merav Passig-Antman, an IEEE student who presented her work at the convention, said, "I was greatly excited to participate in the IEEEIs convention. The convention was stimulating not only academically but socially as well. Overall, the experience was excellent!"

Attendees networking between sessions.


Attendees discussing one of the poster presentations.

After the great success of the first convention, EE students in Israel are looking forward to the next convention!

Overview Of Activities Of The IEEE Comsoc Israel Chapter

The Israeli IEEE COMSOC chapter supports the CE-Club, a Technion Computer Engineering Club that holds a weekly seminar discussing systems and theory of networking, distributed systems, and more. This club is a joint collaboration of Technion's Computer Science and Electrical Engineering faculties. We meet every Wednesday at 11:30, alternating between both faculties, to hear an interesting research talk. We take a broad view of what constitutes "interesting," both systems and theory of networking, distributed systems, storage, computer architectures, operating systems, compilers, etc.

Our COMSOC members are also invited each year (our 10th year now) to "Israeli Networking Day," which brings together members of the Israeli networking community from both academia and industry to discuss recent developments and research in this vibrant area and to foster new collaborations. The presentations in this event are focused on both published research results and on work in progress. In addition, we invite our members to the CYBERDAY workshop on Cyber and Computer Security in the Computer Science department at the Technion, together with the Technion Computer Engineering center. This workshop has been held since 2008, formerly known as the "CRYPTODAY" workshop.

DISTINGUISHED LECTURER TOUR

Lecture of Andrea Goldsmith at the Nanjing Chapter

By Yueming Cai and Shi Jin, IEEE ComSoc Nanjing Chapter, China

Prof. Goldsmith during her lecture at Southeast University.

Prof. Goldsmith with members of the IEEE ComSoc Nanjing Chapter.

Professor Andrea Goldsmith from Stanford University was invited to give a lecture titled "The Road Ahead for Wireless Technology: Dreams and Challenges" at Southeast University, Nanjing, China on June 26, 2015. The lecture was supported by the IEEE Communications Society Nanjing Chapter and hosted by Prof. Nan Liu from Southeast University. More than 260 people attended this seminar, including academic and research staff, professionals, and students from the IEEE ComSoc Nanjing Chapter.

In the lecture, Prof. Goldsmith presented visions about the future of wireless communications and discussed some innovations and breakthroughs in wireless technologies that are required to realize the visions. First she presented an overview of advanced wireless technologies, which include millimeter wave, massive MIMO, non-coherent massive MIMO, small cell, sub-Nyquist sampling, etc. Then Prof. Goldsmith focused on new advances in green wireless communications. She pointed out that the biggest problem with WiFi and the limitation of small cell is resource contention. She elaborated and explained that the shortage of spectrum could be alleviated by research advances in cognitive radios. Further, she mentioned that breakthroughs in energy-efficiency algorithms and hardware would be employed to make wireless systems "green." Finally, Prof. Goldsmith proposed that most of these research advances are interdisciplinary and a synergistic exploration of knowledge bases from multiple technical domains is required for future research. In the Q&A session, Prof. Goldsmith gave detailed answers to questions about full duplex and the relationship between small cell and D2D techniques. The lecture was very successful and gave audiences a better understanding of green wireless communication and also stimulated their interests in exploring this area.

Andrea Goldsmith is the Stephen Harris Professor in the

(Continued on Newsletter page 4)

CONFERENCE REPORT

First IEEE International Workshop on Security and Privacy for Internet of Things and Cyber-Physical Systems (IOT/CPS-Security 2015)

ICC2015, London, UK

By Qinghe Du, Eirini Karapistoli, Heath LeBlanc,
IEEE IoT/CPS-Security 2015 Workshop Publicity Co-Chairs

The 2015 IEEE International Conference on Communications (ICC) was held in London, UK from 8-12 June. Themed "Smart City & Smart World," this flagship conference of the IEEE Communications Society attracted a record-breaking 2,947 attendees. The technical program of IEEE ICC 2015 consisted of 12 symposia, 27 workshops, 20 tutorials, as well as 18 industry panels. ICC 2015 also featured seven keynote speeches. The First IEEE International Workshop on Security and Privacy for Internet of Things and Cyber-Physical Systems (IOT/CPS-Security 2015) was one of 24 full-day workshops in conjunction with ICC 2015.

IoT/CPS-Security 2015 Keynote Speech by Prof. Wei Yu.

IoT/CPS-Security 2015 Keynote Speech by Dr. Jesús Alonso-Zárate.

Recent advances in networking, communications, computation, software, and hardware technologies have revolutionized the way humans, smart things, and engineered systems interact and exchange information. The Internet of Things (IoT) and Cyber-Physical Systems (CPS), which are the major contributors to this area, will fuel the realization of this new, globally interconnected cyber-world. Yet the success, prosperity, and advancement of IoT and CPS systems strongly depend on the security, privacy, and trust of the IoT and cyber-physical devices as well as the sensitive data being exchanged. While these technologies offer many new possibilities, the increasing complexity of hardware and software as well as

(Continued on Newsletter page 4)

IOT/CPS Security/*Continued from page 3*

the worldwide access increase the vulnerability to security attacks. Successful attacks targeted to IoT devices and CPS systems have in common that not only a single computer is affected, but also interconnected technical systems allowing interaction with the physical world are influenced leading to malfunction of devices and control systems with severe financial, environmental, and health losses. This fact highlights the need to develop novel tools that will constitute the heart of a much-needed science of security for IoT and CPS and will assist in building resilient, secure, and dependable networked systems. The aim of the IEEE IOT/CPS-Security workshop series, with the first edition in 2015, is to foster a research community committed to advancing research and education at the confluence of cybersecurity, privacy, Internet of Things, and cyber-physical systems, and to transitioning its findings into engineering practice.

IOT/CPS-Security 2015 was a joint effort among active researchers on three continents: Europe, Asia, and North America. Anastasios A. Economides serves as a professor at the University of Macedonia, Thessaloniki, Greece, and the director of the Computer Networks & Telematics Applications (CONTA) Lab. Minho Jo serves as a professor at Korea University, Sejong Metropolitan City, South Korea, and the Editor-in-Chief of *KSII Transactions on Internet and Information Systems*. Houbing Song serves as an assistant professor in the Department of Electrical and Computer Engineering at West Virginia University, Montgomery, WV, USA, and the director of both the West Virginia Center of Excellence for Cyber-Physical Systems (WVCECPS), sponsored by the West Virginia Higher Education Policy Commission, and the Security and Optimization for Networked Globe Laboratory (SONG Lab). Houbing Song is the lead editor of two books, *Cyber-Physical Systems: Foundations, Principles and Applications* and *Security and Privacy in Cyber-Physical Systems: Foundations and Applications*, both of which will be published in 2016 by Elsevier and Wiley, respectively.

The General Chairs of IoT/CPS-Security 2015 were Anastasios A. Economides, University of Macedonia; Houbing Song, West Virginia University; Minho Jo, Korea University; and Daqiang Zhang, Tongji University. The Technical Program Chairs of IoT/CPS-Security 2015 were Krishna Kumar Venkatasubramanian, Worcester Polytechnic Institute; Eirini Karapistoli, University of Western Macedonia; Vasilis Friderikos, King's College London; João Paulo Miranda, Center for Research and Development (CPqD); Dev Audsin, Orange; and Jianguo Ding, University of Skövde.

IoT/CPS-Security 2015 received 49 papers and accepted 12 papers. Papers were submitted from 24 different countries and regions across five continents: North America, Europe, Asia, Africa and Australia.

IoT/CPS-Security 2015 also featured two keynote speeches by Dr. Jesús Alonso-Zárate, who is a senior research associate and the head of the Machine-to-Machine Communications (M2M) Department at the Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Barcelona, Spain, and Prof. Wei Yu, who is an associate professor in the Department of Computer & Information Sciences at Towson University, Towson, MD, USA. Their talks were entitled "Enabling Autonomous Communications between Machines, Humans, and Things" and "Secure Energy-Based Cyber-Physical Systems," respectively.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

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

IEEE Communications Society

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

Regional Correspondents Who Contributed to This Issue

Ewell Tan, Singapore (ewell.tan@ieee.org)

A publication of the
IEEE Communications Society

www.comsoc.org/gcn
ISSN 2374-1082

Region 5 Activities/*Continued from page 1*

IEEE ComSoc New Orleans Chapter

February: Distinguished Lecturer Tour. Topic: "Realizing FTTH, G-PON is widely adopted, international standards on FTTH, G-PON in particular, have been established and are being further developed." Presenter: Dr. Koichi Asatani, IEEE ComSoc Distinguished Lecturer, IEEE Fellow, IEICE Fellow, Ph.D., lecture professor, Nankai University, Tianjin, China; professor emeritus, Kogakuin University, Tokyo, Japan.

May: Joint Section/ComSoc Chapter Meeting. Topic: "Cellular and Wi-Fi communications systems design that were installed in the Mercedes Benz Louisiana Superdome and Arena prior to the 2013 Super Bowl and the engineering involved with this effort." Presenters: Leo L. Holzenthal, Jr., P.E., ACFE Fellow, president and engineering manager for M S Benbow and Associates; Ken M. Wright, P.E., manager of technology and telecommunications for M S Benbow and Associates.

Conference Coordination: The ComSoc Chapter supported the Section in hosting the IEEE WCNC 2015 conference, as well as the 2015 Region 5 GreenTech Conference in New Orleans.

Distinguished Lecturer Tour/*Continued from page 3*

School of Engineering and a professor of electrical engineering at Stanford University. She is a Fellow of the IEEE and of Stanford, and she has received several awards for her work, including the IEEE Communications Society and Information Theory Society joint paper award, the IEEE Communications Society Best Tutorial Paper Award, the National Academy of Engineering Gilbreth Lecture Award, the IEEE Communication Theory Technical Committee Recognition Award, the IEEE Wireless Communications Technical Committee Recognition Award, the Alfred P. Sloan Fellowship, and the Silicon Valley/San Jose Business Journal's Women of Influence Award. She is author of the book *Wireless Communications* and co-author of the books *MIMO Wireless Communications* and *Principles of Cognitive Radio*, all published by Cambridge University Press. She received the B.S., M.S. and Ph.D. degrees in electrical engineering from U.C. Berkeley.

Conference Calendar

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

2016

JANUARY

COMSNETS 2016 — 8th Int'l. Conference on Communication Systems & Networks, 5–9 Jan.

Bangalore, India
http://www.comsnets.org/index.html

IEEE CCNC 2016 — IEEE Consumer Communications and Networking Conference, 8–11 Jan.

Las Vegas, NV
http://ccnc2016.ieee-ccnc.org/

WONS 2016 — 12th Annual Conference on Wireless On-Demand Network Systems and Services, 20–22 Jan.

Cortina d'Ampezzo, Italy
http://2016.wons-conference.org/

ICACT 2016 — 18th Int'l. Conference on Advanced Communication Technology, 31 Jan.–2 Feb.

Phoenix Park, Pyeongchang, Korea
http://www.icact.org/

FEBRUARY

IEEE BHI 2016 — IEEE Int'l. Conference on Biomedical and Health Informatics, 24–27 Feb.

Las Vegas, NV
http://bhi.embs.org/2016/

MARCH

DRCN 2016 — 12th Int'l. Workshop on Design of Reliable Communication Networks, 14–17 March

Paris, France
https://drcn2016.lip6.fr/

ICBDSC 2016 — 3rd MEC Int'l. Conference on Big Data and Smart City, 15–16 Mar.

Muscat, Oman
http://www.mec.edu.om/conf2016/index.html

OFC 2016 — Optical Fiber Conference, 20–24 Mar.

Anaheim, CA
http://www.ofcconference.org/en-us/home/

IEEE CogSIMA 2016 — IEEE Int'l. Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, 21–25 Mar.

San Diego, CA
http://www.cogsima2016.org/

WD 2016 — Wireless Days 2016, 23–25 Mar.

Toulouse, France
http://wd2015.sciencesconf.org/

IEEE ISPLC 2016 — 2016 IEEE Int'l. Symposium on Power Line Communications and Its Applications, 29 Mar.–1 Apr.

Bottrop, Germany
http://www.ieee-isplc.org/

APRIL

IEEE WCNC 2016 — IEEE Wireless Communications and Networking Conference, 3–6 Apr.

Doha, Qatar
http://wcnc2016.ieee-wcnc.org/

IEEE INFOCOM 2016 — IEEE Int'l. Conference on Computer Communications, 10–15 April

San Francisco, CA
http://infocom2016.ieee-infocom.org/

WTS 2016 — Wireless Telecommunications Symposium, 18–20 Apr.

London, U.K.
http://www.cpp.edu/~wtsi/

IEEE/IFIP NOMS 2016 — IEEE/IFIP Network Operations and Management Symposium, 25–29 Apr.

Istanbul, Turkey
http://noms2016.ieee-noms.org/

MAY

IEEE CQR 2016 — IEEE Int'l. Communications Quality and Reliability Workshop, 9–12 May

Stevenson, WA
http://www.ieee-cqr.org/

ONDM 2016 — Int'l. Conference on Optical Network Design and Modeling, 9–12 May

Cartagena, Spain
http://ondm2016.upct.es/index.php

IEEE CTW 2016 — IEEE Communication Theory Workshop, 15–18 May

Nafplio, Greece
http://www.ieee-ctw.org/

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 LANMAN 2016 — 22nd IEEE Workshop on Local & Metropolitan Area Networks, 13–15 June

Rome, Italy
http://www.ieee-lanman.org/

IEEE HPSR 2016 — IEEE 17th Int'l. Conference on High Performance Switching and Routing, 14–17 June

Yokohama, Japan
http://www.ieee-hpsr.org/

IEEE IWQOS — IEEE Int'l. Symposium on Quality and Service, 20–21 June

Beijing, China
http://www.dongliangxie.com/

EUCNC 2016 — European Conference on Networks and Communications, 27–30 June

Athens, Greece
http://eucnc.eu/

–Communications Society portfolio events appear in bold colored print.

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

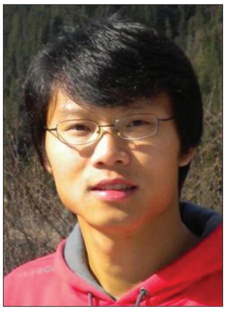
–Sister Society conferences appear in plain black print.

–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.

Guest Editorial

Wireless Physical Layer Security

Walid Saad


Xiangyun Zhou

Mérouane Debbah

H. Vincent Poor

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

In this second issue of the Feature Topic on wireless physical layer security, the first article by Lin *et al.* investigates the impact of channel state information (CSI) on wireless secrecy. In this regard, the authors expose how different levels of CSI may affect confidentiality in terms of information-theoretic secrecy. Then the next article, by Bash *et al.*, studies the use of covert communication techniques that can counter security threats from adversaries that use non-computational methods, such as side-channel analysis, to jeopardize wireless transmissions. Various secrecy signaling and coding schemes have been designed at the physical layer of wireless systems to guarantee confidentiality against information leakage to unauthorized receivers, among which the strategy based on the idea of node cooperation is promising and is discussed in the following three articles. In this regard, the work by Jimenez *et al.* provides a broad overview of this area while discussing one case study to quantify the benefits of relay resource allocation for improving wireless secrecy. Next, Chen *et al.* focus on scenarios in which relays are equipped with multiple antennas. For such settings, the authors discuss how one can exploit MIMO techniques to further enhance cooperation and boost the secrecy of wireless transmission. The next article provides a signal processing approach to the problem of wireless cooperation, and focuses on secrecy signal design and optimization techniques to increase secrecy performance. The privacy of a wireless user and the operation of a wireless network can be threatened by the leakage of transmission signatures, even when encryption and authentication services are employed. The Feature Topic concludes with an article by Rahbari and Krunz that describes various passive (traffic analysis) and active (jamming) attacks that are facilitated by side-channel information. The goal is to highlight the need for novel physical-layer security techniques that can be used to complement classical encryption methods.

In a nutshell, given the significant advances in physical layer security of the past decade, this Feature Topic provides an in-depth exposition of the various challenges faced, and that will continue to be faced, in the field of wireless physical layer security. We hope that these contributions will initiate future research developments in this field and contribute toward introducing physical layer security schemes in practical scenarios.

Acknowledgments

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

Biographies

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

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

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

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

To Avoid or Not to Avoid CSI Leakage in Physical Layer Secret Communication Systems

Ta-Yuan Liu, Pin-Hsun Lin, Shih-Chun Lin, Y.-W. Peter Hong, and Eduard Axel Jorswieck

Ta-Yuan Liu and Y.-W. Peter Hong are with National Tsing Hua University.

Pin-Hsun Lin and Eduard Axel Jorswieck are with Technische Universität Dresden.

Shih-Chun Lin is with National Taiwan University of Science and Technology.

ABSTRACT

Physical layer secrecy has attracted much attention in recent years due to its ability to ensure communication secrecy with the use of channel coding and signal processing techniques (and without the explicit use of secret keys) in the physical layer. It serves as a promising technique for highly dynamic or ad hoc systems such as device-to-device and machine-type communication systems. However, the achievable secrecy performance depends highly on the level of CSI at the transmitter, the receiver, and the eavesdropper. In this article, we discuss how different levels of CSI resulting from conventional and unconventional ways of performing training and channel feedback may affect the confidentiality in terms of the information-theoretic (perfect) secrecy rate. The conventional approach refers to the emission of pilot signals from the transmitter and explicit channel feedback from the receiver. This approach is backward compatible with existing systems and allows the receiver to obtain accurate knowledge of the CSI, but may suffer from CSI leakage toward the eavesdropper. Unconventional approaches capitalize on reverse training to prevent CSI leakage and are shown to achieve significant improvements over conventional schemes in certain cases. For example, in a system with four transmit antennas and a single antenna at both the receiver and the eavesdropper, a secrecy rate gain of approximately 0.8 b/channel use at transmit SNR of 16 dB is observed over the full CSI case by providing CSI only to the transmitter (but not the receiver and the eavesdropper).

INTRODUCTION

Security in wireless communications has always been a major concern due to the broadcast nature of wireless transmissions. Conventionally, these issues have been addressed in the upper layers of the network protocol stack using cryptography-based solutions, which typically rely on the use of confidential secret keys to seal the transmitted messages. However, with the rapid growth of the number of wireless devices, the secret key distribution and management that are required to maintain these operations are becoming increasingly difficult, and are introducing larger overhead and latency to the system. For example, in the Long Term Evolution Advanced (LTE-A) system, the authentication and key agreement (AKA) process takes up to several hundreds of milliseconds for key computations and distribution. The latency and the additional burden on the backhaul may increase even more rapidly with the introduction of machine-type communications and the Internet of things (IoT).

Interestingly, recent information-theoretic studies of the wiretap channel have demonstrated the possibility of achieving secrecy in the physical layer with the sole use of channel coding and signal processing techniques (i.e., without the explicit use of secret keys). However, many of these fundamental studies make ideal assumptions on the channel state information (CSI) at the transmitter, the receiver, and/or the eavesdropper, and ignore the practical issues of training and channel feedback. However, in practice, CSI cannot be obtained for free and is often subject to imperfections. For example, in LTE-A, a downlink pilot time slot (DwPTS) is allocated for the emission of pilot symbols by the base station to enable channel estimation at the user equipment (UE), and a physical uplink control channel (PUCCH) is utilized for CSI feedback from the UEs to facilitate closed-loop transmissions when the coherence time is sufficiently long. Even with the dedicated resources for training and feedback, CSI at the receiver and the transmitter is never perfect due to channel estimation errors and limited feedback bandwidth. These issues must be taken into consideration when employing physical layer secrecy techniques in practice.

In this article, we first discuss how different levels of CSI at the transmitter, which result from conventional ways of performing training and channel feedback, may impact the confidentiality in terms of the information-theoretic (perfect) secrecy rate. We show that the conventional approach of having the transmitter emit the training signal and having the receiver feed back the estimated channel to the transmitter is compatible with existing systems, and allows the receiver

0163-6804/15/$25.00 © 2015 IEEE

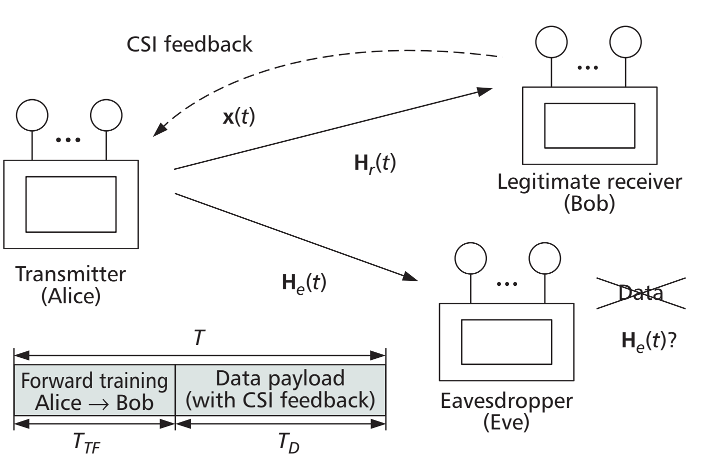


Figure 1. The wiretap channel consists of a transmitter, a receiver, and an eavesdropper with conventional training-based transmission where there is only forward training for Bob to estimate channel state information.

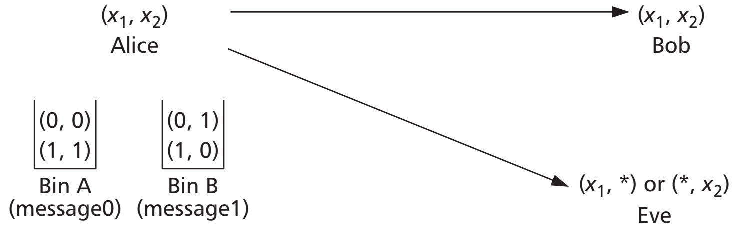


Figure 2. An illustration of secrecy binning.

and the transmitter to obtain accurate knowledge of the CSI. However, this approach does not prevent the eavesdropper from obtaining CSI, and in order may reduce the achievable secrecy rate. To avoid CSI leakage, we then review several novel techniques proposed in the literature that capitalize on the use of reverse training, that is, training with pilot signals emitted by the receiver. Reverse training enables channel estimation directly at the transmitter, but does not benefit estimation of the transmitter-to-eavesdropper channel at the eavesdropper. Finally, numerical simulations are provided to demonstrate the gains that can be obtained with the prevention of CSI leakage in terms of the achievable secrecy rate. However, it is worthwhile to note that the schemes which avoid CSI leakage may not be suitable for all scenarios due to their need for the use of reverse training and artificial noise. Moreover, by assuming that perfect CSI is available at the eavesdropper, schemes that do not avoid CSI leakage can be viewed as worst-case schemes which can be used when the level of CSI at the eavesdropper is uncertain.

Overview of Physical Layer Secrecy

A basic secret communication system (often referred to as the wiretap channel in the information theory literature) consists of three terminals: a transmitter (Alice), a legitimate receiver (Bob), and an eavesdropper (Eve), as illustrated in Fig. 1. For the above channel, the seminal works [1, 2] (and many works that follow) proved the existence of channel codes that can be used to send confidential messages from the transmitter to the receiver with arbitrarily low error probability while ensuring asymptotically zero information rate at Eve (i.e., perfect secrecy). The transmission rate achievable under these conditions is referred to as the secrecy rate, and the maximum of such rates is the secrecy capacity.

Let $\mathbf{x}(t)$ be the symbol vector sent by the transmitter in the tth coherence block, and let

$$\begin{aligned}\mathbf{y}_r(t) &= \mathbf{H}_r(t)\mathbf{x}(t) + \mathbf{z}_r(t), \text{ and}\\ \mathbf{y}_e(t) &= \mathbf{H}_e(t)\mathbf{x}(t) + \mathbf{z}_e(t),\end{aligned} \tag{1}$$

be the received signals at Bob and Eve, respectively, where $\mathbf{H}_r(t)$ and $\mathbf{H}_e(t)$ are the corresponding channel matrices in block t, and $\mathbf{z}_r(t)$ and $\mathbf{z}_e(t)$ are the corresponding additive white Gaussian noise (AWGN) vectors. We assume that the channel is block fading with coherence time T, with a value that depends on whether the channel is fast or slow fading. The ergodic secrecy rate is considered as the main performance criterion throughout this article.

To ensure confidentiality without secret keys, a coding technique called "secrecy binning" is employed at the heart of many physical layer secrecy techniques. The key idea is to insert additional randomness into the codeword of each confidential message to increase the uncertainty at Eve. The amount of randomness required to achieve perfect secrecy is often proportional to the capacity of the eavesdropper channel (Alice-Eve). An example is given as follows.

Example I[1]: We consider the wiretap channel in Fig. 2 with input $\mathbf{x} = (x_1, x_2)$ being a vector of binary entries, that is, $x_1, x_2 \in \{0,1\}$. The channel outputs at Bob and Eve are $\mathbf{y}_r = (x_1, x_2)$ and $\mathbf{y}_e = (x_1, *)$ or $(*, x_2)$, respectively, where $*$ denotes an erasure. The reception at Bob is perfect, but that Eve may have an erasure in one of the two entries. To transmit a secret binary message to Bob, Alice constructs two secrecy bins corresponding to message bits 0 and 1: Bin A and Bin B, respectively. Bins A and B consist of codewords {(0, 0), (1, 1)} and {(0, 1), (1, 0)}, respectively. During each transmission, Alice randomly sends a codeword from the bin corresponding to the secret message. The message can be successfully decoded by Bob since the main channel is noiseless, but cannot be decoded by Eve since Eve receives only one entry of the codeword. By randomly choosing codewords within a bin, Eve faces one bit of uncertainty. This one bit is equal to the capacity of the eavesdropper channel, and is exactly the amount needed to prevent eavesdropping at Eve.

Notice that the efficiency of the secrecy binning technique depends on the discrepancy between the quality of the main and eavesdropper channels. Therefore, to enhance secrecy, signal processing techniques have been employed on top of the secrecy-binning-based coding schemes to further enlarge the channel quality discrepancy by artificially generating a better channel for Bob than for Eve. For example, with multiple antennas at the transmitter, one can employ the so-called secrecy beamforming tech-

[1] Please refer to S. El Rouayheb, E. Soljanin, and A. Sprintson, "Secure Network Coding for Wiretap Networks of Type II," *IEEE Trans. Info. Theory*, vol. 58, no. 3, Mar. 2012, pp. 1361–71.

nique where the message-bearing signal is directed toward a spatial dimension that yields a better channel quality for Bob than for Eve. Secrecy beamforming is known to maximize the secrecy capacity of the multiple-input single-output channel when perfect CSI is available at all terminals [3, 4]. Under non-ideal CSI assumptions, artificial noise (AN) [5] can be super-imposed on top of the message-bearing signal in an appropriately chosen signal subspace to cause additional interference at Eve. In the following sections, we discuss how different CSI assumptions resulting from conventional and unconventional ways of doing training and channel feedback may impact the achievable secrecy rate, especially through the prevention of CSI leakage (or lack thereof).

Physical Layer Secrecy without Protection against CSI Leakage

In this section, we discuss the impact of CSI assumptions that result from conventional training and channel feedback schemes. Conventionally, training is performed in the forward direction by having Alice emit pilot signals at the beginning of each coherence interval to enable channel estimation at Bob. Channel feedback is then sent from Bob to Alice. This approach can easily be applied to current wireless standards without major modifications of the signaling mechanism since it adheres to the traditional frame structure, as illustrated in Fig. 1. Moreover, since these schemes make no attempt to prevent CSI leakage to Eve, they are often devised by assuming perfect CSI at Eve, and thus are suitable for worst-case scenarios.

In the following, we first consider cases where perfect CSI is available at both Bob and Eve (due to perfect forward training), and further categorize the results according to three different assumptions on the CSI at the transmitter (CSIT). The CSIT assumptions are consequences of perfect (without delay), delayed, and partial feedback operations from Bob. Then, by taking into consideration the channel estimation errors, we further discuss the trade-off between training and secret data transmission in the conventional setting.

Secrecy with Perfect Main-Channel CSIT

In this subsection, we consider the case where perfect main-channel (Alice-Bob) CSIT is available instantaneously (without delay), which requires sufficiently large feedback bandwidth from Bob to Alice. Ideally, when perfect eavesdropper-channel CSIT is also available, secrecy beamforming accompanied by the Gaussian (secrecy) binning codebook is known to be secrecy-capacity achieving in a multiple-input single-output multiple-eavesdropper (MISOME) channel [4]. Alice can choose a beamforming direction that maximizes the difference between the capacity of the main channel and that of the eavesdropper channel (i.e., the achievable secrecy rate). The optimal beamforming vector is the generalized eigenvector of the two CSITs when the power is sufficiently large.

With only partial or no eavesdropper-channel CSIT, secrecy beamforming can still be performed if long decoding latency is allowed. The uncertainty of the Alice-Eve channel is averaged out by channel coding over multiple coherent blocks. Usually, the beamforming direction cannot be determined to maximize the instantaneous channel quality difference in each block, and AN is used to help suppress the reception quality at Eve. The AN is often placed in the null space of the main channel to avoid interfering Bob. In this case, the achievable secrecy rate of the AN-assisted beamforming scales with the transmit power at a rate similar to that of the optimal secrecy beamforming with full CSIT [4]. Interestingly, placing AN in the null space of the main channel may not always be optimal since embedding AN partially in the main channel may cause more harm to Eve than to Bob [6]. For the case where allowed decoding latency is short, the so-called secrecy outage probability is considered, which denotes the probability that the target rate cannot be achieved. The optimal beamformer in terms of secret outage probability is shown to be a linear combination of the maximal ratio-combining and zero-forcing beamformers with weight adjusted according to the available CSIT [7]. Finally, if only the set of possible eavesdropper channel realizations is known (instead of the distribution), a secrecy beamformer maximizing the worst-case secrecy rate can be chosen.

Secrecy with Delayed Main-Channel CSIT

In some scenarios, CSIT may be outdated due to insufficient feedback bandwidth, causing the main-channel CSIT in coherence block t to consist of only the past channel matrices $\mathbf{H}_r(t-1)$, … , $\mathbf{H}_r(1)$. In this case, it is beneficial to utilize messages and AN transmitted in past blocks as common randomness between Alice and Bob to help secure the message in the current block. This is demonstrated in the following example from [8], where it is assumed that the past CSI of Eve, that is, $\mathbf{H}_e(t-1)$, … , $\mathbf{H}_e(1)$, is also available at the transmitter.

Example 2 [8]: Let us consider a wiretap channel with two transmit antennas at Alice and only a single antenna at both Bob and Eve. The 1 × 2 channel vectors of the main and eavesdropper channels in coherence block t are denoted by $\mathbf{h}_r(t)$ and $\mathbf{h}_e(t)$, respectively. To utilize the delayed CSIT, this example shows an achievable scheme where Alice transmits two independent message-carrying symbols over three coherence blocks, as illustrated in Fig. 3. This scheme is optimal at high signal-to-noise ratio (SNR); thus, we assume that the reception is noiseless for ease of exposition. In the first time slot, only AN is transmitted by Alice and is received by Bob as $y_r(1)$. In the second slot, Alice is able to obtain the knowledge of $\mathbf{h}_r(1)$ (and, thus, $y_r(1)$) through delayed feedback and transmits a linear combination of $y_r(1)$ and the two message signals x_1 and x_2 to Bob. Here, $y_r(1)$ is treated as common randomness between Alice and Bob that can help secure the message signals. By removing $y_r(1)$ at Bob, one linear equation of x_1 and x_2 is obtained. In the third time slot, Alice utilizes the knowledge of $\mathbf{h}_e(2)$ to reconstruct Eve's past signal $y_e(2)$ and transmits it to provide Bob with

To ensure confidentiality without secret keys, a coding technique called "secrecy binning" is employed at the heart of many physical layer secrecy techniques. The key idea is to insert additional randomness into the codeword of each confidential message to increase the uncertainty at Eve.

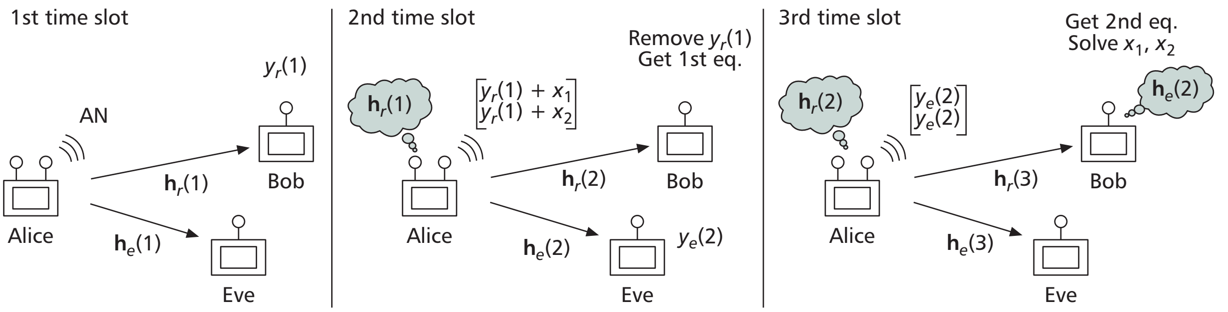


Figure 3. An illustration of the transmission scheme for wiretap channels with delayed CSIT.

the second linear equation needed to solve for x_1 and x_2.

The aforementioned transmission scheme can be modified to incorporate cases without the eavesdropper-channel CSIT [8]. In these cases, Alice will not be able to reconstruct and retransmit Eve's signal in the third time slot. However, the first two time slots can still be employed with only one message transmitted in the second time slot. The delayed CSIT $y_r(1)$ is similarly used as the common randomness to secure the message transmission in slot 2. Even though the above scheme performs well, its optimality is still unknown.

Secrecy with Partial Main-Channel CSIT

In this subsection, we consider the case with instantaneous but partial CSI feedback from Bob and no CSI feedback from Eve. This is most often the case in practical systems, such as LTE-A, where CSI is first quantized, and only the index of the quantized vector is fed back to the transmitter through the PUCCH. In this case, AN-assisted secrecy beamforming can be performed by Alice based on the quantized main-channel CSIT [9, 10]. However, the uncertainty of the CSIT, caused by the quantization noise, will result in the AN leakage problem. That is, the AN placed in the null space of the quantized main channel will leak into the actual main channel causing interference to Bob. To limit the secrecy rate loss due to AN leakage, one can scale logarithmically the number of quantized bits with respect to the transmit SNR.

In the extreme case where no (or infrequent) feedback is provided, Alice may only be able to obtain statistical knowledge of the CSI. In this case, beamforming can still be performed by directing the message-bearing signal toward dimensions that are most likely favorable to Bob. However, when both channels are Gaussian with entries that are independent and identically distributed (i.i.d.), it is optimal to emit signals evenly in all directions. This concept can also be applied to cases with general Nakagami fading channels [11].

Trade-off between Training and Secret Data Transmission

In the previous subsections, we assumed that forward training is perfect, which implies the need for infinite resources for training and channel feedback. However, the cost of utilizing these resources and the trade-off with the amount of resources that can be utilized for a data transmission period is not considered. In fact, when more resources are allocated to training, less resources are left for data transmission, and vice versa.

Let P be the total average power constraint over coherence time $T = T_{TF} + T_D$ so that

$$T_{TF}P_T + T_DP_D \leq PT, \tag{2}$$

where P_T and P_D are the powers utilized for training and data transmission, respectively, and T_{TF} and T_D are the durations of the respective phases. By utilizing the minimum mean square error (MMSE) estimator in the training phase and the AN-assisted secrecy beamforming in the data transmission phase, the power allocation between training and data transmission was examined in [12] at high SNR. With conventional training, and by setting T_{TA} equal to the number of transmit antennas, it was shown that to maximize the achievable secrecy rate, the power ratio P_T/P_D should scale as $\sqrt{T_D}$ as the coherence time increases. This is due to the fact that as the coherence time increases, time and energy resources are sufficient for data transmission, and thus more resources should be devoted to obtaining better channel estimates. Moreover, in data transmission, approximately half the power should be used for the message-bearing signal and half for AN.

Physical Layer Secrecy with Protection against CSI Leakage

Traditional training and channel feedback procedures were often designed without secrecy considerations, and schemes such as forward training and insecure channel feedback may also provide Eve with sufficient CSI to enhance its eavesdropping capability. Therefore, by capitalizing on training in the reverse direction (i.e., from Bob to Alice), several techniques have been proposed in the literature to prevent such (Alice-Eve) CSI leakage.[2] Reverse training can be viewed as an intelligent way to perform channel estimation and feedback simultaneously without benefiting the eavesdropper. By assuming that the Bob-Eve channel is independent of the main (i.e., Alice-Bob) and eavesdropper (i.e., Alice-Eve) channels, as done in [12–14], no

[2] Notice that leakage of the main-channel CSI to Eve may also impact the secrecy performance. In fact, this has also been considered in [3–11] by assuming perfect main-channel CSI at Eve and in [12–15] by assuming that Eve has the same level of main-channel CSI as Bob.

information regarding the latter two channels is revealed to Eve through the reverse training. To adopt reverse training, it is necessary to partition the training phase into two phases: the reverse and the forward training phases, as illustrated in Fig. 4.

Secrecy with Suppressed CSI at Eve

An interesting scheme that utilizes reverse training to suppress the CSI at Eve is the so-called two-way discriminatory channel estimation (DCE) scheme [13]. In this scheme, reverse training is first performed to enable channel estimation directly at Alice. Under the channel reciprocity assumption (e.g., in a time division duplex system), this estimate can be used to directly infer knowledge of the forward channel matrix $\mathbf{H}_r(t)$. Then, in the forward training phase, an AN-assisted training signal, which consists of a pilot signal plus AN in the null space of the estimated main channel, can be transmitted to facilitate channel estimation at Bob while preventing that at Eve. With sufficiently reliable estimation in the reverse training phase, AN can be placed accurately in the desired subspace to avoid interference at Bob. However, under a total power constraint, this reduces the power that can be used for forward training and data transmission. Therefore, a trade-off exists in terms of the resources that should be allocated to reverse training, forward training, and data transmission.

Let P be the average total power constraint over the channel coherence time $T = T_{TF} + T_{TR} + T_D$ so that

$$T_D P_D + (T_{TR} + T_{TF})P_T \leq PT, \quad (3)$$

where P_T and P_D are again the powers utilized for training and data transmission, respectively, and T_{TR}, T_{TF}, and T_D are the durations of the respective phases. By considering again the MMSE estimator in training and the AN-assisted secrecy beamforming scheme in data transmission, the optimal power allocation derived in [12] shows that, compared to the conventional case discussed above, more power should be allocated to training since DCE helps construct a better secrecy channel for data transmission and less power should be utilized for AN in data transmission since the channel quality has already been successfully discriminated through training. This concept can also be extended to non-reciprocal channels as discussed in [13].

Secrecy with Only Main-Channel CSIT and No CSI at Bob and Eve

In this subsection, we consider the case where only reverse training is applied, that is, $T_{TF} = 0$. The key idea is that, since forward training is always performed at the risk of CSI leakage, why not avoid it completely? This can be achieved by employing only reverse training and no forward training, but comes at the price of also not being able to provide CSI to Bob. Conceptually, this can be viewed as a wiretap channel with perfect main-channel CSIT, but no CSI at Bob and Eve [14]. In the wireless scenario, having CSIT allows Alice to pre-compensate for the amplitude and phase variations of the fading channel to enable coherent detection at Bob while leaving Eve confused by the uncertainty of its own channel. It was shown in [14] that under certain conditions, the achievable secrecy rate can actually be significantly higher than that with perfect CSI at all terminals. This implies that under the secrecy scenario, CSI at the transmitter plays a more important role than CSI at the receiver.

Besides the aforementioned schemes, [15] further considered the case where neither training nor feedback is applied, and thus, no instantaneous CSI is available at any node. A constant norm channel input was proposed to exploit the noncoherent nature of the channel and was shown to achieve the optimal performance at high SNR (in terms of the secure degrees of freedom).

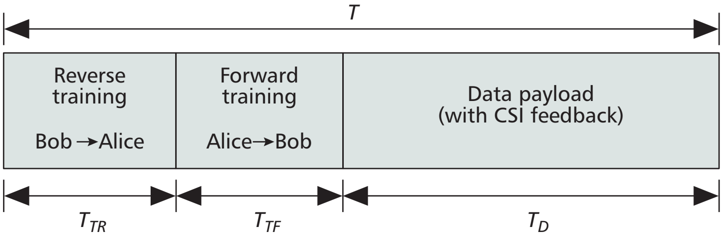


Figure 4. The transmission scheme with both reverse and forward training.

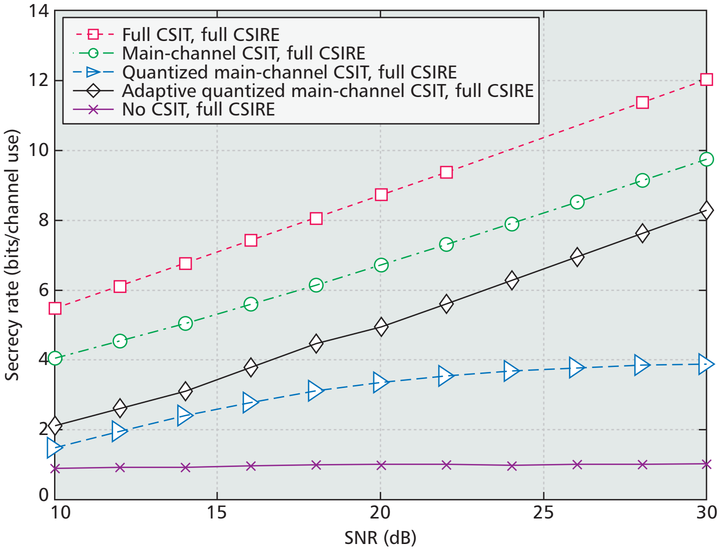


Figure 5. The achievable secrecy rates vs. transmit power for schemes with various channel feedback.

Performance Assessments and Discussions

In this section, we compare the achievable secrecy rates of the aforementioned schemes in wireless fading scenarios. We assume that Alice has four antennas, while both Bob and Eve have only a single antenna each. The 1 × 4 channel vectors $\mathbf{h}_r(t)$ and $\mathbf{h}_e(t)$ have i.i.d. entries with variances 2 and 1, respectively. The variances of the AWGN at both receivers are given by 1.

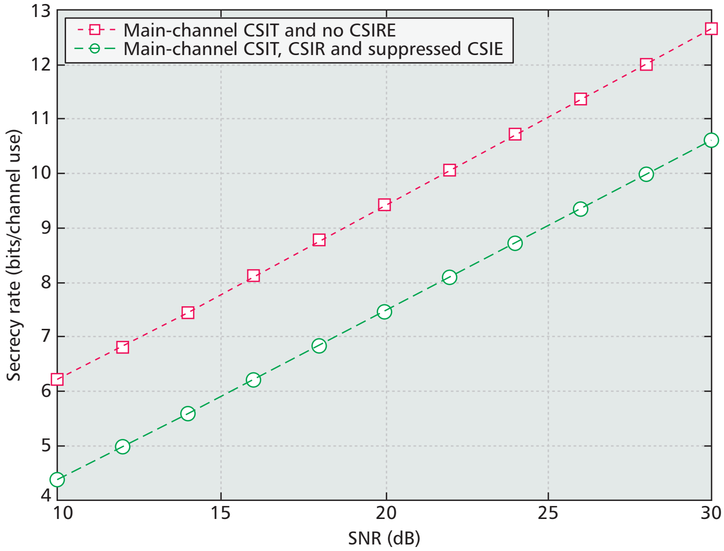


Figure 6. The achievable secrecy rates vs. transmit power for the cases with suppressed CSIE and no CSIRE, respectively.

In Fig. 5, we compare the secrecy rates achievable under four different CSI assumptions introduced earlier. These CSI assumptions result from traditional forward training and channel feedback procedures. In this figure, we can see that the case with full CSIT and full CSI at the receiver and the eavesdropper (CSIRE) achieves the highest secrecy rate. However, the case with only main-channel CSIT and full CSIRE, where AN-assisted secrecy beamforming is adopted, scales at approximately the same rate with respect to the transmit SNR as the full CSIT case. In the case with quantized main-channel CSI, we consider the scenario in [9] where only the channel direction (i.e., the normalized main channel vector) is quantized. The curve is plotted for the case of 10 quantization bits. Due to AN leakage, the achievable secrecy rate eventually saturates as the transmit SNR increases. This can be improved by scaling the number of quantization bits logarithmically with respect to the transmit SNR. When no CSIT is available, the achievable secrecy rate saturates rapidly at low SNR, and thus performs significantly worse than the other schemes, which shows the importance of CSIT.

In Fig. 6, we show the secrecy rates that are achievable with protection against CSI leakage, that is, the case with suppressed CSI at the eavesdropper (CSIE) and the case with only main-channel CSIT, described earlier. For the scheme with suppressed CSI at Eve (i.e., the scheme that utilizes DCE in the training phase), the coherence time is set to be T = 100 (channel uses), and the training length is given as $T_{TF} = N_A = 4$ and $T_{TR} = N_B = 1$, which correspond to the number of antennas at Alice and Bob, respectively. Notice that, as opposed to all other curves, the cost of training and the effect of imperfect channel estimation are both considered in the computation of the secrecy rate. We can see that, even considering the above practical issues, the achievable secrecy rate of the DCE-enabled scheme is still higher than that of the case with main-channel CSIT and full CSIRE (Fig. 5). This shows the advantage of discriminating the quality of the two channels before the secret data is actually transmitted. Moreover, for the case with only main-channel CSIT and no CSIRE, the achievable secrecy rate can actually be higher than that under full CSIT and full CSIRE. This implies that the prevention of CSI leakage is more important than providing Bob with the CSI.

Even though, in the scenarios considered in Figs. 5 and 6, the schemes that avoid CSI leakage are shown to outperform schemes that do not, this is not always the case. In fact, the conventional schemes are still useful due to the following reasons:

- The schemes that avoid CSI leakage do not always lead to better performance. For example, at high SNR, the DCE scheme may not necessarily achieve higher secrecy rates than the conventional AN-assisted secrecy beamforming scheme. This is due to the fact that the reverse training in DCE occupies extra temporal resources without carrying any information, and thus produces a loss in secrecy rate that can be significant at high SNR. Due to similar reasons, the DCE scheme also may not perform well when the coherence time is short.
- Even though the schemes that avoid CSI leakage may be suitable for basic wiretap channels with only a single receiver and eavesdropper, they may not be suitable for other, more general, settings such as the broadcast or interference channels with confidential messages. For example, the reverse training required in several of these schemes may not be efficient in broadcast channels since the time required for all receivers to emit their reverse training signals may occupy too many channel uses. In this case, it may be more efficient to have the transmitter send a single forward training signal to all receivers and utilize the schemes that do not avoid CSI leakage. Moreover, for schemes that utilize AN, such as the DCE scheme, the transmission of AN may cause additional interference to other users, and thus may not be suitable for interference channels.
- The schemes that do not avoid CSI leakage were devised by assuming that perfect CSI of the main and eavesdropper channels are available at Eve. These schemes can be viewed as worst-case schemes that can be used when the level of CSI at Eve is uncertain.

Conclusion

In this article, we discuss how different CSI assumptions resulting from conventional and unconventional ways of doing training and channel feedback may affect the achievable secrecy rate. In the conventional case, training is performed in the forward direction, and channel feedback is provided in an insecure manner, which both benefit Eve in terms of obtaining the CSI and strengthening its eavesdropping capability. In this case, secrecy beamforming can be used to direct the message-bearing signal toward dimensions that are more favorable to Bob, and AN can be used to further suppress the reception quality at Eve. These schemes are devised by assuming perfect CSI at Eve and hence are suitable for the worst-case scenario. On the other hand, if the system setting allows for the prevention of CSI leakage, significant gains may be obtained by adopting new training and channel feedback schemes. In fact, by capitalizing on reverse training, CSI can be provided to Alice without revealing too much (if any) CSI to Eve, causing Eve to be confused by its own channel during reception. Notice that in most schemes, some level of the eavesdropper-channel CSI, such as the statistics of the channel, is required in order to achieve good secrecy performance. However, this is difficult to obtain in practice. Therefore, it is interesting to see how physical layer secrecy can be performed in the absence of any eavesdropper-channel CSI (not even the statistics), which is still an open problem. Moreover, it is also important to construct practical wiretap codes that are suitable for practical systems and determine ways to incorporate physical layer secrecy transmission into current standards.

Even though the schemes that avoid CSI leakage may be suitable for basic wiretap channels with only a single receiver and eavesdropper, it may not be suitable for other, more general, settings such as the broadcast or the interference channels with confidential messages.

References

[1] A. D. Wyner, "The Wire-Tap Channel," *Bell Sys. Tech. J.*, vol. 54, Oct. 1975, pp. 1355–87.

[2] I. Csiszár and J. Körner, "Broadcast Channels with Confidential Messages," *IEEE Trans. Info. Theory*, vol. 24, no. 3, Oct. 1978, pp. 339–48.

[3] S. Shafiee, N. Liu, and S. Ulukus, "Towards the Secrecy Capacity of the Gaussian Mimo Wire-Tap Channel: The 2-2-1 Channel," *IEEE Trans. Info. Theory*, vol. 55, no. 9, Sept. 2009, pp. 4033–39.

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

[5] S. Goel and R. Negi, "Guaranteeing Secrecy Using Artificial Noise," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, June 2008, pp. 2180–89.

[6] P.-H. Lin *et al.*, "On Secrecy Rate ot the Generalized Artificial-Noise Assisted Secure Beamforming for Wiretap Channels," *IEEE JSAC*, vol. 31, no. 9, Sept. 2013, pp. 1728–40.

[7] S. Gerbracht, C. Scheunert, and E. A. Jorswieck, "Secrecy Outage in MISO Systems with Partial Channel Information," *IEEE Trans. Info. Forensics Security*, vol. 7, no. 2, Apr. 2012, pp. 704–16.

[8] S. Yang *et al.*, "Secrecy Degrees of Freedom of MIMO Broadcast Channels with Delayed CSIT," *IEEE Trans. Info. Theory*, vol. 59, no. 9, Sept. 2013, pp. 5244–56.

[9] S. C. Lin *et al.*, "On the Impact of Quantized Channel Feedback in Guaranteeing Secrecy with Artificial Noise: The Noise Leakage Problem," *IEEE Trans. Wireless Commun.*, vol. 10, no. 3, Mar. 2011, pp. 901–15.

[10] Z. Rezki, A. Khisti, and M.-S. Alouini, "Ergodic Secret Message Capacity of the Wiretap Channel with Finite-Rate Feedback," *IEEE Trans. Wireless Commun.*, vol. 13, no. 6, June 2014, pp. 3364–79.

[11] P.-H. Lin and E. A. Jorswieck, "On the Fading Gaussian Wiretap Channel with Statistical Channel State Information at Transmitter," *IEEE Trans. Info. Forensics Security*, vol. 11, no. 1, Jan. 2015, pp. 46–58.

[12] T.-Y. Liu *et al.*, "How Much Training Is Enough for Secrecy Beamforming with Artificial Noise," *Proc. IEEE ICC*, June 2012.

[13] C.-W. Huang *et al.*, "Two-Way Training for Discriminatory Channel Estimation in Wireless MIMO Systems," *IEEE Trans. Signal Process.*, vol. 61, no. 10, May 2013, pp. 2724–38.

[14] P.-C. Lan, Y.-W. P. Hong, and C.-C. J. Kuo, "Enhancing Secrecy in Fading Wiretap Channels with Only Transmitter-Side Channel State Information," IEEE GLOBECOM Wksp. Trusted Commun. with Physical Layer Security, Dec. 2014.

[15] T.-Y. Liu *et al.*, "Secure Degrees of Freedom of MIMO Rayleigh Block Fading Wiretap Channels with No CSI Anywhere," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, May 2015, pp. 2655–69.

Biographies

Ta-Yuan Liu received his B.S. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2009, and is currently pursuing his Ph.D. degree in the Institute of Communications Engineering at National Tsing Hua University. His research interests include physical layer security, information theory, and wireless communications.

Pin-Hsun Lin received his Ph.D. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 2010. Since 2014 he has been a research fellow and packet leader for the DIWINE project with Technische Universität Dresden, Germany, for physical layer security. His research interests include wireless communications and information theory.

Shih-Chun Lin received his Ph.D. degree in electrical engineering from National Taiwan University in 2007. He is currently an assistant professor at National Taiwan University of Science and Technology, Taipei. His research interests include coding/information theory, communications, and signal processing. He has also served as a Technical Program Committee member for the IEEE ICC Workshop on Wireless Physical Layer Security, CNS Workshop on Physical-Layer Method for Wireless Security, and ICC.

Y.-W. Peter Hong received his B.S. from National Taiwan University in 1999 and his Ph.D. from Cornell University in 2005. He joined the Institute of Communications Engineering at National Tsing Hua University in fall 2005 and is now a full professor. His research interests include physical layer secrecy, cooperative communications, and signal processing for sensor networks. He is an Associate Editor for *IEEE Transactions on Signal Processing* and *IEEE Transactions on Information Forensics and Security*.

Eduard Axel Jorswieck (Eduard.Jorswieck@tu-dresden.de) received his Dipl.-Ing. and Dr.-Ing. degree from Technische Universität Berlin in 2000 and 2004. From 2006 until 2008, he was a postdoctoral fellow and later an assistant professor at the Royal Institute of Technology, Sweden. Since 2008, he has been head of the Chair of Communications Theory and a full professor at Technische Universität Dresden, Germany. His research interests include applied information theory, signal processing for communication networks, and communications theory. He serves on the Editorial Boards of *IEEE Transactions on Signal Processing* and *IEEE Transactions on Wireless Communications*.

Hiding Information in Noise: Fundamental Limits of Covert Wireless Communication

Boulat A. Bash, Dennis Goeckel, Don Towsley, and Saikat Guha

Boulat A. Bash and Saikat Guha are with Raytheon BBN Technologies.

Dennis Goeckel and Don Towsley are with the University of Massachusetts, Amherst.

Boulat A. Bash is the corresponding author.

ABSTRACT

Widely deployed encryption-based security prevents unauthorized decoding, but does not ensure undetectability of communication. However, covert, or low probability of detection/intercept communication is crucial in many scenarios ranging from covert military operations and the organization of social unrest, to privacy protection for users of wireless networks. In addition, encrypted data or even just the transmission of a signal can arouse suspicion, and even the most theoretically robust encryption can often be defeated by a determined adversary using non-computational methods such as side-channel analysis. Various covert communication techniques have been developed to address these concerns, including steganography for finite-alphabet noiseless applications and spread-spectrum systems for wireless communications. After reviewing these covert communication systems, this article discusses new results on the fundamental limits of their capabilities, and provides a vision for the future of such systems as well.

INTRODUCTION

Security and privacy are critical in modern-day wireless communication. Widely deployed conventional cryptography presents the adversary with a problem he/she is assumed not to be able to solve because of computational constraints, while information-theoretic secrecy presents the adversary with a signal from which he/she cannot extract information about the message contained therein. However, while these approaches address security in many domains by protecting the content of the message, they do not mitigate the threat to users' privacy from the discovery of the very existence of the message itself.

Indeed, transmission attempts expose connections between the parties involved, and recent disclosures of massive surveillance programs revealed that this "metadata" is widely collected. Furthermore, the transmission of encrypted data can arouse suspicion, and many cryptographic schemes can be defeated by a determined adversary using non-computational means such as side-channel analysis. Anonymous communication tools such as Tor resist metadata collection and traffic analysis by randomly directing encrypted messages through a large network. While these tools conceal the identities of source and destination nodes in a "crowd" of relays, they are designed for the Internet and are not effective in wireless networks, which are typically orders of magnitude smaller. Moreover, such tools offer little protection to users whose communications are already being monitored by adversaries. Thus, secure communication systems should also provide *covert*, stealth, or low probability of detection/intercept (LPD/LPI) communication. Such systems not only protect the information contained in the message from being decoded, but also prevent the adversary from detecting the transmission attempt in the first place and allow communication where it is prohibited.

The overarching goal of covert wireless communication research is the establishment of "shadow networks" like that depicted in Fig. 1. They are assembled from relays that generate, transmit, receive, and consume data, and jammers that generate artificial noise and impair the ability of wardens to detect the presence of communication (we discuss the details of this vision below). However, to create such networks, we must first learn how to connect their component nodes by stealthy communication links. Therefore, in this article we focus on the fundamental limits of such point-to-point links and address the following question: how much information can a sender Alice reliably transmit (if she chooses to transmit) to the intended recipient Bob while hiding it from the adversary, warden Willie?

We begin by briefly reviewing the field of steganography, or the practice of hiding messages in innocuous objects. Steganography is important as it was arguably the first covert communication method devised by man. More recently it has been extensively studied by both the computer science and information theory communities in the context of hiding information in digital media. However, since steganography enables covert communication only at the *application layer*, its analysis has limited use for *physical layer* covert communication techniques such as spread-spectrum. Therefore, following

0163-6804/15/$25.00 © 2015 IEEE

that we examine the fundamental limits of covert communication over analog RF channels, where the information is hidden in the channel artifacts such as additive white Gaussian noise (AWGN), as well as digital communication channels, and briefly touch upon the covert broadcast scenario at the end of the section. We conclude with a discussion of shadow networks and ongoing research in jammer-assisted covert communication.

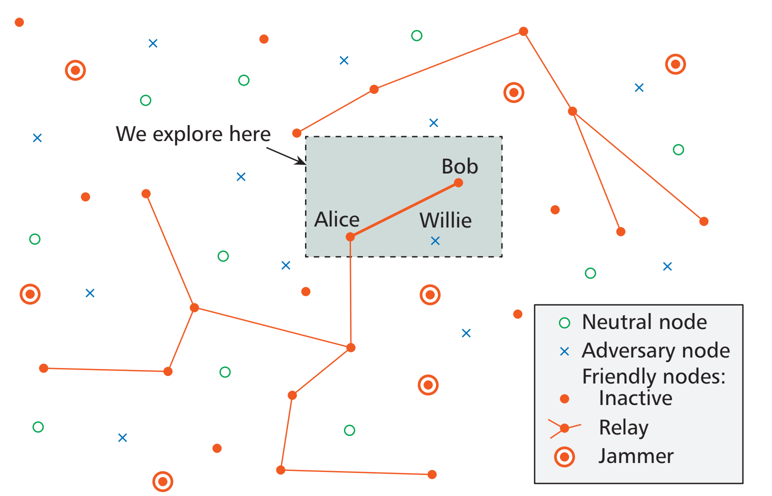


Figure 1. Our vision of a "shadow network." Most of this article focuses on the scenario involving the indicated three nodes: transmitter Alice, receiver Bob, and warden Willie.

Steganography

Covert communication is an ancient discipline. A description of it is given by Herodotus circa 440 BCE in *The Histories*, an account of the Greco-Persian Wars: in Chapter 5, paragraph 35, Histiaeus shaves the head of his slave, tattoos the message on his scalp, waits until the hair grows back, and then sends the slave to Aristagoras with instructions to shave the head and read the message that calls for an anti-Persian revolt in Ionia; in Chapter 7, paragraph 239, Demaratus warns Sparta of an imminent Persian invasion by scraping the wax off a wax tablet, scribbling a message on the exposed wood, and concealing the message by covering the tablet with wax. This practice of hiding sensitive messages in innocuous objects is known as *steganography*.

Modern digital steganography conceals messages in finite-length, finite-alphabet *covertext* objects, such as images or software binary code. Embedding hidden messages in covertext produces *stegotext*, necessarily changing the properties of the covertext. The countermeasure for steganography, *steganalysis* (an analog of cryptanalysis for cryptography), looks for these changes. Covertext is usually unavailable for steganalysis (when it is, steganalysis consists of the trivial comparison between the covertext and the suspected stegotext). However, Willie is assumed to have a complete statistical model of the covertext. The amount of information that can be embedded without being discovered depends on whether Alice also has access to this model. If she does, *positive-rate steganography* is achievable: given an $\mathcal{O}(n)$-bit[1] secret "key" that is shared with Bob prior to the embedding, $\mathcal{O}(n)$ bits can be embedded in an n-symbol covertext without being detected by Willie [1, Ch. 13.1].

Recent work focuses on the more general scenario where the complete statistical model of the covertext is unavailable to Alice. Then Alice can safely embed $\mathcal{O}(\sqrt{n}\log n)$ bits by modifying $\mathcal{O}(\sqrt{n})$ symbols out of n in the covertext, at the cost of pre-sharing $\mathcal{O}(\sqrt{n}\log n)$ secret bits with Bob. Note that this square root law of digital steganography yields zero-rate steganography since $\lim_{n\to\infty}\mathcal{O}(\sqrt{n}\log n)/n = 0$ b/symbol.

The proof is available in Chapter 13.2.1 of the review of pre-2009 work in digital steganography [1]. More recent work shows that an *empirical* model of covertext suffices to break the square root law and achieve positive-rate steganography [2]. Essentially, while embedding at a positive rate lets Willie obtain $\mathcal{O}(n)$ stegotext observations (enabling detection of Alice when statistics of covertext and stegotext differ), the increasing size n of the covertext allows Alice to improve her covertext model and produce statistically matching stegotext.

However, steganography is inherently an application layer covert communication technique. As such, the results for steganography have limited use in physical layer covert communication. First, analysis of the steganographic systems generally assumes that stegotext is not corrupted by a noisy channel. Second, the generalization of the results for steganographic systems is limited because of their finite-alphabet discrete nature. Third, by embedding the hidden messages, Alice *replaces* part of the covertext. While this effectively enables the recent positive rate steganography methods [2], it cannot be done in standard communication systems unless Alice controls Willie's noise source. Finally, the most serious drawback of using steganography for covert communication is the necessity of transmitting the stegotext from Alice to Bob — a potentially unrealizable requirement when all communication is prohibited. We thus consider physical layer covert communication that employs channel artifacts such as noise to hide transmissions.

Physical Layer Covert Communication

We begin the investigation of physical layer covert communication by considering RF wireless communication. Since its emergence in the early 20th century, protecting wireless RF communication from detection, jamming, and eavesdropping has been of paramount concern. *Spread spectrum* techniques, devised between the two World Wars to address this issue, constituted the earliest and, arguably, the most enduring form of physical layer security.

Spread Spectrum Communication

Essentially, the spread spectrum approach involves transmitting a signal that requires a bandwidth W_M on a much wider bandwidth $W_s \gg W_M$,

[1] We use the *Big-O* notation in this article, where $\mathcal{O}(f(n))$ denotes an asymptotic upper bound.

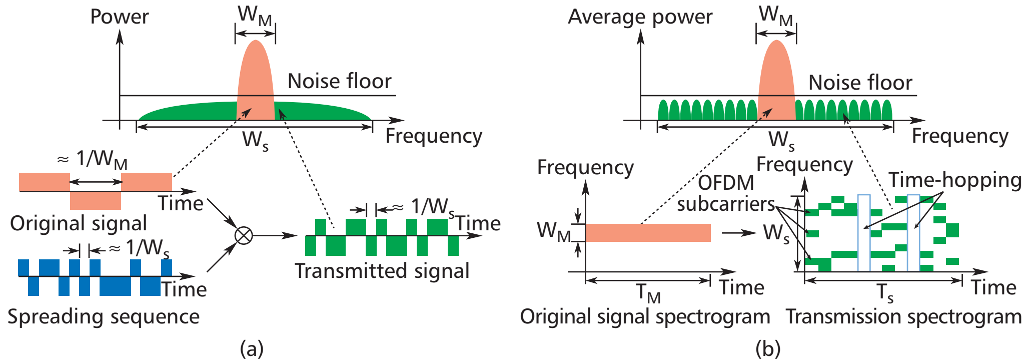


Figure 2. Spread spectrum techniques. a) DSSS; b) FHSS with OFDM and time-hopping.

thereby suppressing the power spectral density of the transmission below the noise floor. Spread spectrum systems provide both covert communication capability as well as resistance to jamming, fading, and other forms of interference. A comprehensive review of this field is available in [3]. Typical spread spectrum techniques include *direct sequence* spread spectrum (DSSS), *frequency-hopping* spread spectrum (FHSS), and their combination.

When Alice uses DSSS, she multiplies the signal waveform by the *spreading sequence* — a randomly generated binary waveform with a substantially higher bandwidth than the original signal. The resulting waveform is thus “spread” over a wider bandwidth, which reduces the power spectral density of the transmitted signal. Bob uses the same spreading sequence to despread the received waveform and obtain the original signal. The spreading sequence is exchanged by Alice and Bob prior to transmission and is kept secret from Willie.[2] Outside of security applications, the use of public uncorrelated spreading sequences between transmitter/receiver pairs enables multiple access; DSSS thus forms the basis of code-division multiple access (CDMA) protocols used in cellular telephony. The operation of DSSS is illustrated in Fig. 2a.

When Alice uses FHSS, she re-tunes the carrier frequency for each transmitted symbol. However, like the spreading sequence in DSSS, the frequency-hopping pattern is also randomly generated and secretly shared between her and Bob prior to the transmission. FHSS can be combined with orthogonal frequency-division multiplexing (OFDM), enabling the use of multiple carrier frequencies. To further reduce the average transmitted symbol power, FHSS can be used with *time-hopping* techniques that randomly vary the duty cycle (the time-hopping pattern is also secretly pre-shared between Alice and Bob prior to the transmission). The operation of FHSS with OFDM and time-hopping is illustrated in Fig. 2b.

Although spread spectrum architectures are well developed, the analytical evaluation of covert communication has been sparse. A. Hero studied secrecy as well as undetectability [4] in a multiple-input multiple-output (MIMO) setting, focusing on the signal processing aspects. He recognized that covert communication systems are constrained by average power, and noted the need to explore the fundamental information-theoretic limits in the conclusion of his work. In fact, knowledge of the limits of any communication system is important, particularly since modern coding techniques (e.g., turbo codes and low-density parity check codes) allow third/fourth generation (3G/4G) cellular systems to operate near their theoretical *channel capacity*, the maximum rate of reliable communication that is unconstrained by the security requirements. However, while the secrecy portion of [4] has drawn significant attention, the covert communication portion was largely overlooked until our work on the square root limit of covert communication, discussed next. We note that the fundamental results that follow apply not only to classical spread spectrum systems, but also to modern covert communication proposals that rely on channel noise and equipment imperfections to hide communications (as is done, e.g., in [5]).

Square Root Law for Covert Communication over AWGN Channels

Spread spectrum systems allow communication where it is prohibited because spreading the signal power over a large time-frequency space substantially reduces Willie’s signal-to-noise ratio (SNR). This impairs his ability to discriminate between the noise and the information-carrying signal corrupted by noise. Here we determine just how small the power has to be for the communication to be fundamentally undetectable, and how much covert information can be transmitted reliably.

Consider an additive white Gaussian noise (AWGN) channel model where the signaling sequence is corrupted by the addition of a sequence of independent and identically distributed zero-mean Gaussian random variables with variance σ^2. This is the standard model for a free-space RF channel. Suppose that the channels from Alice to Bob and to Willie are subject to AWGN with respective variances $\sigma_b^2 > 0$ and $\sigma_w^2 > 0$,[3] as illustrated in Fig. 3a. Let *channel use* denote the unit of communication resource — a fixed time period that is used to transmit a fixed-bandwidth signal — and let n be the total number of channel uses available to Alice and Bob (e.g., $n = W_sT_s$ in Fig. 2b). Willie’s ability to detect Alice’s transmission depends on the amount of total power that she uses. Let us intu-

[2] While an exchange of a secret prior to covert communication is similar to a key exchange in symmetric-key cryptography (e.g., one-time pad), an important distinction is that public-key cryptography techniques cannot be used to exchange this secret on a channel monitored by Willie without revealing the intention to communicate.

[3] If the channel from Alice to Bob is noiseless ($\sigma_b^2 = 0$) and the channel from Alice to Willie is noisy ($\sigma_w^2 > 0$), Alice can transmit an infinite amount of information to Bob; if the channel from Alice to Willie is noiseless ($\sigma_w^2 = 0$), covert communication is impossible.

itively derive[4] Alice's power constraint assuming that Willie observes these n channel uses. When Alice is not transmitting, Willie observes AWGN with total power $\sigma_w^2 n$ over n channel observations on average. By standard statistical arguments, with high probability, observations of the total power lie within $\pm c\sigma_w^2\sqrt{n}$ of this average, where c is a constant. Since Willie observes Alice's signal power when she transmits in addition to the noise power, to prevent Willie from getting suspicious, the total power that Alice can emit over n channel uses is limited to $\mathcal{O}(\sigma_w^2\sqrt{n})$; otherwise, her transmission will be detected (in fact, a standard radiometer suffices for Willie to detect her if she emits more power, provided σ_w^2 is known[5]). This allows her to reliably transmit $\mathcal{O}(\sigma_w^2\sqrt{n}/\sigma_b^2)$ covert bits to Bob in n channel uses, but no more than that [6]. Note that just like the above steganographic square root law, this yields a zero-rate channel (as $\lim_{n\to\infty}\mathcal{O}(\sqrt{n})/n = 0$ b/symbol). The similarity of this *square root law for covert communications* to the steganographic square root law is attributable to the mathematics of statistical hypothesis testing. The additional log n factor in the steganographic square root law comes from the fact that the steganographic "channel" to Bob is noiseless.

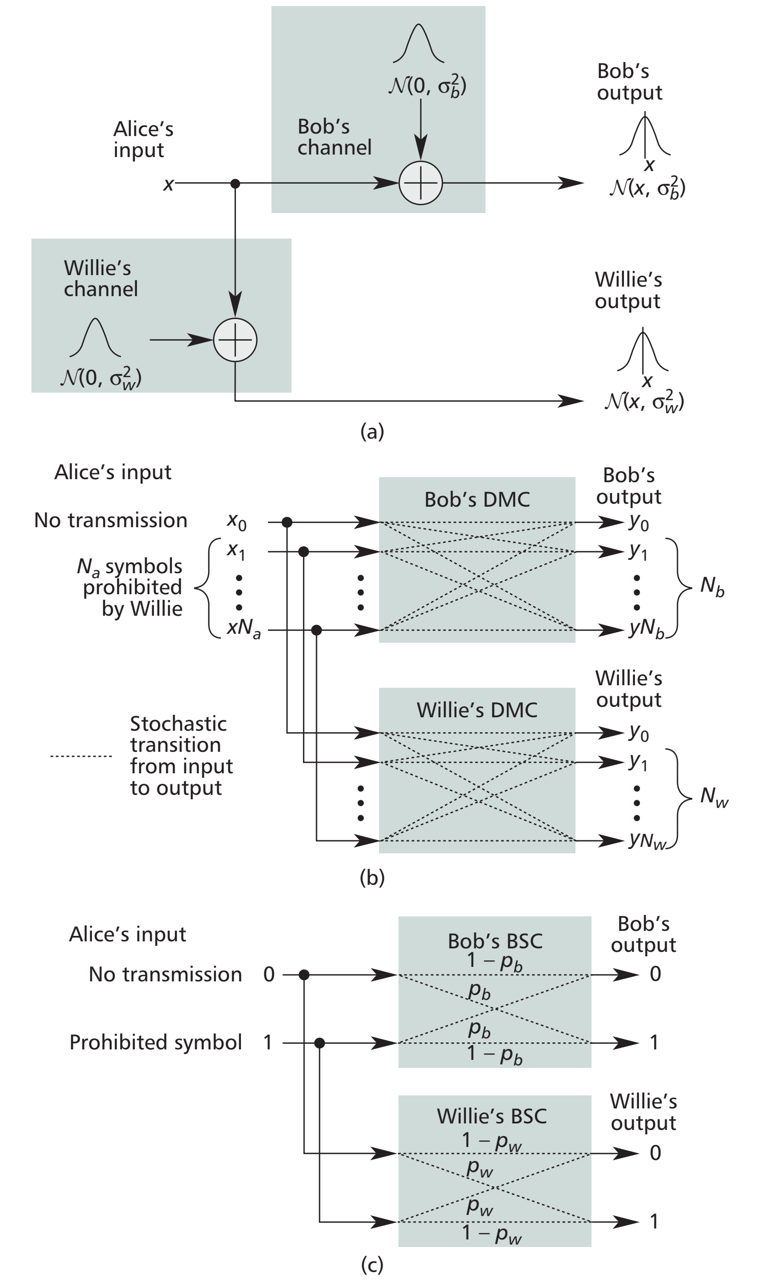


Figure 3. a) AWGN channel; b) DMC; c) BSC.

As in steganography and spread spectrum communication, prior to communicating, Alice and Bob may share a secret. For example, a scheme described in [6] and depicted in Fig. 4 allows Alice and Bob to reliably transmit $\mathcal{O}(\sigma_w^2\sqrt{n}/\sigma_b^2)$ covert bits using binary amplitude modulation, any error correction code (which can be known to Willie), and $\mathcal{O}(\sqrt{n}\log n)$ pre-shared secret bits. The secret contains a random subset $\mathcal{S}$ of n available channel uses (effectively a frequency/time-hopping pattern), and a random one-time pad of size $|\mathcal{S}|$. $\mathcal{S}$ is generated by flipping a biased random coin n times with probability of heads $\mathcal{O}(1/\sqrt{n})$: the ith channel use is selected for transmission if the ith flip is heads; on average, $|\mathcal{S}| = \mathcal{O}(\sqrt{n})$. Knowledge of $\mathcal{S}$ allows Bob to discard the observations that are not in $\mathcal{S}$ and decode Alice's message; Willie observes mostly noise since he does not have $\mathcal{S}$. Rather than protecting the message content, the one-time pad prevents Willie's exploitation of the error correction code's structure to detect Alice.

While the size of the key is asymptotically larger than the size of the transmitted message, there are many real-world scenarios where this is an acceptable trade-off to being detected. Furthermore, the recent extension of [6] to digital covert communication described next demonstrates that the pre-shared secret can be eliminated in some scenarios.

Digital Covert Communication

The *discrete memoryless channel* (DMC) model describing digital communication often sheds light on what is feasible in practical communication systems. The DMC model assumes discrete input and output, which allows the DMC to be represented using a bipartite graph where the two sets of vertices correspond to input and output alphabets, and edges correspond to the stochastic transitions from input to output symbols. The memoryless nature of the DMC means that its output is statistically independent from any symbol other than the input at that time. We illustrate this model in Fig. 3b, which we augment by designating one of Alice's inputs as "no transmission" — a necessary default channel input permitted by Willie.[6]

We first consider the *binary symmetric channel* (BSC) illustrated in Fig. 3c, which restricts the DMC to binary input and output alphabet {0, 1}, and the probability of a crossover from zero at the input to one at the output being equal to that of a crossover from one to zero. Denote by $p_b > 0$ and $p_w > 0$ the crossover probabilities on Bob's and Willie's BSCs, respectively. It has been shown that, while no more than $\mathcal{O}(\sqrt{n})$ covert bits can be reliably transmitted in n BSC uses, if $p_w > p_b$, the pre-shared secret is unnecessary [7].

Channel *resolvability* can be employed to

[4] The formal proof is in [6, Sec. III].

[5] See [6, Sec. IV] for the proof.

[6] For example, this could be the zero-signal in the AWGN channel scenario.

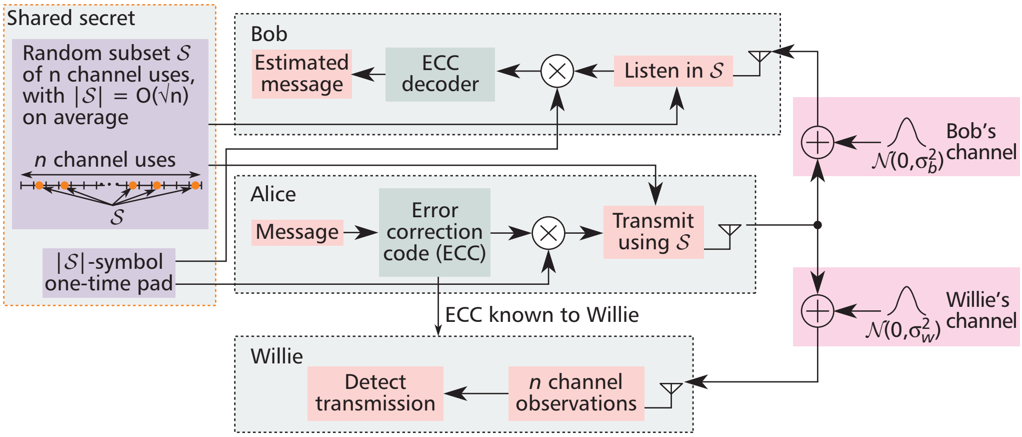

Shared secret
Random subset $\mathcal{S}$ of n channel uses, with $|\mathcal{S}| = O(\sqrt{n})$ on average
n channel uses
$\mathcal{S}$
$|\mathcal{S}|$-symbol one-time pad
Bob
Estimated message
ECC decoder
Listen in $\mathcal{S}$
Bob's channel
$\mathcal{N}(0,\sigma_b^2)$
Alice
Message
Error correction code (ECC)
Transmit using $\mathcal{S}$
Willie's channel
$\mathcal{N}(0,\sigma_w^2)$
ECC known to Willie
Willie
Detect transmission
n channel observations

Figure 4. Design of a covert communication system that allows Alice and Bob to use any error-correction codes (including those known to Willie) to reliably transmit $\mathcal{O}(\sqrt{n})$ covert bits using $\mathcal{O}(\sqrt{n} \log n)$ pre-shared secret bits.

generalize the square root law in [7] to DMCs. Channel resolvability is the minimum input entropy[7] needed to generate a channel output distribution that is "close" (by some measure of closeness between probability distributions[8]) to the channel output distribution for a given input; resolvability has been used to obtain new, stronger results for the information-theoretic secrecy capacity [8]. If the channels from Alice to both Willie and Bob are DMCs, and Willie's channel is worse than Bob's, techniques in [7, 9] can be used to demonstrate the square root law without a pre-shared secret [10]. Furthermore, as long as the Alice-to-Willie channel is known to Alice, $\mathcal{O}(\sqrt{n})$ pre-shared secret bits are sufficient for covert communication when Willie's channel capacity is greater than or equal to Bob's [10]. The results in [10] apply to AWGN channels as well: a covert communication scheme exists[9] that uses $\mathcal{O}(\sqrt{n})$ pre-shared secret bits, and, if the noise power at Willie's receiver is greater than that at Bob's receiver, secret-less covert communication is achievable. Finally, channel resolvability techniques enable the analysis of the constant hidden by the Big-O notation, completely characterizing the square root law for DMCs and AWGN channels [10].

Willie's Ignorance of Transmission Time Helps Alice

When deriving the square root laws, we assume that Willie knows *when* the transmission takes place, if it does. However, in many practical scenarios Alice and Bob have a pre-arranged time for communication that is unknown to Willie (e.g., a certain time and day). The transmission might also be short relative to the total time during which it may take place (e.g., a few seconds out of the day). If Willie does not know when the message may be transmitted, he has to monitor a much longer time period than the time required for the transmission. It turns out that Willie's ignorance of Alice's transmission time allows her to transmit additional information to Bob. Surprisingly, under some mild conditions on the relationship between the total available transmission time and the transmission duration, Alice and Bob do not even have to pre-arrange the communication time. The technical details of this work are provided in [11].

Positive-Rate Covert Communication

The covert communication channels described above are zero-rate, since the average number of bits that can be covertly transmitted per channel use tends to zero as the number of channel uses n gets large. Here we discuss the possibility of positive-rate covert communication, that is, reliable transmission of $\mathcal{O}(n)$ covert bits in n channel uses. In general, the circumstances that allow Alice to covertly communicate with Bob at positive rates occur when Willie either *allows* Alice to transmit messages containing information (rather than zero-signal) or is ignorant of the probabilistic structure of the noise on his channel (note that the applicability of the steganographic results [2] here is limited since estimation of the probabilistic structure of the noise on Willie's channel is insufficient unless Alice can "replace" this noise rather than add to it). When Willie allows transmissions, the covert capacity is the same as the information-theoretic secrecy capacity (see [9] for treatment of the DMCs). Incompleteness of Willie's noise model can also allow positive-rate covert communication: in the noisy digital channel setting, Willie's ignorance of the channel model is a special case of the scenario in [9], while in the AWGN channel setting, random noise power fluctuations have been shown to yield positive-rate covert communication [12]. The latter result holds even when the noise power can be bounded; a positive rate is achieved because Willie does not have a constant baseline of noise for comparison.

Covert Broadcast

Some of the results for point-to-point covert communication in the presence of a single warden that are discussed in this section can easily be extended to scenarios with multiple independently controlled receivers. For example, covert communication over an AWGN channel effectively imposes a power constraint on Alice.

[7] Essentially, entropy measures "surprise" associated with a random variable, or its "uncertainty." For example, a binary random variable describing a flip of a fair coin with equal probabilities of heads and tails has higher entropy than the binary random variable describing a flip of a biased coin with probability of heads larger than tails. The output of the biased coin is more predictable, and less surprising, as one should observe more heads. Introductory texts on information theory provide in-depth discussion of entropy and other information-theoretic concepts.

[8] Examples of measures of closeness are variational distance and relative entropy.

[9] Conceptually, the covert communication scheme that uses $\mathcal{O}(\sqrt{n})$ secret bits resembles the method that uses $\mathcal{O}(\sqrt{n} \log n)$ secret bits as described in Fig. 4; however, its mathematical analysis is highly technical and is outside the scope of this article.

Since a pre-shared secret enables covert communication in this setting, if each receiver obtains it prior to communication, Alice can use standard techniques from network information theory to encode covert messages to multiple recipients. The extension to a multi-warden setting as well as other networked scenarios is the ongoing work discussed next.

Conclusion: Toward Shadow Networks

Our ultimate objective is to enable a wireless "shadow network," illustrated in Fig. 1, composed of transmitters, receivers, and friendly jammers that generate artificial noise, impairing wardens' ability to detect transmissions. While the relays are valuable and require protection, the jammers can be cheap, numerous, and disposable (i.e., the adversary can silence a particular jammer easily, but, because of their great numbers, silencing enough of them to produce a significant impact is infeasible). Thus, jammers have been shown to facilitate information-theoretical secrecy by confusing the eavesdropper even while being completely ignorant of the messages exchanged by legitimate communicating parties [13].

In covert networks jammer activities are independent of the relay transmission states; that is, wardens cannot detect transmissions by listening to the jammers. Thus, jammers have a parasitic effect on the wardens' SNRs and are a nuisance. It is important to characterize the scaling behavior of such a network, akin to the recent results for the secure (but not covert) multipath unicast communication in large wireless networks [14]. The first step toward this goal is extending the covert communication scenario of this article to point-to-point jammer-assisted covert communication in the presence of multiple wardens. Preliminary results [15] assume that jammers operate at a constant power, and the signal propagation model accounts only for path loss and AWGN. However, as [12] demonstrates, uncertainty in noise experienced by the warden is beneficial to Alice. Thus, variable jamming power and multipath fading should be incorporated into the jammer-assisted covert communication model, as it may enable covert communication at a positive rate. Completing the characterization of the point-to-point covert link in a multi-warden multi-jammer environment is an important step toward understanding the behavior of "shadow networks" and their eventual implementation.

Completing the characterization of the point-to-point covert link in a multi-warden multi-jammer environment is an important step toward understanding the behavior of "shadow networks," and their eventual implementation.

Acknowledgment

We are grateful to Matthieu Bloch, Gerhard Kramer, and Sid Jaggi for discussions and comments on a draft of this article. This research was sponsored by the National Science Foundation under grants CNS-1018464 and ECCS-1309573.

References

[1] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*, 1st ed., Cambridge Univ. Press, 2009.

[2] S. Craver and J. Yu, "Subset Selection Circumvents the Square Root Law," *Proc. SPIE 7541 Media Forensics and Security II*, 2010, pp. 754,103–06.

[3] M. K. Simon *et al.*, *Spread Spectrum Communications Handbook*, McGraw-Hill, 1994.

[4] A. O. Hero, "Secure Space-Time Communication," *IEEE Trans. Info. Theory*, vol. 49, no. 12, Dec. 2003, pp. 3235–49.

[5] A. Dutta *et al.*, "Secret Agent Radio: Covert Communication through Dirty Constellations," *Proc. 14th Int'l. Conf. Info. Hiding, ser. IH'12*, Berkeley, CA, 2013, pp. 160–75.

[6] B. A. Bash, D. Goeckel, and D. Towsley, "Limits of Reliable Communication with Low Probability of Detection on AWGN Channels," *IEEE JSAC*, vol. 31, no. 9, 2013, pp. 1921–30; originally presented at ISIT 2012, Cambridge MA.

[7] P. H. Che, M. Bakshi, and S. Jaggi, "Reliable Deniable Communication: Hiding Messages in Noise," *Proc. IEEE Int'l. Symp. Info. Theory*, Instanbul, Turkey, July 2013, arXiv:1304.6693, 2013.

[8] M. Bloch and J. Laneman, "Strong Secrecy from Channel Resolvability," *IEEE Trans. Info. Theory*, vol. 59, no. 12, Dec 2013, pp. 8077–98.

[9] J. Hou and G. Kramer, "Effective Secrecy: Reliability, Confusion and Stealth," *Proc. IEEE Int'l. Symp. Info. Theory*, Honolulu, HI, July 2014, arXiv:1311.1411.

[10] M. Bloch, "Covert Communication over Noisy Memoryless Channels: A Resolvability Perspective," *Proc. IEEE Int'l. Symp. Info. Theory*, Hong Kong, China, June 2015, arXiv:1503.08778.

[11] B. A. Bash, D. Goeckel, and D. Towsley, "LPD Communication when the Warden Does Not Know When," *Proc. IEEE Int'l. Symp. Info. Theory*, Honolulu, HI, July 2014, arXiv:1403.1013.

[12] S. Lee and R. Baxley, "Achieving Positive Rate with Undetectable Communication over AWGN and Rayleigh Channels," *Proc. IEEE ICC*, June 2014, pp. 780–85.

[13] L. Lai and H. El Gamal, "The Relay-Eavesdropper Channel: Cooperation for Secrecy," *IEEE Trans. Info. Theory*, vol. 54, no. 99, Sept. 2008, pp. 4005–19.

[14] C. Capar *et al.*, "Secret Communication in Large Wireless Networks without Eavesdropper Location Information," *Proc. IEEE INFOCOM 2012*, Mar. 2012, pp. 1152–60.

[15] R. Soltani *et al.*, "Covert Single-Hop Communication in a Wireless Network with Distributed Artificial Noise Generation," *Proc. Conf. Commun., Control, Comp.*, Monticello, IL, 2014.

Biographies

Boulat A. Bash (boulat@cs.umass.edu) holds a B.A. (2001) in economics from Dartmouth College, and an M.S. (2008) and a Ph.D. (2015) in computer science from the University of Massachusetts, Amherst. He is a scientist in the Quantum Information Processing (QuIP) group at Raytheon BBN Technologies. His research interests include security, privacy, communications, signal processing, and information theory.

Dennis Goeckel [F] holds a B.S. (1992) from Purdue University, and an M.S. (1993) and a Ph.D. (1996) from the University of Michigan. He is a professor in the Electrical and Computer Engineering Department at the University of Massachusetts, Amherst. His research interests include physical layer communications and wireless network theory. He received the NSF CAREER Award (1999). He was a Lilly Teaching Fellow (2000–2001), and received the University of Massachusetts Distinguished Teaching Award (2007).

Saikat Guha holds a B.Tech. in electrical engineering (2002) from the Indian Institute of Technology, Kanpur, and an S.M. (2004) and a PhD (2008) in electrical engineering and computer science from the Massachusetts Institute of Technology. He is a senior scientist in the Quantum Information Processing (QuIP) group at Raytheon BBN Technologies. His research interests span quantum-optical communication and sensing, optical quantum computing, and network information theory. He received a NASA Tech Brief Award in 2010.

Don Towsley [F] holds a B.A. (1971) in physics and a Ph.D. (1975) in computer science from the University of Texas. He is a Distinguished Professor at the University of Massachusetts in the College of Information and Computer Sciences. His research interests include networks and network science. He has received several achievement awards including the 2007 IEEE Koji Kobayashi Award and numerous paper awards. He is a Fellow of the ACM.

Physical Layer Security in Wireless Cooperative Relay Networks: State of the Art and Beyond

Leonardo Jiménez Rodríguez, Nghi H. Tran, Trung Q. Duong, Tho Le-Ngoc, Maged Elkashlan, and Sachin Shetty

Leonardo Jiménez Rodríguez and T. Le-Ngoc are with McGill University.

Nghi H. Tran is with the University of Akron.

Trung Q. Duong is with Queen's University Belfast.

Maged Elkashlan is with Queen Mary University of London.

Sachin Shetty is with Tennessee State University.

ABSTRACT

Cooperative relaying is an effective method of increasing the range and reliability of wireless networks, and several relaying strategies have been adopted in major wireless standards. Recently, cooperative relaying has also been considered in the context of PHY security, which is a new security paradigm to supplement traditional cryptographic schemes that usually handle security at the upper layers. In wireless PHY security, relay nodes can be used to exploit the physical layer properties of wireless channels in order to support a secured transmission from a source to a destination in the presence of one or more eavesdroppers. While some breakthroughs have been made in this emerging research area, to date, the problem of how to effectively adopt advanced relaying protocols to enhance PHY security is still far from being fully understood. In this article, we present a comprehensive summary of current state-of-the-art PHY security concepts in wireless relay networks. A case study is then provided to quantify the benefits of power allocation and relay location for enhanced security. We finally outline important future research directions in relaying topologies, full-duplex relaying, and cross-layer design that can ignite new interests and ideas on the topic.

INTRODUCTION

Wireless communications has grown explosively and plays an important role in the daily life of human beings. Over the years, significant efforts have been made to address the primary challenge in the design of wireless communication systems: how to increase the data transmission rate over a bandwidth-limited wireless radio channel with high reliability and, at the same time, with as low power consumption as possible. Among various solutions, cooperative relaying has been considered as an effective method to increase the range and reliability in wireless networks. While research on cooperative relaying is still an active area, several relaying strategies have been adopted in major wireless standards because of the tremendous benefits that relaying offers.

Due to the broadcast nature of wireless channels, security and privacy are of utmost concern for future wireless technologies. However, securely transferring confidential information over a wireless network in the presence of adversaries still remains a challenging task. Although security was originally viewed as a high-layer problem to be solved using cryptographic methods, physical layer (PHY) security based on information theory has been gaining increasing research attention, especially for wireless networks [1]. In wireless PHY security, the breakthrough idea is to exploit the characteristics of wireless channels, such as fading or noise, to transmit a message from a source to an intended destination while trying to keep this message confidential from eavesdroppers. Different from cryptographic methods, no computational constraints are placed on the eavesdroppers. The theoretical foundations of PHY security were laid by Wyner [2], who introduced the wiretap channel shown in Fig. 1a. In this channel, a transmitter wants to send a confidential message to a receiver in the presence of an eavesdropper. Wyner characterized the trade-off between achievable rate at the destination and the level of ignorance at the eavesdropper. In particular, he showed that a non-zero rate can be achieved in perfect secrecy. Such a rate is defined as the secrecy rate, and the maximum secrecy rate is called the secrecy capacity. For instance, for a degraded channel, the secrecy capacity is given by

$$C_s = \max I(x, y) - I(x, z), \tag{1}$$

where $I(x, y)$ is the mutual information between the transmitted signal x and the signal received at the legitimate receiver y, $I(x, z)$ is the mutual information between the transmitted signal and the signal overheard at the eavesdropper z, and the maximization is carried over the distribution of x.

Benefiting from information-theoretic studies in cooperative communications, relaying

0163-6804/15/$25.00 © 2015 IEEE

strategies have also recently received considerable attention in the context of PHY security over wireless networks [3]. As shown in Fig. 1b, in wireless PHY security, relay nodes can be deployed to support a secured transmission from a source to a destination in the presence of one or more eavesdroppers. For instance, similar to cooperative communications, relay nodes can be used as trusted nodes to retransmit an amplified version of the signal received from the source with a suitable power amplification coefficient, that is, amplify-and-forward (AF). The trusted relay can also transmit a weighted version of the decoded signal, that is, decode-and-forward (DF), or forward a compressed copy of the received signal, that is, compress-and-forward (CF). Another method is to generate a weighted jamming signal from the relay to confound the adversary. This technique is usually referred to as cooperative jamming. Different combinations of these techniques are also possible, as discussed later. In any case, the key issue is how to exploit channel characteristics, such as channel state information (CSI), to optimize the weighted signals so that the secrecy performance can be enhanced.

While the use of relay nodes to transmit confidential information between the source and destination has gained considerable effort, attention has also been paid to untrusted relay networks in the context of PHY security. In this kind of network, a relay might attempt to try to decode the source's confidential signal. In this case, a very important question can be raised: Can cooperation with an untrusted relay be beneficial? Interestingly, the answer is positive [4]. Specifically, we can achieve a higher secrecy rate by treating the untrusted relay as both a helper to relay the information and an eavesdropper to overhear the information, rather than just considering the untrusted relay as an eavesdropper.

There is no doubt that benefits offered by cooperative relaying to enhance the security at the physical layer of wireless networks are significant. However, while quite a few advancements have been made recently in this emerging area, there are still a number of issues and challenges that need to be addressed for a novel PHY security treatment in a wireless relay network. For example, all current applications of relaying to secure communications assume that the source and relay transmit information over orthogonal channels. By allowing the source and relay(s) to transmit simultaneously in a non-orthogonal manner, the degrees of broadcasting and receiving collision are maximized, and security performance can be further improved. Unfortunately, adopting such advanced relaying protocols to enhance PHY security is not a straightforward task due to the fact that for quite a few non-orthogonal relaying protocols, the corresponding maximum achievable rates are still not known.

It is clear that research on PHY security for wireless relay networks is only at its early stage and the opportunity for innovation and research remains tremendous. Therefore, the aim of this article is two-fold:

- To present a comprehensive summary on current state of the art in this emerging research area
- To provide a high-level scope for future research directions.

In the remainder of the article, we first highlight the development of PHY security issues in untrusted relay networks. Then important issues and current state-of-the-art solutions in trusted relay networks are discussed. Following that we present a case study of AF relaying and jamming to illustrate in further detail the importance of power allocation and relay location for secrecy enhancement. Finally, we provide concluding remarks and outline important future research directions.

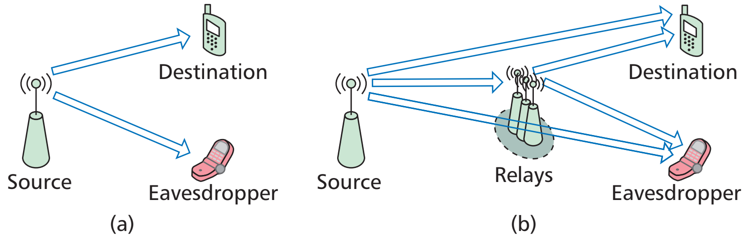


Figure 1. a) Wyner's wire-tap channel; b) wire-tap channel with cooperative relaying for enhanced security.

Untrusted Relays

Untrusted relaying is motivated by several cooperative networks where the source S and destination D seek help from one or multiple relay nodes R to relay the information, but at the same time, the source-destination pair wishes to keep the information confidential from these nodes. Examples of such a network include networks belonging to a government or a financial institution where not every node has the same level of security clearance. In a similar manner, in ad hoc networks, relay nodes are needed for connectivity, but they are not authenticated. In these networks, while the relay is willing to carry out the designated relaying scheme, the relay's observation should not be able to infer information about the message. Given the nature of the problem, AF and CF relaying are of particular interest. DF is precluded since it requires the relay to decode the message from its observation. Under this line of research, a very interesting question has been raised: Can we improve the security performance by exploiting relay cooperation with untrusted nodes?

Reference [4], which focuses on the secrecy capacity, appears to be one of the first studies tackling this issue. By considering a three-node model, as shown in Fig. 2a, which includes a source, a destination, and an untrusted relay, it was demonstrated in [4] that the untrusted relay can be beneficial for some specific relaying topologies. Specifically, when there is an orthogonal link in the second hop from the relay to the destination, one achieves higher secrecy rate by treating the relay as an eavesdropper E as well as a helper rather than considering the relay as an eavesdropper only. This interesting result holds true for both AF and CF relaying. On the other hand, when the source and relay transmit to the destination via a multiple access channel, while

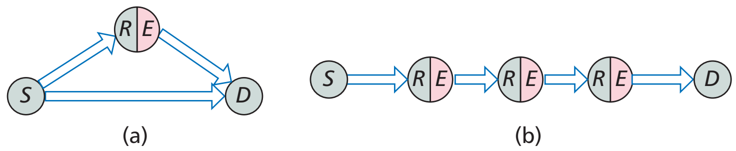


Figure 2. Wireless relay networks with untrusted relays where a relay R acts as both a helper and an eavesdropper E: a) three-node model; b) multihop model.

there is an orthogonal link from the source to the relay, the secrecy capacity is equal to zero [5]. It is because in this case, randomness at the source's encoder is not necessary, and the relay-destination link is not useful in improving the secrecy rate. The results in [4] have also been extended to multi-antenna setups in [6] where the source, destination, and relay are equipped with multiple antennas. In particular, by jointly optimizing the source beamforming vector and the relay beamforming matrix, the cooperative scheme achieves a better secrecy rate than the non-cooperative scheme. However, the proposed beamforming scheme can only be applied to AF relaying. To our knowledge, the corresponding problem associated with CF relaying remains a challenging task. The benefit of secure beamforming in multi-antenna systems was also investigated in [7] for two-way AF relaying. Recently, confidential message transfer over multihop communication with a chain of connected untrusted relays, as illustrated in Fig. 2b, was examined in [8]. Interestingly, under this line network model, end-to-end secrecy can still be achieved, and the secrecy rate has been shown to be independent of the number of hops. Specifically, in this network, interference is created for each relay from its next hop neighbor while it receives a confidential message from the previous hop neighbor. As such, each relay receives a superposition of the message and another signal that is intended for cooperative jamming. Via a coding scheme utilizing nested lattice codes, each relay cannot infer the message from the combination of the message and jamming that corresponds to another codeword.

The secrecy capacity relies on the assumption of ergodic channels and has been considered one of the foremost system benchmarks. However, in some certain fading scenarios, the channel gains change slowly over time. This corresponds to a situation where the coherent time of the channels is sufficiently long compared to the delay requirement. For such cases, the secrecy outage probability can be used as the main performance metric. In [9], the secrecy outage probability has been studied for a three-node non-regenerative AF relay network with an untrusted relay. It is then shown in [9] that secrecy can be achieved as long as the source and destination keep their CSI secret from the untrusted relay.

Trusted Relays

We now turn our attention to the case of trusted nodes for security improvements. In trusted relay scenarios, the source is assisted by a single or multiple *trustworthy* relays to transmit confidential information to the destination in the presence of a passive eavesdropper, in addition to the legitimate parties. Different from the untrusted case, the relays are trusted nodes and can be fully exploited to significantly enhance security. This trusted scenario is of more interest and has received considerable attention in the literature. In the following, we introduce different ways that trusted relays can be used to enhance security.

Strategies

For the scenario of trusted relays, several strategies to improve security have been proposed in the literature. The main techniques are schematically illustrated in Fig. 3 and explained in detail below.

Relaying: Consider first the relaying strategy where the helper nodes aid in transmission by simply relaying information between legitimate nodes (e.g., [3, 10]). Depending on how the information flows, one-way (OW) and two-way (TW) relay protocols have been considered in the literature. In OW relaying, a source node wants to communicate to a destination node with the help of relays, so information flows in a unidirectional fashion (i.e., from source to destination). This is usually carried over two transmission phases: the source communicates with the relays in the first phase, and the relays communicate with the destination in the second one. In TW relaying, two nodes want to exchange data and information flows in a bidirectional manner. This is carried over two or three phases: the nodes communicate to the relay simultaneously or by turns in the first one or two phases, respectively, and the relay broadcasts in the third. An eavesdropper might overhear the information in one or multiple transmission phases.

When only one relay is available, the conventional DF or AF techniques are usually considered in the literature along with OW or TW relay protocols. On the other hand, when multiple relays are available, the most common relaying approach is *distributed beamforming*. In this approach, multiple relays transmit a weighted version of the decoded signal (for DF relays) or the noisy received signal (for AF relays). The weights are designed to steer the information vector away from the eavesdropper and in the direction of the intended destination. Assuming CSI of the links to the eavesdropper at the legitimate nodes, complete *nulling* of the information vector at the eavesdropper can be achieved. Such a beamforming/nulling scheme is applicable to both OW and TW relaying.

Jamming: Consider now the strategy in which the helper nodes do not relay information but instead transmit jamming signals to confound the eavesdropper (e.g., [3]). This is commonly referred to as *cooperative jamming*. Generally speaking, in this approach, two nodes communicate directly with each other while the relays transmit jamming signals independent of the nodes' information. The objective of these signals is to degrade the signal-to-noise ratio at the eavesdropper without degrading that at the intended receiver. For instance, when multiple relays are available, complete *nulling* of the jam-

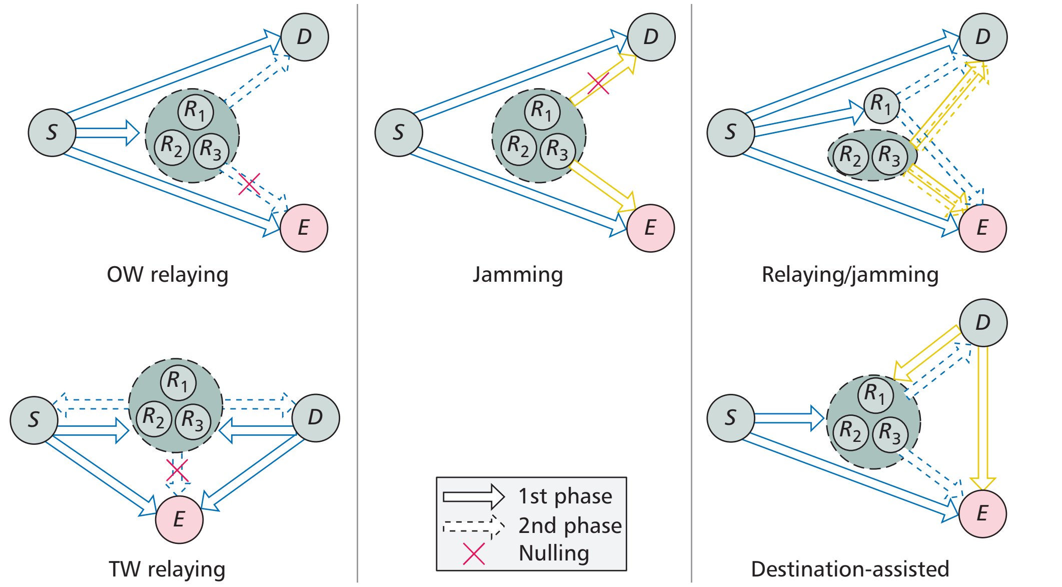


Figure 3. Relay-assisted techniques.

Different from the untrusted case, the relays are trusted nodes and can be fully exploited to significantly enhance the security. This trusted scenario is of more interest and has received considerable attention in the literature. In the following, we introduce different ways that trusted relays can be used to enhance security.

ming signal at the intended receiver is possible by proper weighting of the jamming signals. When information on the links to the eavesdropper can be acquired, the signal-to-noise ratio at the eavesdropper can be further degraded while still achieving nulling of the jamming signal at the destination. Note that different from relaying approaches, the relay nodes do not need to know any information about the signal being transmitted by the source node.

Pure Relaying/Jamming Combinations: The above jamming and relaying approaches can also be combined into a single strategy (e.g., [11]). In this case, a subset of nodes act as relays, while another subset does jamming. Similar to the previous two strategies, beamforming and nulling can be used at any of the subsets for performance enhancement. However, different from jamming techniques, nulling of the jamming signal might be needed not only at the destination node, but also at the relay subset. One special case of relaying/jamming combinations is the so-called *destination-assisted* schemes. In these schemes, the destination node has the double duty of being a receiver and a jammer. Due to the half-duplex constraint, the destination cannot perform both tasks at the same time, and thus the source must communicate through relaying. Specifically, the source can transmit information to a relay subset in the first phase, while the destination and jamming subset transmits noise signals to the eavesdropper. In the second phase, the relay subset simply forwards the information to the destination, which must then remain silent and listen. Note that the techniques described here are referred to as "pure" combinations in that each node acts as either a jammer or a relay at any given time.

Hybrid Relaying/Jamming Combinations: All the above techniques can be said to be part of a more general hybrid strategy in which all nodes are allowed to send superpositions of information and jamming signals, that is, the nodes can simultaneously perform jamming and relaying (e.g., [12, 13]). One of the most well-known hybrid schemes is perhaps the destination-assisted *artificial noise* protocol [12]. In this protocol, the source and destination send jamming signals in the first phase to the relays. In the second phase, the relays transmit a weighted version of the signal received in the previous phase. At the same time, the source sends a superposition of jamming and information signals. The jamming signal in this superposition is designed to cancel the jamming component due to the source at the destination, whereas the jamming component due to the destination can readily be cancelled off since it is known. This artificial noise concept has also been extended to TW relaying. Another destination-assisted hybrid protocol is the one in [13]. In that protocol, the source transmits a combination of data and jamming to the relay, while the destination cooperates with the source by also transmitting a jamming signal. The jamming signals from the source and destination are designed such that their addition will be cancelled at the relay. In the second slot, the relay sends a superposition of information and jamming signals, while the source transmits a different jamming signal. As in the first phase, the source and relay jamming signals are designed to be nulled at the destination. Hybrid protocols that do not require assistance from the destination have also been proposed.

Criteria and Enhancements

Similar to the case of untrusted relays, different criteria have been considered in the literature to optimize the performance of the above strategies. Most works have concentrated on maximizing the secrecy rate, while fewer studies have been carried to minimize the outage performance. For either of these criteria, three aspects have been considered to enhance security.

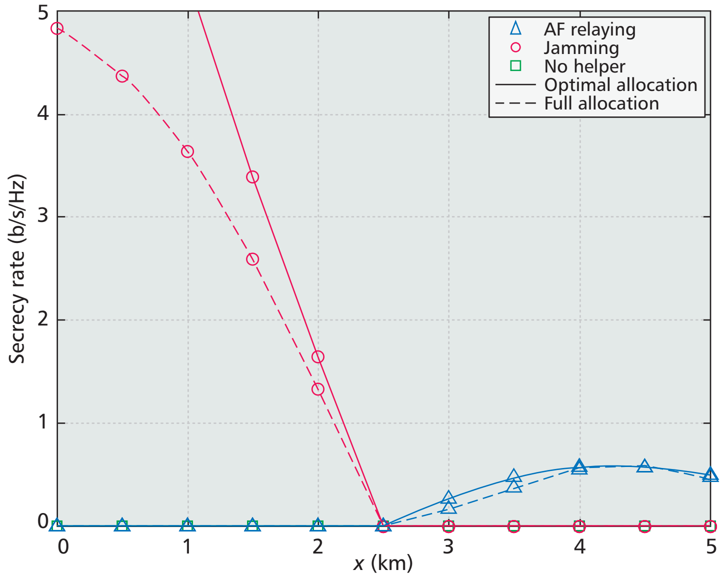


Figure 4. Secrecy rate of different jamming and AF relaying schemes.

The first aspect is *power allocation*, where the total system power must be optimally shared among the nodes (i.e., sum power constraint scenario) or where each node has an individual power budget (i.e., per-node power constraint scenario). In the latter case, we should mention that using full power at any of the nodes might not always be beneficial in certain configurations. For instance, full power at a relay could result in too much information being leaked to the eavesdropper, whereas full power at a jammer might cause too much interference at the destination. For hybrid protocols, further splitting the power between information and jamming signals at a given node is also of great importance.

The second aspect considered in the literature is *weight optimization* at the relays. As discussed above, such optimization is needed for beamforming or nulling of the relaying and jamming signals. It should be emphasized that complete nulling of the jamming vector at the destination or of the information vector at the eavesdropper is in general not always optimal. This is because such a tight constraint could potentially limit the degrees of freedom and the overall performance of the system.

The third and final aspect is *relay selection*, where a subset of all available nodes must be selected for relaying or jamming. Generally speaking, the nodes that increase the interference to the eavesdropper while protecting the destination must be selected for jamming. Likewise, the nodes that improve the quality of the signal received at the destination without increasing that at the eavesdropper must be selected for relaying. Different selection techniques can be applied depending on the availability of the channel information at the controller.

A Case Study

Most of the strategies presented in the previous section require the network to have multiple friendly relays. However, the benefits offered by such multi-relay techniques might be severely undermined by coordination, synchronization, and heavy signaling/feedback issues. This is especially true for techniques in which the channel information among all the links in the network is required at all legitimate nodes. Although the single-relay schemes are simpler to study, their analysis is still very challenging, and such networks have not been thoroughly investigated in the literature, especially for jamming or AF relaying. For instance, although power allocation schemes have been derived in closed form for some DF networks, the globally optimal power allocation schemes at the source and relay that maximize the secrecy rate for the jamming or AF relaying strategies have not been addressed in the literature. This is due to the difficulty in solving the related non-convex optimization problems. Given that jamming and AF relaying present reduced complexity, and the latter has been shown to provide larger secrecy service areas than DF [10], a thorough investigation of these schemes is required.

In this case study, we quantify the gain that can be achieved using optimal power allocation in single-relay networks. We also investigate the effect of relay location on security. Both the jamming and AF relaying strategies according to Fig. 3 are adopted. Specifically, for the jamming strategy, the source communicates directly to the destination while the jammer sends Gaussian noise to both the eavesdropper and the destination. For AF relaying, the source transmits a signal to the relay and destination in the first phase. In the second phase, the relay amplifies what it received in the previous phase and forwards it to the destination. The eavesdropper overhears in both phases. In this case study, the source, eavesdropper, and destination are placed at the corners of a square with sides of 5 km, while the relay can be anywhere inside the square. Specifically, the source, destination, eavesdropper, and relay have coordinates of (0, 0), (5 km, 0), (0, 5 km), and (x, y), where $0 <= x, y <= 5$ km. In addition, the source and relay have a power budget of 40 dBm and 30 dBm, respectively, and the noise power at all nodes is –100 dBm. A path loss model is considered such that the power received at any node is given by $P_{\text{Rx}} = P_{\text{Tx}}/d^{\alpha}$, where P_{Tx} is the transmitted power, d is the distance between the transmitter and the receiver, and α is the path loss exponent, which is set to 3.

To analyze the effect of power allocation on security, Fig. 4 shows the secrecy rate of the jamming and AF relaying strategies when the relay moves along the diagonal of the square from eavesdropper to destination. Two power allocation schemes are considered in this figure: full power at both nodes and the optimal power allocation scheme. The optimal power allocation maximizes the instantaneous secrecy rate for the considered protocols under per-node power constraints and under the assumption of full channel information at the legitimate nodes (similar to [3, 10, 11, 13]). First, note from Fig. 4 that using full power at source and relay is not necessarily optimal. For instance, when the jammer is close to the eavesdropper, a small amount of power is needed to jam it, and further increasing the power affects the destination and thus the overall performance. Similarly, when the relay is close to the destination, using full power might lead to too much information leakage. From this example, power control appears to be more beneficial for the jamming strategy. As expected, it can be seen from Fig. 4 that jamming is preferred when the relay is closer to the eavesdropper, whereas relaying is a better choice when it is closer to the destination. It should also be noted from Fig. 4 that the secrecy rate for AF relaying is zero when the relay is closer to the eavesdropper, whereas that for jamming is zero when the relay is closer to the destination.

To analyze the joint effect of relay location and optimal power allocation, Fig. 5 shows the contour of the secrecy rate when the helping node acts as a relay and is placed at a given (x, y) location. Only the optimal power allocation is considered in Fig. 5. Note from this figure that relaying can achieve a positive rate when the node is closer to the destination than to the eavesdropper. More importantly, the optimal relay location appears to be on the line from source to destination. This is because in this location, the relay is far from the eavesdropper while still being relatively close to the source for listening and the destination for forwarding.

To analyze the optimal location for the jamming strategy, Fig. 6 shows a similar contour plot as above but now assuming that the helping node is a jammer. The optimal power allocation at the jammer is again considered. We can see in Fig. 6 that a positive rate is achieved when the jammer is closer to the eavesdropper than to destination. In this case, the performance of jamming improves as the jammer approaches the eavesdropper.

By comparing the rates in Figs. 5 and 6, we observe that relaying is again preferable when the helping node is closer to the destination ($x > y$), whereas jamming is better when the node is closer to the eavesdropper ($y > x$). Similar trends have also been observed when the eavesdropper moves closer to the destination while keeping the same distance from the source. Finally, it is important to note that using a helping node in this configuration is crucial to achieving a positive secrecy rate. This is because the destination and eavesdropper are at the same distance from the source, so the secrecy capacity without such help would be zero.

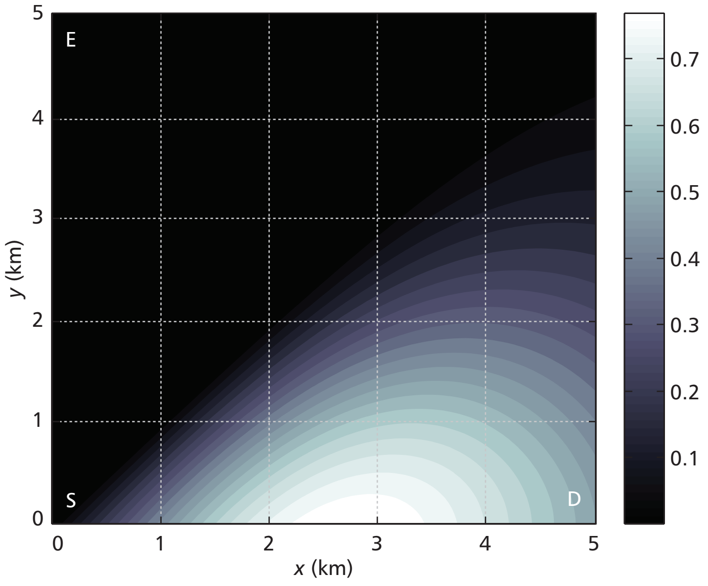


Figure 5. Secrecy rate of AF relaying with different locations.

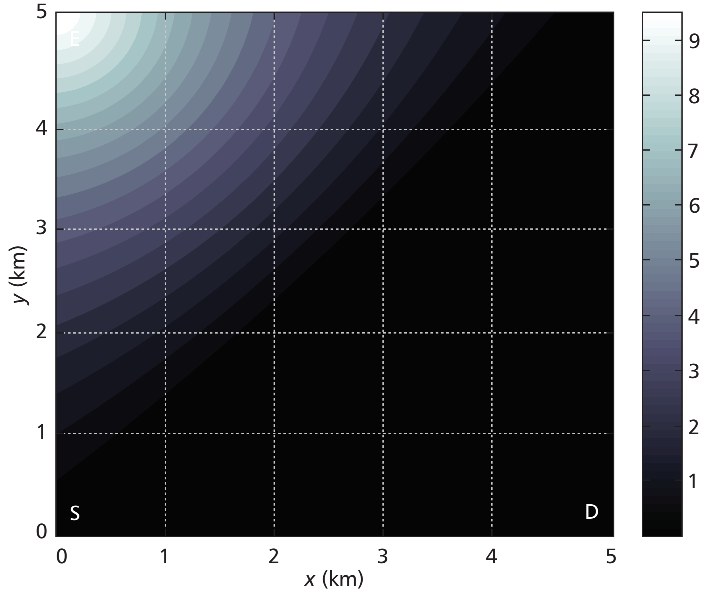


Figure 6. Secrecy rate of jamming with different locations.

Concluding Remarks and Future Research Directions

This article has provided a comprehensive overview of the area of physical layer security in wireless cooperative relay networks. The focus was on both untrusted and trusted relay networks to illustrate that cooperative relaying plays an important role in enhanced security. While the discussion has been at a high level, we hope that the article can motivate further research on PHY security for such important networks. The scope of future research in this direction is broad, and we have no doubt that novel relaying topologies and scenarios along with the corresponding security schemes shall be developed. Therefore, in the following, we would like to present only a few interesting and challenging research topics we believe are worth further investigation.

Thus far, all current relaying strategies considered under the context of PHY security are based on orthogonal or multihop mechanisms. That is, the source and relay transmit information over orthogonal channels. By using this “cake-cutting” approach, only a fraction of the channel degrees of freedom can be exploited. As a consequence, it might fail to realize the full benefits offered by cooperative relaying for enhanced secrecy. To understand the true limitation of cooperative relaying to improve security, more advanced non-orthogonal relay protocols in which the source and relay transmit information simultaneously should be considered. Such

> *Another research challenge is to define a totally new security metric that has both information-theoretic and cryptographic flavor [15] that might lead to a more efficient encryption scheme using cooperative relaying. This research direction shall certainly offer a rich set of challenges.*

a study will certainly result in a more complete picture of the benefits of relaying for security enhancement in wireless networks. Besides relaying, strategies such as jamming or jamming/relaying combinations can also be adopted. Among all these strategies, it is in general not clear which one is better for a given topology. Thus, another interesting research direction is to provide a comparative study to analyze the circumstances under which any one transmission strategy is preferable. Lastly, security enhancements such as power allocation, weight optimization, and relay selection require channel knowledge. Therefore, investigating enhancements that rely on partial or statistical channel information is also of great importance.

Current relaying technologies in wireless communications have been developed under the constraint of half-duplex (HD) communication, where a relay node can either transmit or receive on a single channel, but not both simultaneously. This is because the transmitted signal power in wireless systems is usually many orders of magnitude larger than the received signal power, thus rendering simultaneous transmission and reception over the same frequency band impractical. This HD constraint results in inefficient use of resources as a dedicated bandwidth or time slot is required for relay transmissions. Recently, a number of encouraging full-duplex (FD) designs have been proposed to overcome the self-interference problem using novel combinations of antenna, analog, and digital cancellations. As one important aspect of FD transmission, FD relaying can be exploited to enhance secrecy. For instance, an FD relay node can generate a jamming signal to degrade the eavesdropper channel, while at the same time assisting the transmission from the source to destination. While the potential benefits of FD relaying for enhanced security are undoubted, it is important to investigate jointly cooperative relay and jamming protocols to optimize the secrecy capacity of wireless FD relay networks. To this end, the residual self-interference of FD operation must also be taken into account, which makes the related problems much more challenging [14].

Finally, we note that while PHY security techniques are promising, the security of communication networks has traditionally relied on cryptographic schemes in upper layers, such as the application and presentation layers. Therefore, cross-layer analysis of secrecy to find how best to combine the PHY security and cryptographic schemes in wireless relay networks to guarantee the security of the whole system is another interesting research area. To find such a combination approach, it is important to investigate how the PHY security and traditional cryptographic methods interact with each other to enhance the security of the system. For example, one interesting question is how to combine PHY security techniques in cooperative relaying and cryptographic techniques to build a secret-key agreement protocol, which is to generate a secret key that can be used in a cryptosystem at an upper level. Another research challenge is to define a totally new security metric that has a both information-theoretic and cryptographic flavor [15] that might lead to a more efficient encryption scheme using cooperative relaying. This research direction shall certainly offer a rich set of challenges.

References

[1] X. Zhou, L. Song, and Y. Zhang, *Physical Layer Security in Wireless Communications*, CRC Press, 2013.

[2] A. Wyner, "The Wire-Tap Channel," *Bell Sys. Tech. J.*, vol. 54, no. 87, Oct. 1975, pp. 1355–87.

[3] L. Dong *et al.*, "Improving Wireless Physical Layer Security via Cooperating Relays," *IEEE Trans. Signal Processing*, vol. 58, no. 3, Mar. 2010, pp. 1875–88.

[4] X. He and A. Yener, "Cooperation with an Untrusted Relay: A Secrecy Perspective," *IEEE Trans. Info. Theory*, vol. 56, no. 8, Aug. 2010, pp. 3807–27.

[5] L. Sun *et al.*, "Performance Study of Two-Hop Amplify-and-Forward Systems with Untrustworthy Relay Nodes," *IEEE Trans. Vehic. Tech.*, vol. 61, no. 8, Oct. 2012, pp. 3801–07.

[6] C. Jeong, I.-M. Kim, and D. I. Kim, "Joint Secure Beamforming Design at the Source and the Relay for an Amplify-and-Forward MIMO Untrusted Relay System," *IEEE Trans. Signal Processing*, vol. 60, no. 1, Jan. 2012, pp. 310–25.

[7] J. Mo *et al*., "Secure Beamforming for MIMO Two-Way Communications with an Untrusted Relay," *IEEE Trans. Signal Processing*, vol. 62, no. 9, May 2014, pp. 2185–99.

[8] X. He and A. Yener, "End-to-End Secure Multi-Hop Communication with Untrusted Relays," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, Jan. 2013, pp. 1–11.

[9] J. Huang, A. Mukherjee, and A. L. Swindlehurst, "Secure Communication via an Untrusted Non-Regenerative Relay in Fading Channels," *IEEE Trans. Signal Processing*, vol. 61, no. 10, May 2013, pp. 2536–50.

[10] P. Zhang *et al*., "Analyzing Amplify-and-Forward and Decode-and-Forward Cooperative Strategies in Wyner's Channel Model," *Proc. IEEE WCNC*, Apr. 2009, pp. 1–5.

[11] J. Chen *et al*., "Joint Relay and Jammer Selection for Secure Two-Way Relay Networks," *IEEE Trans. Info. Forensics Security*, vol. 7, no. 1, Feb. 2012, pp. 310–20.

[12] S. Goel and R. Negi, "Guaranteeing Secrecy Using Artificial Noise," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, June 2008, pp. 2180–89.

[13] Y. Liu, J. Li, and A. Petropulu, "Destination Assisted Cooperative Jamming for Wireless Physical-Layer Security," *IEEE Trans. Info. Forensics Security*, vol. 8, no. 4, Apr. 2013, pp. 682–94.

[14] S. Parsaeefard and T. Le-Ngoc, "Improving Wireless Secrecy Rate via Full-Duplex Relay-Assisted Protocols", *IEEE Trans. Info. Forensics & Security*, vol. 10, no. 10, Oct. 2015, pp. 2095-2107.

[15] M. Bellare, S. Tessaro, and A. Vardy, "Semantic Security for the Wiretap Channel," *Advances in Cryptology—CRYPTO 2012*, Lecture Notes in Computer Science, vol. 7417, Springer-Verlag, pp. 294–311.

Acknowledgment

The work of Nghi H. Tran and Sachin Shetty was partially supported by the Office of Naval Research, under Grant No. N00014-15-1-2444.

Biographies

Leonardo Jiménez Rodríguez (S'09) received his B.Eng. degree (with honors) in electrical engineering from Ryerson University, Toronto, Ontario, Canada, in 2008, and his M.Eng. and Ph.D. degrees in electrical engineering from McGill University, Montreal, Quebec, Canada, in 2010 and 2014, respectively. His research interests include cooperative communications, physical layer security, full-duplex transmission, and coded modulation techniques.

Nghi H. Tran (S'05, M'08, SM'15) received a B.Eng. degree from Hanoi University of Technology, Vietnam, in 2002, and M.Sc. (with Graduate Thesis Award) and Ph.D. degrees from the University of Saskatchewan, Saskatoon, Canada, in 2004 and 2008, respectively, all in electrical and computer engineering. Since August 2011, he has been an assistant professor with the Department of Electrical and Computer Engineering, University of Akron, Ohio. His research interests include signal processing, communication, and information theories for wireless systems and networks.

TRUNG Q. DUONG (S'05, M'12, SM'13) received his Ph.D. degree in telecommunications systems from Blekinge Institute of Technology, Sweden in 2012. In 2013, he joined Queen's University Belfast, United Kingdom as a lecturer (assistant professor). His current research interests include cooperative communications, cognitive radio networks, physical layer security, massive MIMO, cross-layer design, mmWave communications, and localization for radios and networks.

THO LE-NGOC (F'97) is a professor in the Department of Electrical and Computer Engineering at McGill University. His research interest is in the area of broadband access communications. He is a Fellow of the Engineering Institute of Canada, the Canadian Academy of Engineering, and the Royal Society of Canada. He was the recipient of the 2004 Canadian Award in Telecommunications Research and the IEEE Canada Fessenden Award 2005. He holds a Canada Research Chair (Tier I) on Broadband Access Communications.

MAGED ELKASHLAN (M'06) received a Ph.D. degree in electrical engineering from the University of British Columbia, Canada, in 2006. From 2007 to 2011, he was with the Wireless and Networking Technologies Laboratory at Commonwealth Scientific and Industrial Research Organization, Australia. During this time, he held an adjunct appointment at the University of Technology Sydney, Australia. In 2011, he joined the School of Electronic Engineering and Computer Science at Queen Mary University of London, United Kingdom, as an assistant professor. He serves as an Editor of *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Vehicular Technology*, and *IEEE Communications Letters*. His research interests fall into the broad areas of communication theory, wireless communications, and statistical signal processing for distributed information processing, security, cognitive radio, millimeter wave communications, and 5G HetNets.

SACHIN SHETTY received a Ph.D. degree in modeling and simulation from Old Dominion University. He also serves as director of the Cyber Security Laboratory and the associate director of the TSU Interdisciplinary Graduate Engineering Research Institute. His research interests lie at the intersection of computer networking, network security, and machine learning. He has published over 70 refereed conference, workshop, and journal articles, and book chapters in research and pedagogical techniques. He has secured over $5 million external funding from several federal agencies. He is the recipient of a DHS Scientific Leadership Award and a TSU Research Mentorship Award.

Multi-Antenna Relay Aided Wireless Physical Layer Security

Xiaoming Chen, Caijun Zhong, Chau Yuen, and Hsiao-Hwa Chen

Xiaoming Chen is with Nanjing University of Aeronautics and Astronautics.

Caijun Zhong is with Zhejiang University.

Chau Yuen is with Singapore University of Technology and Design

Hsiao-Hwa Chen is with National Cheng Kung University.

ABSTRACT

With the growing popularity of mobile Internet, providing secure wireless services has become a critical issue. Physical layer security (PHY-security) has been recognized as an effective means to enhance wireless security by exploiting wireless medium characteristics, for example, fading, noise, and interference. A particularly interesting PHY-security technology is cooperative relay due to the fact that it helps to provide distributed diversity and shorten access distance. This article offers a tutorial on various multi-antenna relaying technologies to improve security at physical layer. The state-of-the-art research results on multi-antenna relay aided PHY-security as well as some secrecy performance optimization schemes are presented. In particular, we focus on large-scale MIMO relaying technology, which is effective in tackling various challenging issues for implementing wireless PHY-security, such as short-distance interception without eavesdropper CSI and with imperfect legitimate CSI. Moreover, the future directions are identified for further enhancement of secrecy performance.

INTRODUCTION

We have witnessed significant growth in wireless communications due to the rapid technological advancements in cellular, sensor, cyber-physical, and machine-to-machine (M2M) communication networks. Applications based on these diverse wireless systems are used to transmit and receive confidential/private data (e.g., credit card information, energy pricing, e-health data, command and control messages). Therefore, it is important to guarantee secure communications in the presence of possible undesired third parties, for example, attackers, eavesdroppers, adversaries with malicious data injection capability, and so on. Traditionally, secure communication systems were implemented using upper layer protocols and tools, such as cryptography. However, cryptography requires an extra secure channel for exchange of private keys. Note that for mobile or unstructured networks, it is difficult to provide a reliably secure channel. Recently, a new paradigm known as physical layer security (PHY-security), which exploits the randomness of the wireless propagation medium, has emerged [1]. The benefits of PHY-security are two-fold. First, heavy dependence on complex higher-layer encryption may not be necessary, leaving more computation resources for communications. Second, PHY-security avoids the use of private keys, and thus can be made more applicable. And in practical systems, PHY-security can serve as an additional layer of protection on top of the existing security features.

From an information-theoretic viewpoint, the essence of PHY-security is to maximize the performance difference between legitimate and eavesdropper channels [2]. Generally speaking, it aims to enhance the legitimate signal and impair the eavesdropper signal simultaneously, thus realizing secure, reliable, and QoS-guaranteed communications. In this context, a variety of physical layer techniques can be utilized to enhance wireless security. The multi-antenna technique is one of the most powerful tools for secure communications. Making use of spatial degrees of freedom, it is possible for us to increase the legitimate channel rate and concurrently decrease the eavesdropper channel rate. As a simple example, if a signal is transmitted in the null space of the eavesdropper channel, the eavesdropper cannot receive any information, and thus information leakage is avoided. It is worth pointing out that the quality of both legitimate and interception signals are related largely to the propagation distance. If the interception distance is short, it is difficult to provide a high quality of service (QoS)-guaranteed secure communication, even exploiting the benefit of the multi-antenna technique. This is because the gain from multi-antenna is small compared to path loss of signal propagation. To address this challenge, relaying technology was introduced into PHY-security to shorten the propagation distance of the legitimate signal [3]. In particular, multi-antenna relaying technology has attracted considerable attention as it has the advantages of both multi-antenna and relaying technologies.

To fully exploit the benefits of multi-antenna relaying technology for PHY-security, it is necessary to adaptively adjust the transmit parameters, such as transmit beams, transmit powers, transmit durations, and relaying protocols [4]. Intuitively, in order to implement these secrecy performance optimization schemes, the transmitters require full or at least partial channel state information (CSI). Unlike traditional relay-

0163-6804/15/$25.00 © 2015 IEEE

ing systems, the secrecy relaying system involves different types of CSI. In addition to the legitimate CSI about source-relay and relay-destination channels, there is the eavesdropper CSI about source-eavesdropper and relay-eavesdropper channels. As revealed in the literature, the CSI has a great impact on the performance of adaptive transmission techniques. If full CSI is available at the source and relay, it is possible to attain a steady secrecy rate, or even achieve the secrecy capacity. However, eavesdropper CSI is usually unavailable, since an eavesdropper can be passive and keep silent. In this case, it is impossible to provide a steady secrecy rate over all realizations of fading channels. To this end, some new performance metrics, that is, ergodic secrecy rate, secrecy outage probability, and interception probability, are proposed accordingly to evaluate wireless security in a statistical sense [5]. Moreover, legitimate CSI may also be imperfect as normally it is obtained through limited feedback or by making use of channel reciprocity. Under this condition, it is nontrivial to design adaptive performance optimization schemes.

In this article, we intend to provide an overview of various state-of-the-art multi-antenna relaying technologies from the perspective of PHY-security. Especially, we investigate viable secrecy performance optimization schemes in the framework of multi-antenna secure relaying systems. Then we discuss and analyze an up-to-date multi-antenna relaying technology, large-scale multiple-input multiple-output (LS-MIMO) relaying, to show the benefits of cooperative schemes for wireless security. At the end, we conclude the whole article with a discussion on future research directions in secure relaying systems.

State-of-the-Art Multi-Antenna Secure Relaying Technologies

Some pioneering works on multi-antenna relaying technologies for PHY-security revealed the fundamental functions of wireless security. Specifically, the multi-antenna relay plays two roles:

- To help the source by enhancing channel quality to the legitimate destination
- To repress the interception by deteriorating the channel condition to the eavesdropper

The performance of multi-antenna relay for PHY-security depends mainly on relaying protocols and schemes. For example, amplify-and-forward (AF) and decode-and-forward (DF) are two commonly used relaying protocols [6, 7]. AF forwards the signal polluted by noise, while DF forwards the original signal by decoding the received signal at the relay. From the perspective of PHY-security, it is not easy to judge which protocol is better. In general, the relaying protocol is selected according to relaying scheme and channel condition. In what follows, we provide an overview of various multi-antenna secure relaying schemes.

One-Way Relaying

One-way multi-antenna secure relaying technology is the most popular relaying scheme. In this case, to accomplish a transmission two time slots are required. As shown in Fig. 1, during the first time slot, the source sends a message to the relay, and then the relay forwards the post-processed signal to the legitimate destination within the second time slot. Meanwhile, the eavesdropper also receives the signals and tries to decode them. In order to improve the secrecy performance, a feasible way is the use of multi-antenna techniques at the relay. For both AF and DF protocols, zero-forcing (ZF), minimum mean square error (MMSE), or a match filter (MF) receiver can be utilized in the first time slot. Similarly, ZF, MMSE, or an MF transmitter is applicable within the second time slot [8]. Then, with different transceivers and relaying protocols at the relay, there are 18 combinations in total. According to channel conditions, it is possible to select an optimal combination. For example, the MMSE receiver can mitigate the noise, and then AF is used due to its low complexity. Moreover, a ZF transmitter can effectively decrease the information leakage if eavesdropper CSI is available. Even without eavesdropper CSI, the transceiver designed based only on legitimate CSI is beneficial for secrecy performance enhancement.

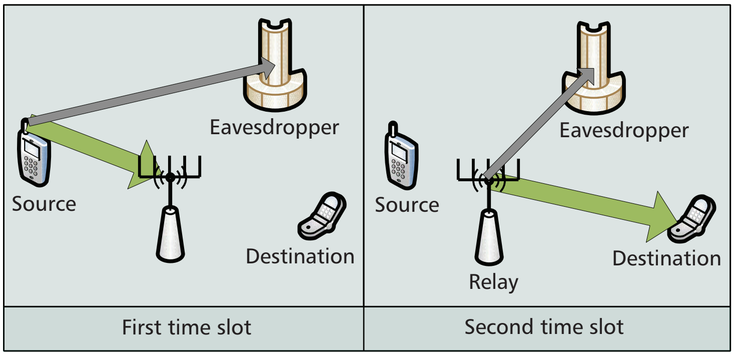


Figure 1. A model of a one-way secure relaying system.

Moreover, with multiple antennas at the relay, cooperative jamming can also be used to further improve the secrecy performance. Specifically speaking, a relay generates interference independent of the source message (e.g., artificial noise) toward an eavesdropper. In order to avoid interference to the destination, the jamming signal is transmitted in the null space of the relay-destination channel, making use of spatial degrees of freedom of the multi-antenna relay. Similarly, the source can also send the jamming signal to interfere with the eavesdropper in the second time slot. It is worth pointing out that there are two potential problems for cooperative jamming. First, if CSI is imperfect, cooperative jamming may result in residual interference to the destination. However, even with residual interference, it may still be beneficial for wireless security to adopt cooperative jamming as long as legitimate CSI is sufficiently accurate. Second, the jamming signal consumes extra power. Thus, in power-limited secure systems, it makes sense to design an energy-efficient cooperative jamming scheme.

Two-Way Relaying

In a two-way relaying case, the source and destination exchange messages with the aid of a multi-antenna relay in two time slots. Specifical-

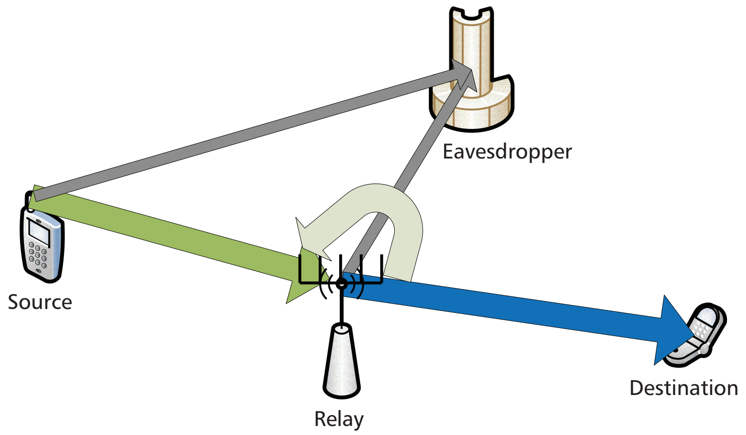


Figure 2. A model of a full-duplex secure relaying system.

ly, two nodes send their signals simultaneously to the relay during the first time slot. Then the relay broadcasts the post-processed mixed signal based on AF or DF relaying protocol. Each node subtracts its transmitted signal from the received signal, and then recovers the information from the other node. Compared to one-way relaying, two-way relaying has two advantages from the perspective of wireless security. First, two-way relaying doubles the spectral efficiency of the legitimate signal transmission. Second, the current transmission of two signals may degrade the quality of the interception signal, since there is no interference cancellation at the eavesdropper.

The key to two-way relaying for PHY-security lies in the design of the transceiver at the multi-antenna relay. On one hand, interference between two legitimate signals should be avoided, while still guaranteeing high spectral efficiency. To this end, some advanced network coding techniques can be used at the relay [9]. For example, physical layer network coding performs an XOR operation to the two signals at the bit level after decoding the two signals from the mixed signal, and then the desired signal can be recovered at each source using XOR operation to the received signal based on its own transmit signal. Moreover, ZF beamforming can also be adopted to separate the two signals in space. On the other hand, wireless security should be fulfilled by decreasing information leakage to the eavesdropper. If full or partial eavesdropper CSI is available, ZF or MMSE beamforming is an effective way to reduce the information leakage. Otherwise, if there is no eavesdropper CSI, cooperative jamming can be used to enhanced wireless security.

Note that in order to decrease the complexity of separating the mixed signal at a relay, it is likely to transform the traditional two-slot two-way relaying to a three-slot scheme. Specifically, one source first sends a message, and then the other source transmits its signal. Finally, the relay broadcasts the post-processed signal. This transformed scheme may weaken the wireless security, since there is no self-interference during the first two time slots. Moreover, it requires a longer transmission time. However, it may achieve a balance between security and complexity. In addition, if the compute-and-forward protocol is used, the relay does not need to decode each signal from the mixed signal. Instead, it can decode a function of the signals and forward it, which can further reduce the complexity.

Full-Duplex Relaying

Both one-way and two-way relaying adopt a half duplex scheme, which separates the processes of transmitting and receiving in time. However, if the relay can simultaneously transmit and receive signals (i.e., full-duplex relaying [10]), as seen in Fig. 2, the spectral efficiency can be doubled with respect to one-way relaying. In addition, the signals from the source and relay may produce extra interference to the eavesdropper, and thus improve the secrecy performance.

Although full-duplex brings great benefits for wireless security, it still faces many challenging issues. The biggest problem is self-interference from the transmitted signal from the relay to the received signal at the relay [10]. Due to relatively short propagation distance, self-interference may severely degrade performance. Intuitively, it is possible to cancel the interference from the received signal, since the relay knows the transmit signal perfectly. However, self-interference is also affected by the loop channel from the transmitter to the receiver. If the CSI for the loop channel is imperfect, the interference cannot be cancelled completely. More importantly, since interference has its pros and cons in PHY-security, it may not be optimal to cancel interference completely for full-duplex relaying. A feasible way is a joint design of transmit and receive beams at the relay in order to achieve a fine balance between the effects of self-interference on the legitimate and interception signals.

Cooperative Relaying

If there is a strict spatial limitation at the relay, it may be impossible to deploy multiple antennas. In this case, multiple single-antenna relays can cooperatively assist secure communications [11]. The advantages of cooperative relaying for PHY-security are two-fold. First, these relays are geographically distributed, so the access distance of the destination may be shortened, and thus the secrecy performance improved. Second, these relays can play different roles according to channel conditions. For example, the relays close to the eavesdropper may act as cooperative jammers so as to generate strong interference to the eavesdropper. The other relays still forward the legitimate signal cooperatively. Compared to cooperative jamming in a co-located multi-antenna relay, the one with cooperative relaying has lower complexity.

However, cooperative relaying also faces some implementation difficulties. Specifically, cooperative relaying is in general carried out in a distributed way. Thus, the synchronization for multiple relays is a nontrivial task, especially for the relays with different roles. Moreover, CSI exchange between the relays is also challenging. It may increase overheads, and an intelligent eavesdropper can obtain the CSI. If it succeeds, these kinds of disruptive attacks can be a serious thread and will significantly impair the secrecy performance as a whole.

Untrusted Relaying

A key feature in relaying systems as described above is that they all assume the relay can be trusted. In other words, the relay will assist secure transmissions in the best way they can. However, from recent research works, several papers have considered the use of untrusted relays [12]. In an untrusted relay model, although the relay is a cooperative node, information intended for the destination must be kept secret from it. Another line of works assumes that the relay is "malicious," that is, the relay may try to modify the retransmitted signal toward the destination. The use of untrusted relay may occur in several cases. For example, in public networks, the relays that are used for connectivity may belong to a third party. Such relays can operate with standard protocols, although they can be unauthenticated. Malicious relay scenarios can occur in military applications as well, where an enemy can "pretend" to be a cooperative node forwarding the malicious data to the destination.

Untrusted relaying has a great impact on secrecy performance. The achievable secrecy rate of the DF protocol is zero, while the AF protocol can achieve a nonzero secrecy rate. A feasible solution to untrusted relaying is the use of cooperative jamming. A friend sends a jamming signal to interfere with the relay, but the destination can completely cancel the interference with prior knowledge. Thus, the secrecy performance in the case of untrusted relaying can be improved.

Adaptive Resource Allocation for Multi-Antenna Secure Relaying

In multi-antenna relay networks, there are different types of resources, such as power, time, space, and antenna resources. These resources affect the quality of both legitimate and interception signals; thus, it makes sense to allocate them according to channel conditions and system parameters [13]. However, resource allocation in secure communications is a nontrivial task. In what follows, we discuss several key issues on resource allocation in multi-antenna secure relaying systems.

Adaptive Beamforming

Beamforming has a great impact on secrecy performance. As aforementioned, if the legitimate signal is transmitted in the null space of the eavesdropper channel, the eavesdropper cannot receive any information. However, implementation of beamforming in secure relaying systems is not easy, especially in the case without eavesdropper CSI. In general, the source and relay design the beams independently. Then, the source constructs a beam aimed at the relay only if it knows legitimate CSI. However, the beamforming design at the relay involves multiple factors. It is quite complex and can only be suboptimal. On one hand, the beamforming scheme is related to the relaying protocols. For example, AF will forward the noise, so it is better to adopt a beam that may achieve a trade-off between enhancing the signal and mitigating the noise (i.e., ZF and MMSE). The DF forwards the original signal; hence, MF beamforming can maximize the signal-to-noise ratio (SNR) at the destination. On the other hand, the receive beam in the first time slot and the transmit beam in the second time slot should be designed jointly. Generally speaking, the receive beam will determine the quality of the legitimate signal, while the transmit beam can impair the interception signal. In addition, if full-duplex relaying is adopted, the receive and transmit beams should be designed carefully to deal with the effect of self-interference.

Power Allocation

In traditional communications without security requirements, the communication quality (e.g., transmission rate) is usually an increasing function of transmit power. However, the power has a side effect in secure communications. This is because increasing the power would simultaneously improve the performance of the legitimate and eavesdropper channels. Thus, the power should be allocated adaptively to the conditions of the legitimate and eavesdropper channels.

In secure relaying systems, power allocation becomes more complicated, since the powers at the source and relay are inter-related. For example, based on the DF relaying protocol, the legitimate channel rate is determined by the smaller of the rates of the source-relay and relay-destination channels. Therefore, it does not make sense to increase the power on one side but fix the other. For the AF relaying protocol, increasing the relay power may amplify the noise, resulting in performance saturation. In addition, if a more advanced relaying scheme is adopted, power allocation should be adjusted accordingly. As a simple example, in full-duplex relaying systems, the relay power directly determines self-interference, so the power allocation should consider the interference cancellation scheme and the effect of the interference on the interception signal. Moreover, for secure relaying systems with cooperative jamming, if the total relay power is constrained, it is necessary to distribute the power between the forwarding signal and the jamming signal.

Time Allocation

In general, the durations of the first and second time slots in relaying systems are equally allocated. Such an allocation scheme is simple and asymptotically optima if the relay is at the middle of the source and destination. However, in secure relaying systems, since the channels from the source to the eavesdropper and from the relay to the eavesdropper may be quite different, equal duration allocation may result in obvious secrecy performance loss. Specifically, if the eavesdropper is closer to the relay, it makes sense to distribute a longer duration to the first time slot. Intuitively, time allocation is also related to the other system parameters (i.e., relaying protocol and transmit power). Hence, time allocation can effectively enhance the secrecy performance.

Antenna Selection

In multi-antenna secure relaying systems, the antennas at the relay have different effects on the secrecy performance if the channels expe-

Untrusted relaying has a great impact on the secrecy performance. The achievable secrecy rate of the DF protocol is zero, while the AF protocol can achieve a nonzero secrecy rate. A feasible solution to the untrusted relaying is the use of cooperative jamming.

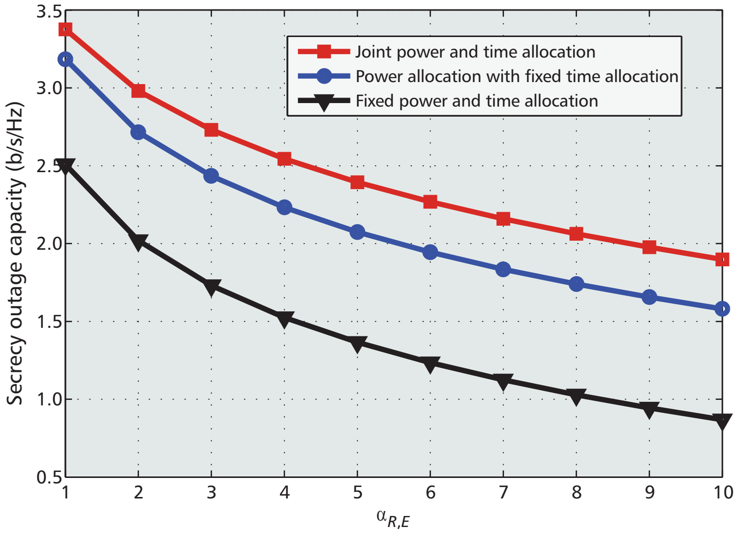


Figure 3. Secrecy performance comparison with joint and fixed power allocation.

rience independent fadings. As mentioned in cooperative relaying, some relays may be closer to the eavesdropper, and thus the forwarding of these relays may lead to information leakage. Even in a co-located multi-antenna relaying system, certain channels from a relay antenna to the destination may experience deep fading, but the channel to the eavesdropper may have high gain. In this case, the use of these antennas not only wastes power, but also degrades the secrecy performance.

Antenna selection in secure relaying systems is not a trivial issue, since it is a combinatorial optimization problem from a pure mathematical viewpoint. If the number of antennas is not very large, it is possible to select the optimal antennas by exhaustive searching. Otherwise, some suboptimal scheme may be used to select the antennas. For example, if eavesdropper CSI is unavailable, the antennas can be selected only according to the quality of the legitimate channels.

Relaying Protocol Switch

There are various relaying protocols, where AF and DF are the two most commonly used ones. In secure relaying systems, there is no dominant protocol. As channel conditions change, the optimal relaying protocol may also vary. Thus, it makes sense to switch the relaying protocols according to channel conditions in order to optimize the secrecy performance.

It is worth pointing out that the above resource allocation schemes are interactive. For example, a power allocation scheme may affect time allocation. Thus, it is better to optimize these resources jointly in order to maximize secrecy performance.

Large-Scale MIMO Relaying for PLS

In secure communications, there may be some adverse conditions (e.g., no eavesdropper CSI and imperfect legitimate CSI). In this context, if the interception distance is relatively short, even with a multi-antenna relay the secrecy performance may be very poor. As a result, it is difficult to provide secure, reliable, and QoS-guaranteed communications.

To solve the problem with short-distance interception in secure communications, we recently proposedto using LS-MIMO relaying technology to enhance wireless security significantly [14]. LS-MIMO can generate a very high-resolution spatial beam, making use of a large number of antennas. Thus, on one hand, the performance of the legitimate channel can be improved enormously due to the high array gain. On the other hand, the information leakage to unintended users can be made very small. In particular, as the number of antennas tends to be infinite, the information leakage is negligible. Then, even under adverse conditions, it is still likely to achieve good secrecy performance. Additionally, compared to traditional multi-antenna secure relaying technologies, LS-MIMO secure relaying technology offers several appealing advantages. First, LS-MIMO simplifies the signal processing, and even with a low-complexity transceiver at the relay, that is, maximum ratio combination (MRC) and maximum ratio transmission (MRT), it is still able to achieve good performance. Second, it is easy to improve secrecy performance by adding antennas only at the relay. Third, due to channel hardening in LS-MIMO systems, performance analysis and optimization become simpler. In what follows, we show the performance gain of several adaptive resource allocation schemes in LS-MIMO secure relaying systems through numerical simulations.

Let us consider a one-way secure relaying system, where the source communicates with the destination with the aid of an LS-MIMO relay. The number of antennas N_R at the relay is very large (e.g., $N_R = 100$ or even bigger). The relay has full CSI about the source-relay channel through channel estimation, imperfect CSI of the relay-destination channel due to channel reciprocity, but no CSI of the relay-eavesdropper channel. The eavesdropper is closer to the relay, but not the source, since it assumes that the signal is from the relay directly. We use $\alpha_{S,R}$, $\alpha_{R,D}$, and $\alpha_{R,E}$ to denote the normalized path loss of the source-relay channel, the relay-destination channel, and the relay-eavesdropper channel, respectively. Note that we take secrecy outage capacity as the performance metric, since eavesdropper CSI is unavailable. Secrecy outage capacity is defined as the maximum transmission rate, while secrecy outage probability needs to satisfy a given constraint. In this manuscript, the bound on secrecy outage probability is set to 0.05.

First, we show the performance gain of joint resource allocation over fixed resource allocation scheme in a DF LS-MIMO secure relaying system. For analysis convenience, we normalize $\alpha_{S,R} = \alpha_{R,D} = 1$, and use$\alpha_{R,E} >> 1$ to represent short-distance interception. We consider the optimization of source power, relay power, and duration ratio between the first and second hops. As seen in Fig. 3, a joint power and time allocation scheme obviously performs better than a power allocation with fixed time allocation scheme. This is because the duration ratio between the two hops also has a great impact on the secrecy performance. For example, if the eavesdropper is close to the relay, it is better to use a small duration in the second hop. Meanwhile, transmit powers at the source and relay also affect the duration ratio. Thus, it makes sense to jointly optimize power and time. Moreover, if both power and time are fixed regardless of channel conditions and system parameters, there will be more performance loss. Thus, joint resource allocation can effectively improve the secrecy performance.

Next, we examine the impact of the number of antennas at the relay on the secrecy performance in an AF LS-MIMO secure relaying system with $\alpha_{S,R} = \alpha_{R,D} = \alpha_{R,D} = 1$. Intuitively, adding more antennas can always improve secrecy outage capacity, but also increases resource consumption, such as power. In this case, we take secrecy energy efficiency as the performance metric, which is defined as the ratio of secrecy outage capacity and total consumed power, including transmit power, circuitry power per antenna, and basic power independent of the number of antennas. In Fig. 4, we use P_C to denote the circuitry power per antenna. It is found that if P_C is very small (i.e., P_C = –20 dB), adding more antennas is always helpful to increase the secrecy energy efficiency. However, with P_C = –10 dB, the energy efficiency first increases and then decreases as the number of antennas increases. This is because when the number of antennas is small, adding more antennas can increase the secrecy outage capacity significantly. However, when the number of antennas is relatively large, although adding more antennas can further increase the secrecy outage capacity, the consumed power increases sharply. Thus, it makes sense to select the optimal number of antennas in order to maximize the energy efficiency.

In summary, adaptive resource allocation can effectively improve the secrecy performance. LS-MIMO secure relaying technology simplifies the signal processing, so it is possible to optimize the utilization of different resources jointly, such as power and time. Therefore, wireless security can be enhanced significantly.

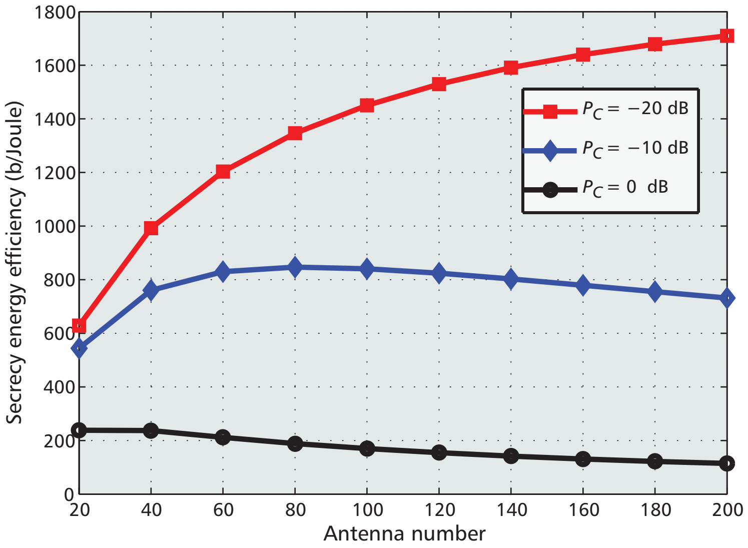


Figure 4. Secrecy energy efficiency of a secure relaying system with different numbers of antennas.

Future Research Directions

Wireless security is always a critical issue. Although the introduction of multi-antenna relay can improve the secrecy performance effectively, there are many challenges that remain to be tackled. As our future work, we intend to solve these problems in the following directions to enhance wireless security further.

Mobile Relay: The position of the relay has a great impact on performance, especially in secure mobile communications. As channel conditions change, the optimal position of the relay may also need to vary accordingly. Hence, a fixed relay may result in obvious performance loss. If it is a vehicular relay, it should be able to flexibly move the position and select the secrecy scheme. For example, the relay moves closer to the eavesdropper to strengthen the interference to the eavesdropper through cooperative jamming. However, it is not a trivial task to design a mobile relay scheme. First, it requires full CSI, which increases the overhead. Second, there is a balance between secrecy performance and implementing complexity, which is again an open issue.

Multiuser Access: In modern communications, multiuser concurrent transmission is commonly used to improve the spectral efficiency. For example, the Long Term Evolution (LTE) system supports multiple users' access through a relay. In multiuser secure relaying systems, multiuser transmission faces several challenges. On one hand, inter-user interference degrades secrecy performance. On the other hand, inter-user interference can be used to impair the interception signal. Thus, it is necessary to design effective user scheduling and precoding schemes to optimize the secrecy performance.

Combination of Encryption and PHY-Security: PHY-security mainly emphasizes pure signal processing techniques, while high-layer cryptographic techniques work well independent of channel conditions. In secure relaying systems, the CSI may be imperfect or even unavailable, and then we can integrate cryptographic techniques into the transceiver design. Combining cryptographic techniques and PHY-security offers another way to improve secrecy performance significantly.

Conclusion

This article provides an overview of multi-antenna relaying technologies in PHY-security, and discusses the opportunities and challenges in the design of secure relaying systems. Through analyzing the characteristics of secure relaying communications, we give a comprehensive tutorial on adaptive resource allocation schemes to further improve the secrecy performance. To solve the problem with short-distance interception under adverse conditions, we propose using LS-MIMO relaying technology and show its effectiveness through simulations. Finally, we identify several research directions for our future work.

In secure relaying systems, the CSI may be imperfect or even unavailable, and then we can integrate cryptographic techniques into the transceiver design. Combining cryptographic techniques and PHY-security offers another way to improve the secrecy performance significantly.

Acknowledgment

This work was supported by the Natural Science Foundation of China (Nos. 61201229, 61301102), the Natural Science Foundation of Jiangsu Province (No. BK20130820), the Zhejiang Science and Technology Department Public Project (2014C31051), the Fundamental Research Funds for Central Universities (2014QNA5019), the iTrust, and Taiwan Ministry of Science & Technology (No. 102-2221-E-006-008-MY3).

References

[1] A. D. Wyner, "The Wire-Tap Channel," *Bell Sys. Tech. J.*, vol. 54, Oct. 1975, pp. 1355–87.

[2] P. K. Gopala, L. Lai, and H. El. Gamal, "On the Secrecy Capacity of Fading Channels," *IEEE Trans. Info. Theory*, vol. 54, no. 10, Oct. 2008, pp. 4687–98.

[3] L. Dong *et al.*, "Improving Wireless Physical Layer Security via Cooperating Relays," *IEEE Trans. Signal Processing*, vol. 58, no. 3, Mar. 2010, pp. 1875–88.

[4] X. Wang, K. Wang, and X. Zhang, "Secure Relay Beamforming with Imperfect Channel Side Information," *IEEE Trans. Vehic. Tech*., vol. 62, no. 5, June 2013, pp. 2140–55.

[5] O. Gungor *et al.*, "Secrecy Outage Capacity of Fading Channels," *IEEE Trans. Info. Theory*, vol. 59, no. 9, Sept. 2013, pp. 5379–97.

[6] K. Park, T. Wang, and M. Alouini, "On the Jamming Power Allocation for Secure Amplify-and-Forward Relaying via Cooperative Jamming," *IEEE JSAC*, vol. 31, no. 9, Sept. 2013, pp. 1741–50.

[7] D. Ng, E. Lo, and R. Schober, "Secure Resource Allocation and Scheduling for OFDMA Decode-and-Forward Relay Networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 10, Oct. 2011, pp. 3528–40.

[8] G. Zhu *et al.*, "Ergodic Capacity Comparison of Different Relay Precoding Schemes in Dual-Hop AF Systems with Co-Channel Interference," *IEEE Trans. Commun*., vol. 62, no. 7, July 2014, pp. 2324–28.

[9] C. Zhang *et al*., "Beamforming for Secure Two-Way Relay Networks with Physical Layer Network Coding," *Proc. IEEE GLOBECOM*, Dec. 2014, pp. 1734–39.

[10] H. Alves *et al*., "On the Performance of Full-Duplex Relaying under Phy Security Constraints," *Proc. IEEE ICASSP*, 2014, pp. 3978–81.

[11] H. Long *et al*., "Secrecy Capacity Enhancement with Distributed Precoding in Multirelay Wiretap Systems," *IEEE Trans. Info. Forensic Security*, vol. 8, no. 1, Jan. 2013, pp. 229–38.

[12] X. He and A. Yener, "Cooperative with an Untrusted Relay: A Secrecy Perspective," *IEEE Trans. Info. Theory*, vol. 56, no.8, Aug. 2010, pp. 3807–27.

[13] H-M. Wang, F. Liu, and X-G. Xia, "Joint Source-Relay Precoding and Power Allocation for Secure Amplify-and-Forward MIMO Relay Networks," *IEEE Trans. Info. Forensic Security*, vol. 9, no. 8, Aug. 2014, pp. 1240–50.

[14] X. Chen et al., "On the Secrecy Outage Capacity of Physical Layer Security in Large-Scale MIMO Relaying Systems with Imperfect CSI," *Proc. IEEE ICC*, Jun. 2014, pp. 1–6.

Biographies

Xiaoming Chen [M'10, SM'14] received his B.Sc. degree from Hohai University in 2005, his M.Sc. degree from Nanjing University of Science and Technology in 2007, and his Ph. D. degree from Zhejiang University in 2011, all in electronic engineering. He is now with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, China. His research interests mainly focus on cognitive radio, multi-antenna techniques, wireless security, interference network and wireless power transfer, and others.

Caijun Zhong [S'07, M'10, SM'14] received his B.S. degree in information engineering from Xiían Jiaotong University, China, in 2004, and his M.S. degree in information security in 2006 and Ph.D. degree in telecommunications in 2010, both from University College London, United Kingdom. From September 2009 to September 2011, he was a research fellow at the Institute for Electronics, Communications and Information Technologies (ECIT), Queens University Belfast, United Kingdom. Since September 2011, he has been with Zhejiang University, Hangzhou, China, where he is currently an associate professor. His research interests include multivariate statistical theory, MIMO communications systems, cooperative communications, and cognitive radio systems. He is an Editor of *IEEE Communications Letters*. He was the recipient of the 2013 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award. He and his coauthors were awarded a Best Paper Award at WCSP 2013. He was an Exemplary Reviewer for *IEEE Communnications Letters* in 2012 and for *IEEE Wireless Communications Letters* in 2012.

Chau Yuen [SM'12] received B.Eng. and Ph.D. degrees from Nanyang Technological University, Singapore, in 2000 and 2004, respectively. In 2005, he was a postdoctoral fellow with Lucent Technologies Bell Labs, Murray Hill, New Jersey. In 2008, he was a visiting assistant professor at Hong Kong Polytechnic University, Kowloon. From 2006 to 2010, he was with the Institute for Infocomm Research, Singapore, as a senior research engineer. Since 2010, he has been an assistant professor with the Singapore University of Technology and Design. He serves as an Associate Editor for *IEEE Transactions on Vehicular Technology*. He received the IEEE Asia-Pacific Outstanding Young Researcher Award in 2012.

Hsiao-Hwa Chen [S'89, M'91, SM'00, F'10] is currently a Distinguished Professor in the Department of Engineering Science, National Cheng Kung University, Taiwan. He obtained his B.Sc. and M.Sc. degrees from Zhejiang University, China, and a Ph.D. degree from the University of Oulu, Finland, in 1982, 1985, and 1991, respectively. He is the founding Editor-in-Chief of Wiley's *Security and Communication Networks Journal* (http://www.interscience.wiley.com/security). Until recently, he served as Editor-in-Chief of *IEEE Wireless Communications*. He is a a Fellow of IET and an elected Member at Large of IEEE ComSoc.

Enhancing Wireless Secrecy via Cooperation: Signal Design and Optimization

Hui-Ming Wang and Xiang-Gen Xia

ABSTRACT

Physical layer security, or information-theoretic security, has attracted considerable attention recently, due to its potential to enhance the transmission secrecy of wireless communications. Various secrecy signaling and coding schemes have been designed at the physical layer of wireless systems to guarantee confidentiality against information leakage to unauthorized receivers, among which the strategy based on the idea of node cooperation is promising. This article provides an overview of the recent research on enhancing wireless transmission secrecy via cooperation. We take a signal processing perspective and focus on the secrecy signal design and optimization techniques to increase secrecy performance. We also propose some future research directions on this topic.

INTRODUCTION

Ensuring secrecy, or privacy, is a fundamental issue in data communications. Specifically, protecting the confidentiality of wireless communications is believed to be more challenging compared to its wireline counterpart, due to the openness of the wireless propagation channel and the broadcast nature of the radio transmission medium. Any receiver located in the covered area of a transmitter can intercept a transmitted signal, putting the information at risk of being decoded by adversarial users. Therefore, enhancing secrecy at the physical layer becomes critical for wireless communications.

The research on physical layer security was pioneered by Wyner in his seminal work [1], where the wiretap channel model was introduced, and the concept of secrecy capacity was defined to evaluate the efficiency of the secrecy transmission against eavesdropping. Roughly speaking, secrecy capacity is the maximum rate at which transmission can achieve reliability and secrecy at the destination, that is, a legitimate receiver can decode the information successfully while an eavesdropper fails to decode it. Mathematically, secrecy capacity is the supremum of the achievable secrecy rate, which is defined as the rate difference between the main channel (from source to destination) and the wiretap channel (from source to eavesdropper). Therefore, a positive secrecy capacity can only be achieved when the main channel is more "advantageous" than the wiretap channel. However, this cannot always be guaranteed in wireless communications due to uncontrollable channel fading. If the main channel is worse than the wiretap channel, the secrecy rate is typically zero, and secure transmission fails.

In such a context, cooperation emerges as a promising way to enhance wireless physical layer security. With the help of multiple cooperative nodes, one can increase the achievable rate of the main channel while decreasing that of the wiretap channel such that the secrecy performance can be greatly improved. In recent years, various secrecy cooperation strategies and techniques have been developed, aimed at providing transmission secrecy via cooperative signal design and optimization. This article is motivated to provide an overview of recent progress on this topic. Note that we only focus on the *signal processing perspective* of the physical layer security of cooperative systems, such as distributed beamforming, cooperative jamming, power allocation techniques, and optimization algorithms, rather than providing a comprehensive survey of the state of the art in the whole field of cooperative physical layer security [2].

The article is organized as follows. We first briefly introduce three information theoretic metrics for wireless physical layer security, which are the objective functions of secrecy scheme design and signal optimization. Then we focus on recent research efforts on enhancing the secrecy via cooperation, and also discuss the impact of channel state information (CSI). Finally, we propose some future research directions on this topic.

INFORMATION-THEORETIC METRICS FOR PHYSICAL LAYER SECRECY

Wyner's original work considered discrete memoryless channels. When investigating wireless transmission, various researchers have generalized the idea of physical layer security to fading channels. So far, there are three metrics used to evaluate the effectiveness of the secrecy schemes:

Hui-Ming Wang is with Xi'an Jiaotong University.

Xiang-Gen Xia is with the University of Delaware.

0163-6804/15/$25.00 © 2015 IEEE

This achievable ergodic secrecy rate is strictly smaller than the secrecy capacity. Nevertheless, in most cases it is more computationally efficient, which is usually taken as the optimization objective function as a lower bound of the ergodic secrecy capacity.

secrecy capacity/rate, ergodic secrecy capacity/rate, and secrecy outage/throughput. Here, we give a brief introduction of these metrics, since they are taken as objective functions of secrecy signal design and optimization applied in different scenarios.

Secrecy Capacity/Rate

Secrecy capacity is a fundamental evaluation metric of the physical layer security. It is shown by Wyner that the secrecy capacity of a degraded wiretap channel is precisely expressed as

$$C_s = \max_{p(X)}\{I(X;Y) - I(X;Z)\}. \tag{1}$$

where X is the source input, Y and Z are the channel outputs at the intended destination and eavesdropper, respectively, $I(\cdot;\cdot)$ is the mutual information, and the maximization is over the input probability distribution $p(X)$. Obviously, to get the secrecy capacity we have to solve the above optimization problem. However, since both $I(X; Y)$ and $I(X; Z)$ are convex functions over $p(X)$, the difference of two convex functions is generally not convex, and consequently, a non-convex optimization problem needs to be solved, which is generally very difficult. To evaluate the secrecy more conveniently and computation affordably, sometimes we simply use the Gaussian signal for X and evaluate the *achievable secrecy rate*, which is defined as the difference between the achievable rates of the main channel and the wiretap channel with Gaussian codebook. Obviously, the achievable secrecy rate is a lower bound of the secrecy capacity, which has usually been taken as the performance metric.

Ergodic Secrecy Capacity/Rate

Secrecy capacity/rate is defined for fixed channels, and does not take the *fading* of the wireless medium into consideration. To characterize the time varying feature of a channel, the *ergodic secrecy capacity* should be considered if the secrecy message can be coded across a sufficiently large number of varying channel states, that is, under delay-tolerant applications. Ergodic secrecy capacity measures the average ability of secrecy transmission over fading channels, which can be achieved by performing rate and power adaptions according to CSI.

However, the optimal transmission scheme achieving ergodic secrecy capacity is very difficult to obtain for various fading channels. Therefore, the achievable ergodic secrecy rate is usually used to evaluate the secrecy performance. An achievable ergodic secrecy *rate* is defined as the difference between the ergodic rates of the main and wiretap channels with Gaussian codebook. This achievable ergodic secrecy rate is strictly smaller than the secrecy capacity. Nevertheless, in most cases it is more computationally efficient, which is usually taken as the optimization objective function as a lower bound of the ergodic secrecy capacity.

Secrecy Outage/Throughput

When the channel undergoes quasi-static fading, encoding over multiple channel states may not be acceptable for delay-limited applications. In this case, one should consider the secrecy outage probability or secrecy outage capacity as the performance measure. A secrecy outage happens when the instantaneous secrecy capacity C_s is less than a target secrecy rate R_s, that is, the target secrecy rate is too high to be supported by the current channel state, and the information security is compromised. The *secrecy outage probability* $\mathcal{P}_{\text{out}}(R_s)$ is defined as the probability that a secrecy outage happens. For an acceptable secrecy outage probability, the largest secrecy rate that can be supported is secrecy outage capacity. Therefore, perfect secrecy transmission at a rate R_s can only be guaranteed by a probability $1 - \mathcal{P}_{\text{out}}(R_s)$ under quasi-static fading.

Since the source may have access to the instantaneous CSI of the main channel, it may not have to fix its target secrecy rate R_s. When the main channel cannot support a secrecy transmission under rate R_s, it would certainly reduce R_s to avoid such an outage. Therefore, R_s can be adjusted adaptively according to the instantaneous CSI of the main channel to maintain a required outage probability, instead of transmitting at a fixed rate. In this case, *secrecy throughput* should be adopted to evaluate the secrecy performance, which is defined as the average achievable secrecy rate over all channel realizations, subject to a required secrecy outage probability. Secrecy throughput has been widely taken as the optimization objective function for signal design under some acceptable secrecy outage constraint in slow fading channels.

Cooperation for Wireless Secrecy

We can see that the metric of secrecy capacity/rate plays a critical role in evaluating the secrecy level of physical layer transmission. Basically, the secrecy performance depends heavily on the superiority of the main channel to the wiretap channel, just as shown in Eq. 1. However, this superiority cannot always be guaranteed in wireless propagation environments. Fortunately, although the physical channels cannot be controlled, by appropriate signal design and optimization, we can construct *equivalent channels*, which may guarantee or enhance such superiority. This is indeed the motivation of applying cooperation to enhance physical layer security.

Node cooperation, originally proposed to provide diversity gain to combat fading for single-antenna wireless communications systems, has become one promising technique to enhance the physical layer security. In this scheme, one/multiple cooperative nodes help the source deliver confidential signals to the destination against one/multiple eavesdroppers. In [3, 4], information theoretical strict deviations have shown that cooperative helpers provide great potential to secure wireless transmissions, which triggered significant research interest in this topic [5–15]. For more on information theoretical security issues, we refer the reader to [2]. Roughly speaking, cooperation is able to improve the quality of the main channel and/or degrade that of the wiretap channel to establish and enhance the superiority of the equivalent main channel to the equivalent wiretap channel, and to enhance the secrecy.

Generally, the roles of cooperative nodes securing legitimate transmissions can be divided into two categories: cooperative relaying and cooperative jamming. When cooperative nodes help to relay, they employ amplify-and-forward (AF) or decode-and-forward (DF) protocol to forward the confidential information in a collaborative manner, such as cooperative beamforming (CB) and relay selection techniques. Alternatively, cooperative nodes can also transmit jamming signals (also known as artificial noise) collaboratively to prevent any efficient interception of eavesdroppers. Relaying is to improve the quality of the main channel, and jamming is to degrade that of the wiretap channel, both of which enhance the secrecy metrics mentioned above. Figure 1 provides a figure of the methodologies.

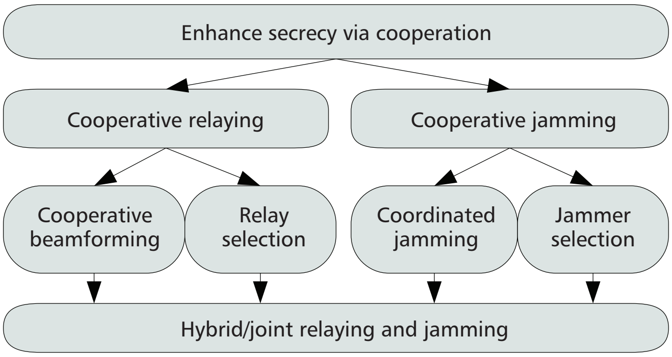


Figure 1. Cooperation schemes for physical layer secrecy.

Mathematically, based on the expression of the secrecy capacity/rate in Eq. 1, a common problem all these secrecy schemes face is to solve a non-convex optimization problem. Therefore, the global optimal solution is generally not available. Efforts have been made to find some suboptimal design with a tractable complexity, and various optimization algorithms have been proposed in [5, 6], which are detailed in the following.

Cooperative Relaying

In a conventional cooperative network, relay nodes are used to expand coverage/distance and/or combat fading. Therefore, in a secrecy transmission, cooperative nodes can act as relays to connect source and destination. Consider a secrecy transmission of one source-destination pair in the presence of one or more eavesdroppers with the help of multiple relay nodes. The cooperation is divided into two phases. In the first phase, the source broadcasts confidential signals to all the relay nodes and possibly the eavesdroppers. In the second phase, the relay nodes forward the received signals to the destination in an AF or DF manner, which gives eavesdroppers another chance to intercept the signals. Both destination and eavesdroppers combine signals received in the two phases to do decoding.

Cooperative beamforming is an efficient technique adopted at relay nodes to forward confidential signals, which can provide both diversity and power gains for the destination to greatly enhance the rate of the legitimate channel (source-relays-destination) (Fig. 2a). Concurrently, the forward signals can also be designed to superimpose destructively or even null out at the eavesdroppers. In such a way, the secrecy rate can be greatly increased. In [5, 6], weight coefficients and power allocations are optimized to maximize the achievable secrecy rate

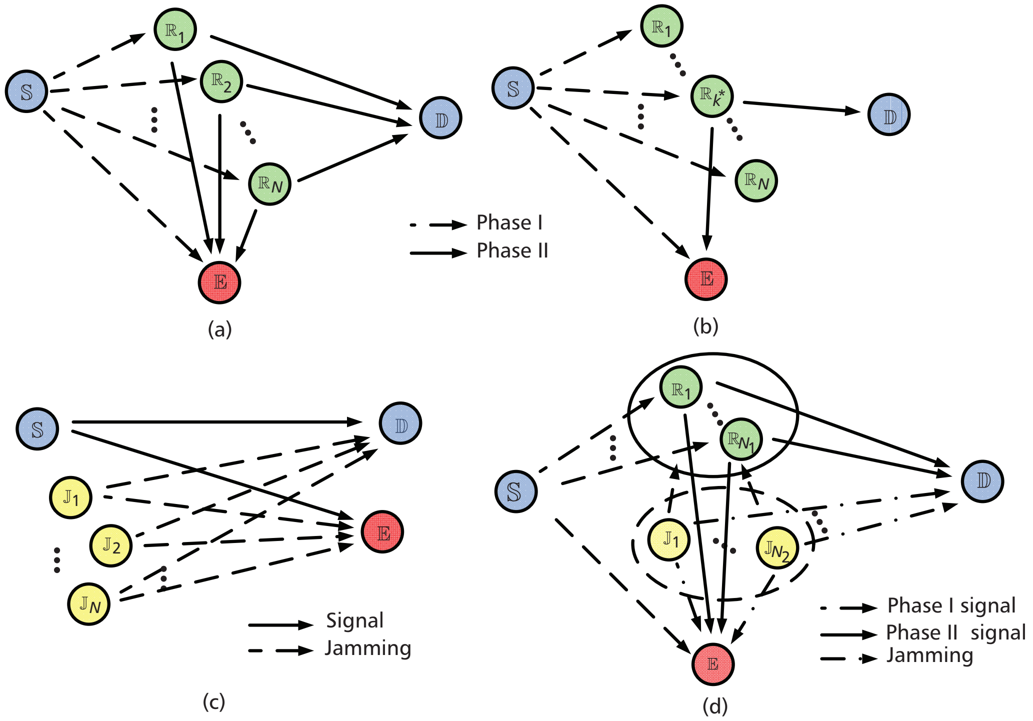


Figure 2. Cooperative schemes for physical layer security: a) cooperative beamforming; b) relay selection; c) cooperative jamming; d) hybrid beamforming and jamming.

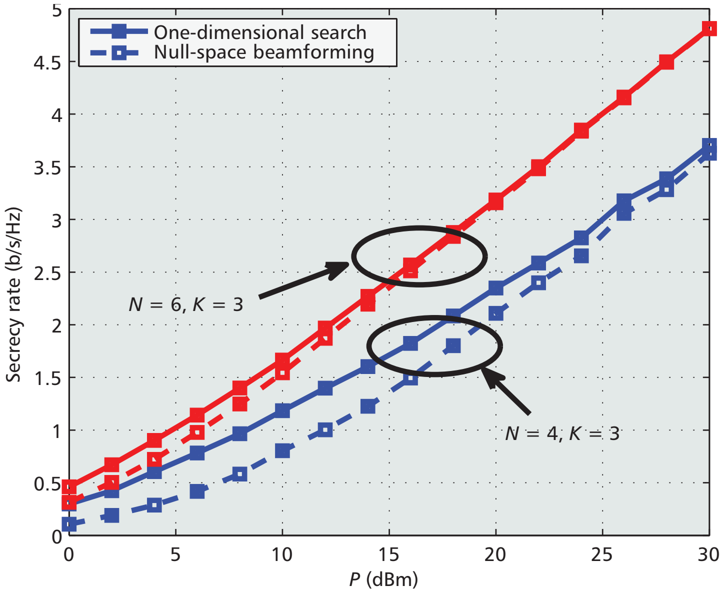


Figure 3. Comparison of null-space beamforming and optimal beamforming via one-dimensional search. K is the number of eavesdroppers, and N is the number of relay nodes. P is the total power budget of all the cooperative nodes, and each node has equal power constraint P/N. The noise power is normalized to 0 dBm.

under a total power constraint or minimize the total power consumption under an achievable secrecy rate constraint, with AF and DF relay nodes, respectively. In [7], a similar idea has been extended to AF two-way relay networks. A common problem involved in maximizing the achievable secrecy rate of a cooperative beamforming network is to maximize a difference of two concave functions over a convex set, which is generally non-convex; thus, there is no efficient algorithm to find the global optimum directly. To get a more tractable solution, constraining the forwarded signals in the null space of the channels from relays to the eavesdroppers is a possible approach. In such a way, information leakage in the second phase will be completely eliminated, and closed forms or semi-closed forms of the optimal weights and power allocations can be conveniently obtained [5–7]. More importantly, the suboptimal solutions are asymptotically optimal in the high signal-to-noise ratio (SNR) regime when the number of relay nodes is much larger than that of eavesdroppers. The reason is as follows. The extra zero-forcing constraint reduces the degrees of freedom provided by relay nodes, and thus decreases the diversity gain they can provide for the legitimate channel (there is no power gain loss due to the total power constraint). When the number of relay nodes is much larger than that of eavesdroppers, this reduction of degrees of freedom will not impact the performance significantly since the diversity gain diminishes quickly as the degrees of freedom increase, as shown in Fig. 3.

When there is only one relay to help the source, cooperative beamforming degrades to power allocation for that relay. In [8], how to maximize the secrecy throughput under a given secrecy outage probability constraint via power allocation optimization and secrecy rate design is addressed. Adaptive and non-adaptive design schemes are investigated, and both of the optimization problems are non-convex. Fortunately, at high and low SNR regimes, closed-form solutions very approximate to the optimal one can be obtained.

Another cooperative relaying scheme is to just select a single relay out of a bunch of relay nodes that can yield the best secrecy performance for cooperation in the second phase, that is, *relay selection* (Fig. 2b). Intuitively, the node with the best channel to the destination and the worst channel to the eavesdropper should be selected as the active relay, which can increase the legitimate channel rate and decrease the wiretap channel rate simultaneously. However, we should have a trade-off between these two. In [9], three opportunistic relay selection schemes are investigated: selecting the node with the lowest instantaneous SNR to the eavesdroppers, selecting the node with the highest SNR to the destination, and selecting the node with the maximal ratio between its SNR and the maximum among the corresponding SNRs to the eavesdroppers. The system performance in terms of probability of non-zero achievable secrecy rate, secrecy outage probability, and achievable secrecy rate of the three schemes has been analyzed. Obviously, compared to cooperative beamforming schemes, the relay selection scheme only provides diversity gain (no power gain) for the legitimate channel, and is not able to eliminate any leakage in the second phase. It sacrifices secrecy performance for lower complexity since cooperative beamforming requires time and frequency synchronization, which results in extra overhead.

Cooperative Jamming

Cooperative jamming is also known as cooperative artificial noise transmission, or noise forwarding [4]. Cooperative nodes become jammers to transmit no-information-bearing signals (artificial noise) to cover confidential signals (Fig. 2c). Since the jamming signals will interfere with both destination and eavesdroppers, they should be designed carefully to establish the superiority of the equivalent legitimate channel to the equivalent wiretap channel in terms of signal-to-interference-plus-noise ratio (SINR). Similarly, there are two kinds of cooperative jamming schemes: coordinated jamming and jammer selection.

In coordinated jamming, jamming signals should be designed coordinately to focus on the wiretap channels while bypassing the legitimate channels. In [5, 6], a scenario where one single antenna source-destination pair communicates with the help of multiple cooperative jammers in the presence of one or multiple eavesdroppers has been investigated. With full CSI, these jammers transmit coherent jamming signals (each jammer transmits a weighted version of a common jamming signal) concurrently with the secrecy signal transmission from the source to the destination. The goal is to design these weight coefficients to maximize the achievable secrecy rate under the total power constraint for the jammers. A similar problem is considered in [10] where the jammers are imposed with individual power constraints. Just as in cooperative beamforming schemes, the objective functions in both cases are non-convex. Therefore, to get a suboptimal but more tractable solution, the null-space transmission scheme is adopted again, where the jamming signals are completely null at the destination. In this case, a closed-form solution can be obtained when only one eavesdropper exists [5, 6]. Without the zero-forcing constraint, a two-level numerical optimization is required, where the inner problem is a convex problem, and the outer one requires a one-dimensional search to obtain the solution, which is more computationally complex [10].

Jammer selection is a low-complexity alternative to coordinated jamming, where only one node is selected as a jammer. Without coordination, the broadcast jamming signal will interfere with both the destination and the eavesdropper. Therefore, intuitively the jammer should be selected to have the worst channel to the destination and the best channel to the eavesdropper. In [11], jammer selection is investigated in a two-way relay network, where one jammer is selected in each phase, and secrecy gain is analyzed compared to no jammer case.

When the transceivers are equipped with multiple antennas, how to design covariance matrices of secrecy signal and jamming signals to maximize the secrecy rate becomes even more difficult. The problem now becomes a non-convex optimization with matrix arguments, which makes the numerical search very unaffordable. In [12], the authors consider a source-destination pair coexisting with one eavesdropper and one jammer, all with multiple antennas. They propose a scheme that can guarantee to provide a secrecy rate larger than or at least equal to the secrecy capacity of the wiretap channel without jamming; that is, a jamming scheme guarantees to improve the secrecy rate. In this case, a closed-form solution of the covariance matrix for the jamming signal is obtained. However, how far the gap is to the optimal solution is not clear. So far, a systematic way to find a suboptimal solution is sill unavailable.

Hybrid Cooperative Relaying and Jamming

If we compare the cooperative relaying and cooperative jamming schemes, we find two interesting observations:

First, for cooperative relaying schemes, due to the half-duplex constraint of the transceivers, two phases are required for one round of data transmission from the source to the destination. Each phase grants the eavesdropper an opportunity to wiretap the information, which in fact improves the quality of the wiretap channel as well. Cooperative beamforming, or relay selection, operates only in the second phase, while during the first phase, all the relay nodes listen to the source, so no one helps protect the transmission. The first phase becomes a secrecy bottleneck of the whole system, especially when the source is only equipped with a single antenna.

Second, for cooperative jamming schemes, although the eavesdropper may only have one chance to intercept the secrecy signal, the quality of the legitimate channel from source to destination has not been improved via cooperation. Since the secrecy capacity is always smaller than the capacity of the legitimate channel, its poor quality due to fading greatly degrades the secrecy capacity, especially when both the source and destination are only equipped with a single antenna.

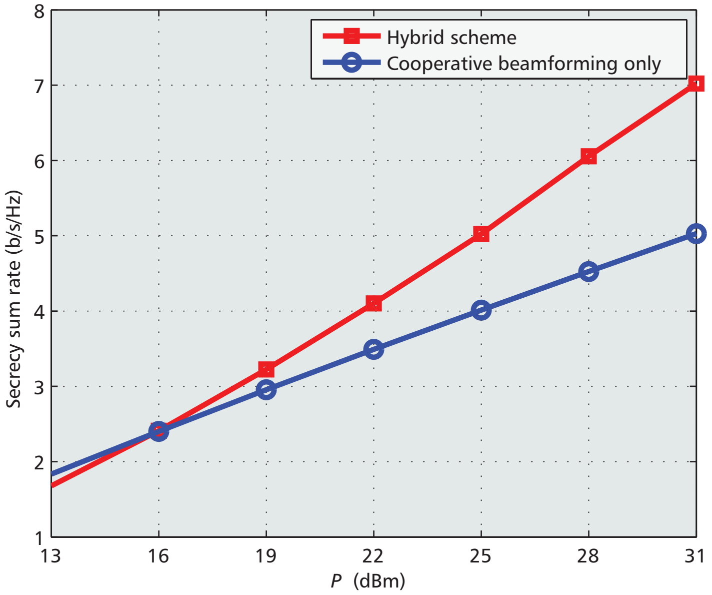


Figure 4. Comparison of the hybrid scheme and the cooperative beamforming scheme in a two-way relay network with $N = 16$. P is the total power budget of all the cooperative nodes, and each node (both the relays and the jammer) has equal power constraint P/N. The noise power is normalized to 0 dBm. The detailed simulation parameters are in [14].

Based on the observations, some *hybrid* schemes are proposed to combine the advantages of both the relaying and jamming strategies to enhance secrecy capacity further. In a hybrid relaying and jamming scheme, multiple cooperative nodes are divided into two groups: relay group and jammer group. The nodes in the relay group help relay confidential signals, while those in the jammer group perform cooperative jamming (Fig. 2d). In such a way, the legitimate channel is improved, and concurrently the eavesdropper is perturbed, which enlarges the difference between these two channels and enhances secrecy capacity. For example, in [13], a hybrid cooperative beamforming and jamming scheme for single-antenna AF two-way relay networks is proposed, where one node is selected as a jammer, and the remaining ones are relay nodes. In the first phase, the relay nodes listen to the sources, while the jammer broadcasts jamming signals to protect the transmissions. In the second phase, the relay nodes forward the received signals via cooperative beamforming, and the jammer keeps silence to avoid interfering with the destinations. This hybrid scheme puts data transmissions in both phases under protection. However, the secrecy sum rate maximization problem by optimizing the cooperative beamformer under the individual power constraint of each relay node is a non-convex one, even when we restrict the beamformer in the null space of the eavesdroppers' channel. With the rate-split technique, an iterative algorithm based on a penalty function method is proposed in [13], which guarantees achieving a stationary solution. A special case is when the two-way channel is reciprocal such as in the time-division duplex

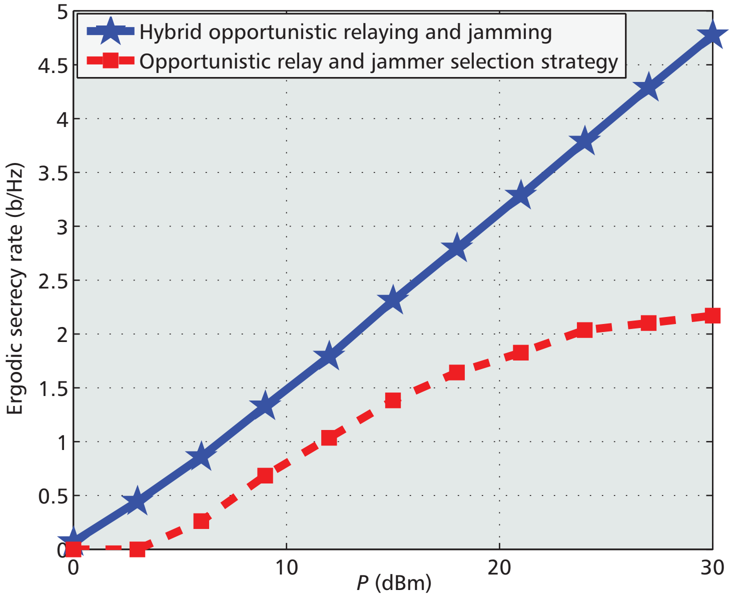


Figure 5. Performance comparison between the hybrid opportunistic relaying and jamming strategy with the relay and jammer selection strategy with $N = 30$. P is the total power budget of all the cooperative nodes, and the noise power is normalized to 0 dBm. The detailed simulation parameters are in [15].

(TDD) mode, the problem becomes a second order convex cone programming (SOCP), which is easier to handle. In Fig. 4, we compare this hybrid scheme with the cooperative beamforming only scheme, where in the latter scheme all the cooperative nodes are relays to forward the confidential signal. We can see that the hybrid scheme indeed improves the secrecy sum rate.

In [14], another hybrid opportunistic relaying and jamming scheme has been proposed for a DF one-way relay network with one eavesdropper, where one "best" relay node is chosen to forward the desired signal, and the remaining nodes are jammers. In both phases, the jammers transmit jamming signals collaboratively in the null space of the channel to the receiver (selected relay or destination). The objective is to maximize the ergodic secrecy rate by optimizing the power allocation to the secrecy signals and jamming signals under a total power budget. This problem is quite difficult since the objective function is also non-convex, and does not even have a closed-form expression. Using the limiting distribution technique of extreme order statistics when the number of jammers is sufficiently large, an asymptotically tight closed-form lower bound has been built and optimized. A sequential parametric convex approximation (SPCA) algorithm is proposed to solve the problem, which provides a Karush-Kuhn-Tucker (KKT) solution. In Fig. 5, we compare the hybrid scheme with the relay and jammer selection strategy, where in the latter scheme a jammer is only active in the second phase. We can see that this hybrid scheme greatly improves the ergodic secrecy rate again.

When the transceivers are equipped with multiple antennas, the additional degrees of freedom facilitate more sophisticated secrecy transmission schemes. In [15], the authors consider a scenario where a source-destination pair communicates with the help of a DF relay in the presence of an eavesdropper, and all four terminals are equipped with multiple antennas. In the first phase, the source transmits secrecy signals together with jamming to the relay, while concurrently the destination transmits jamming signals. In the second phase, the relay forwards the secrecy signals together with the jamming signals, and simultaneously the source transmits jamming signals. The problem is to maximize the secrecy rate via optimizing the linear precoder/decoder matrices at the source and relay under the total or individual power constraints. For the case of single stream transmission, closed-form jamming beamformers and the corresponding optimal power allocation can be obtained. For the transmission of multiple streams, a generalized singular value decomposition (GSVD)-based relaying scheme has been proposed, and the optimal power allocation is found via geometric programming (GP).

Channel State Information

Prior knowledge about the CSI of the main and wiretap channels is critical in both the choice of secrecy metric and the design of the secrecy scheme. In particular, to maximize the achievable secrecy rate requires both instantaneous main and wiretap CSI. The acquisition of the main channel CSI is similar to the conventional methods such as training/estimation and feedback. However, in practice, perfect CSI may not be obtained due to estimation error and feedback delay, which will impact the secrecy performance. For example, when CSI estimation error exists, jamming signals cannot be perfectly nulled out at the destination, which causes interference to the main channel and harms secrecy. Therefore, a robust design is required.

Regarding the wiretap channel CSI, in a scenario where the "eavesdropper" is also a legitimate terminal in the network but not a desired receiver of the current transmission, the assumption that its instantaneous CSI can be available is reasonable. On the other hand, when the eavesdropper is malevolent or even hostile, to get its CSI is very difficult, if not impossible, since it usually works in a passive way without transmitting anything. In this case, only some channel distribution information (CDI) of the wiretap channel may be assumed according to the propagation environment, to maximize the ergodic secrecy rate or secrecy throughput subject to a secrecy outage constraint.

When the eavesdroppers' CSI is completely absent, we cannot do any optimization since the expression of the secrecy rate is not available. In this case, a practical approach is a QoS-based scheme, where a transmitter first allocates part of its resources (power) to guarantee a required target rate for the destination, and then uses the remaining resources to jam the eavesdropper. The optimization problem now becomes how to consume as few resources as possible to fulfill the QoS requirement of the destination so that as much power as possible can be used to jam eavesdroppers. In [8, 6, 14], this QoS-based secrecy strategy has been applied in various scenarios.

Future Directions

The above mentioned research works already show that node cooperation has great potential to secure wireless transmission. However, further investigations are still needed to provide more comprehensive understanding of this strategy and make it more practical. In the following we propose some possible future research directions.

First, the impact of relative spatial locations of the source, destination, eavesdropper, and cooperative helper on the secrecy schemes and secrecy performance needs further exploration. Obviously, if the source is very close to the destination or the eavesdropper is very close to the cooperative helper, direct transmission with cooperative jamming should be more secure. However, most current works neglect the relative spatial locations of these nodes and the large-scale fading between them, simply assuming that the channels are only subject to small-scale fading. Recently, stochastic geometry tools have been adopted to describe the spatial location distribution of random networks. It may be interesting to use it to do secrecy analysis.

Second, the robustness of a secrecy scheme to the imperfection of CSI needs further investigation. We already know that secrecy performance depends heavily on CSI. Since CSI should be estimated/fed back in practice, the accuracy of the channel estimation and feedback delay will definitely impact the secrecy performance. In [15] the impact of the outdated CSI to the ergodic secrecy rate has been investigated. More research is still needed for other schemes.

Third, uncoordinated jamming schemes need to be explored. When multiple jammers help to attack an eavesdropper, they should coordinate their transmissions to avoid interfering with the destination. Since they are spatially distributed, this implies that the global CSI of multiple jammers should be collected, and the collaborative jamming weights should be optimized in coordination, which requires extra overhead. Therefore, it is attractive to find some uncoordinated jamming schemes such that each jammer operates only according to its local CSI.

Fourth, the impact of full-duplex transceiver techniques on cooperative physical layer security must be further investigated. Full-duplex transceiver techniques have received great interest in 5G mobile communications systems. With full-duplex relays and/or jammers, secrecy signals and jamming signals can be received and transmitted simultaneously in the same frequency band, which will significantly impact the secrecy performance.

Acknowledgment

This work was partially supported by the NSFC under Grants No. 61102081 and No. 61221063, the Foundation for the Author of National Excellent Doctoral Dissertation of China under Grant 201340, the National High-Tech Research and Development Program of China under Grant No. 2015AA011306, HuaWei Development Fund under Grant YB2013110059, the New Century Excellent Talents Support Fund of China under Grant NCET-13-0458, the Fok Ying Tong Education Foundation under Grant 141063, the Fundamental Research Funds for the Central University under Grant No. 2013jdgz11, and the Young Talent Support Fund of Science and Technology of Shanxi Province under Grant 2015KJXX-01.

References

[1] A. D. Wyner, "The Wire-Tap Channel," *Bell Sys. Tech. J.*, vol. 54, 1975, pp. 1355–87.

[2] R. Bassily *et al.*, "Cooperative Security at the Physical Layer: A Summary of Recent Advances," *IEEE Signal Processing Mag*., vol. 30, no. 5, Sept. 2013, pp. 16–28.

[3] E. Tekin and A. Yener, "The General Gaussian Multiple Access and Two-Way Wire-Tap Channels: Achievable Rates and Cooperative Jamming," *IEEE Trans. Info. Theory*, vol. 54, no. 6, June 2008, pp. 2735–51.

[4] L. Lai and H. El Gamal, "The Relay-Eavesdropper Channel: Cooperation for Secrecy," *IEEE Trans. Info. Theory*, vol. 54, no. 9, Sept. 2008, pp. 4005–19.

[5] L. Dong *et al*., "Improving Wireless Physical Layer Security via Cooperating Relays," *IEEE Trans. Signal Processing*, vol. 58, no. 3, Mar. 2010, pp. 1875–88.

[6] J. Li, A. P. Petropulu, and S.Weber, "On Cooperative Relaying Schemes for Wireless Physical Layer Security," *IEEE Trans. Signal Processing*, vol. 59, no. 10, Oct. 2011, pp. 4985–97.

[7] H.-M. Wang, Q. Yin, and X.-G. Xia, "Distributed Beamforming for Physical-Layer Security of Two-Way Relay Networks," *IEEE Trans. Signal Processing*, vol. 60, no. 7, July 2012, pp. 3532–45.

[8] T.-X. Zheng *et al*., "Outage Constrained Secrecy Throughput Maximization for DF Relay Networks," *IEEE Trans. Commun*., vol. 63, no. 5, May 2015, pp. 1741–55.

[9] V. N. Q. Bao, N. L.-Trung, and M. Debbah, "Relay Selection Schemes for Dual-Hop Networks under Security Constraints with Multiple Eavesdroppers," *IEEE Trans. Wireless Commun*., vol. 12, no. 12, Dec. 2013, pp. 6076–85.

[10] G. Zheng, L.-C. Choo, and K.-K. Wong, "Optimal Cooperative Jamming to Enhance Physical Layer Security using Relays," *IEEE Trans. Signal Processing*, vol. 59, no. 3, Mar. 2011, pp. 1317–22.

[11] J. C. Chen *et al*., "Joint Relay and Jammer Selection for Secure Two-Way Relay Networks," *IEEE Trans. Info. Forensics and Security*, vol. 7, no. 1, Feb. 2012, pp. 310–20.

[12] S. Fakoorian and A. Swindlehurst, "Solutions for the MIMO Gaussian Wiretap Channel with a Cooperative Jammer," *IEEE Trans. Signal Processing*, vol. 59, no. 10, Oct. 2011, pp. 5013–22.

[13] H.-M. Wang et al., "Hybrid Cooperative Beamforming and Jamming for Physical-Layer Security of Two-Way Relay Networks," *IEEE Trans. Info. Forensics and Security*, vol. 8, no. 12, Dec. 2013, pp. 2007–20.

[14] C. Wang, H.-M. Wang, and X.-G. Xia, "Hybrid Opportunistic Relaying and Jamming with Power Allocation for Secure Cooperative Networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, Feb. 2015, pp. 589–605.

[15] J. Huang and A. L. Swindlehurst, "Cooperative Jamming for Secure Communications in MIMO Relay Networks," *IEEE Trans. Signal Processing*, vol. 59, no. 10, Oct. 2011, pp. 4871–85.

Biographies

Hui-Ming Wang [S'07, M'10] (xjbswhm@gmail.com) is a professor with the Department of Information and Communications Engineering, Xi'an Jiaotong University, and also with the Ministry of Education Key Lab for Intelligent Networks and Network Security, China. From 2007 to 2008 and 2009 to 2010, he was a visiting scholar at the Department of Electrical and Computer Engineering, University of Delaware. His research interests include cooperative communication systems, physical-layer security of wireless communications, MIMO, and space-timing coding.

Xiang-Gen Xia [M'97, SM'00 'F'09] (xxia@ee.udel.edu) received his Ph.D. in electrical engineering from the University of Southern California, Los Angeles, in 1992. He is currently the Charles Black Evans Professor in the Department of Electrical and Computer Engineering, University of Delaware. His current research interests include space-time coding, MIMO and OFDM systems, digital signal processing, and SAR and ISAR imaging. He is the author of the book *Modulated Coding for Intersymbol Interference Channels* (Marcel Dekker, 2000).

Full-duplex transceiver techniques have received great interest in 5G mobile communications system. With full-duplex relays and/or jammers, secrecy signals and jamming signals can be received and transmitted simultaneously in the same frequency-band, which will significantly impact the secrecy performance.

Secrecy beyond Encryption: Obfuscating Transmission Signatures in Wireless Communications

Hanif Rahbari and Marwan Krunz

The authors are with the University of Arizona.

ABSTRACT

The privacy of a wireless user and the operation of a wireless network can be threatened by the leakage of side-channel information (SCI), even when encryption and authentication are employed. In this article, we describe various passive (traffic analysis) and active (jamming) attacks that are facilitated by SCI. Our goal is to highlight the need for novel PHY-layer security techniques that can be used to complement classical encryption methods. We discuss several of these techniques along with advanced hardware that exhibits promising capabilities for countering privacy and SCI-related attacks.

INTRODUCTION

In 1960, the British Secret Intelligence Service (MI6) was under pressure to break a cipher related to the French position on the issue of Britain's membership in the European Economic Community. The officers were unable to break the code. However, Peter Wright, an MI6 scientist, noticed that the intercepted encrypted telex coming out of the French embassy in London carried a faint secondary signal. This signal turned out to be an electromagnetic "echo" of the plaintext message that was being entered to the cipher machine. Wright exploited this signal to disclose the content of the ciphertext without having to break the code. Decades later, features of communicated traffic in the form of echoes or footprints of encrypted messages have been widely used to disclose clues about the traffic content (e.g., the spoken language in an encrypted VoIP session).

In computer networks, a given layer in the protocol stack is secured independent of other layers. Encryption is the common way to provide message confidentiality. For example, at the application layer, encryption algorithms and protocols, such as HTTPS and SSH, provide message confidentiality. At the transport and network layers, the corresponding headers and payloads are encrypted using protocols such as TLS and IPSec. At the data link (medium access control, MAC) layer, WPA2 is used for 802.11 frames. Third generation/Universal Mobile Telecommunications System (3G/UMTS) and 4G Long Term Evolution (LTE) technologies for wireless communications also provide message confidentiality through encryption.

However, even when the payload of any protocol data unit (PDU) is encrypted, various transmission features such as the packet size and inter-packet times can still be determined by eavesdropping on the physical (PHY) layer frame. Wireless traffic is particularly vulnerable to eavesdropping because of the broadcast nature of wireless communications. For example, the 128-bit AES block cipher used in WPA2 preserves the size of the plaintext and does not impact PHY-layer parameters, such as the modulation scheme. At the transmitter (Tx) side, the PHY layer is responsible for receiving the (encrypted) payload from the MAC layer, prepending the required unencrypted PHY header and preamble, converting the entire frame to an analog signal, and then transmitting it over the air. Furthermore, user privacy can be threatened by the exposure of unencrypted header fields. In many wireless security standards such as WPA2 (802.11i), MAC and PHY headers are not encrypted (Fig. 1). Therefore, various transmission features remain visible to eavesdroppers. These features include the received signal strength (RSS), modulation scheme, traffic direction (uplink/downlink), and traffic statistics (e.g., frame size, inter-frame time, data rate). Collectively, these features are referred to as side-channel information (SCI).

In this article, we explain how adversaries can exploit SCI of encrypted wireless traffic to launch various attacks against user privacy and functionality of a practical wireless network. We then discuss some PHY-layer solutions that have been proposed to counter such attacks and complement conventional message encryption at upper layers.

SCI-BASED ATTACKS IN WIRELESS NETWORKS

In this section, we present two types of SCI-enabled attacks: passive and active attacks. Passive attacks refer to SCI analysis performed by

0163-6804/15/$25.00 © 2015 IEEE

an eavesdropper (Eve) to disclose some private information about a user. Active attacks refer to selective jamming of specific packets or parts of a packet, where "significance" is determined based on leaked SCI. The mechanisms for acquiring and analyzing SCI are discussed below.

Passive (Privacy) Attacks

The privacy of a wireless user can be violated by overhearing and analyzing encrypted traffic at the PHY layer. We categorize the types of leaked information and privacy violation into two groups.

Device Identification and User Tracking: An eavesdropper can fingerprint a wireless device or its user by exploiting device identifiers embedded in unencrypted headers, the device's intrinsic signature, or captured SCI. Using a device's fingerprint, the adversary can easily track the user's geographical location or determine his online activity. For example, *Snoopy* is a software program that can be deployed on a low-altitude flying drone to track users based on their fingerprints, steal their confidential information, or launch a man-in-the-middle attack by spoofing already trusted access points. It does not require a visual sensor; instead, it uses an antenna to observe WiFi encrypted communications.

Background activities of installed apps on a smartphone/tablet or the specific implementation of its wireless card driver can be used to construct a fingerprint. For instance, Eve can create a device-specific traffic fingerprint by analyzing the SCI of software programs running for six hours in the background of a 3G smartphone [1]. This is because more than 70 percent of a smartphone's traffic is independent of user interactions and depends only on installed apps. In fact, by monitoring 15 minutes' worth of traffic, it is possible to identify a particular device with 90 percent success rate among 20 devices running different sets of apps [1]. Similarly, traffic statistics can characterize an 802.11 device with high probability. Apart from apps/user-generated traffic, different vendors often have different protocol implementations and data rate distributions on their wireless cards that result in vendor-specific inter-frame times, medium access wait (backoff) times, and transmission times [2]. Together, these parameters constitute a vendor-specific fingerprint of the device.

Besides traffic statistics, hardware-specific and electromagnetic characteristics of an RF emitter form a "radiometric" identity of a particular Tx. The analog components of a wireless card's transmit path (e.g., oscillator, baseband filter, amplifier, and antenna) exhibit inherent manufacturing impairments that differ from one card to another. Small variations in these components create distinct artifacts in the emitted signal (e.g., frequency offset and amplitude clipping). The distortions in the captured modulation symbols due to hardware impairments can be exploited to detect a signal's originating device [3].

User's Activities and Browsing Interests: An eavesdropper can also exploit SCI to discern the online activities of a user, his/her interests, or his/her search queries. For example, through captured SCI, Eve can identify not only the website a user is browsing, but also the currently active page within a specific website. A typical website is characterized by a nominal uplink/downlink traffic volume and duration. These coarse-grained traffic features are sufficient to classify websites [4]. Even within a given website in which different pages are designed for different users, analyzing the packets size distribution allows a specific page to be identified. As a result, the attacker may be able to conclude the user's product of interest and may overwhelm him/her with many commercial ads.

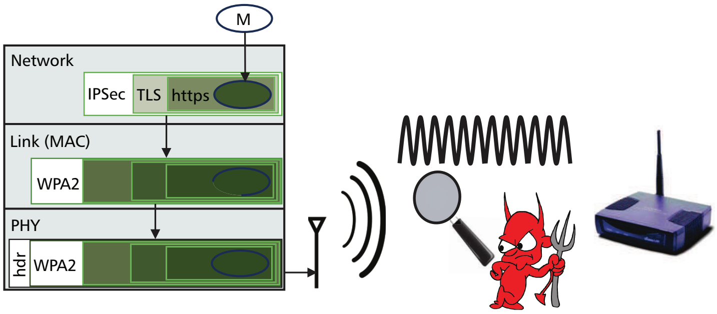


Figure 1. Encryption of a message (shown in shaded area) at different layers of the protocol stack. An upper-layer packet is considered the payload for the next lower layer.

The leakage of private information is not limited to online browsing. An adversary can determine with 80 percent accuracy the type of user activity (gaming, video streaming, Skype, browsing, etc.) by only eavesdropping for 5 s on that user's WiFi traffic [5]. Differences between the traffic statistics of different applications are often large enough to distinguish these applications. Furthermore, the adversary can find out the user's specific actions during an activity, such as posting a status on Facebook or opening a chat window in Gmail, based on the statistics of the sequence of packets generated by the user. Along the same lines, tracking the traffic of two users can reveal if they are communicating with each other.

The sizes (in bytes) and directionality (uplink/downlink) of a sequence of packets exchanged between a mobile user and an access point can also reveal what the user is searching for. Google, Bing, and other search engines provide users with suggestions for a searched phrase (i.e., auto-suggestion feature). When a user types the first letter of a keyword, the search engine quickly responds with a list of suggested words. Typing the second letter updates the list of suggestions, and so on. The size of the packet that contains the list of suggestions is highly correlated with the typed letters [6]. Eve can construct a table of different keywords and associate them with the sizes of per-keystroke suggested lists. She can then match the sizes of an observed sequence of packets to one of the entries in the table and determine the queried word [6]. Even the message length and the language used in an encrypted instant messaging application can be determined based on packet sizes only.

Unencrypted PHY and MAC headers are the main sources for extracting important traffic parameters. At the frame level, the PHY header contains the packet size/duration and transmission rate/modulation scheme fields. Parameters such as source and destination MAC addresses, direction of the packet, and packet type are specified in the MAC header.

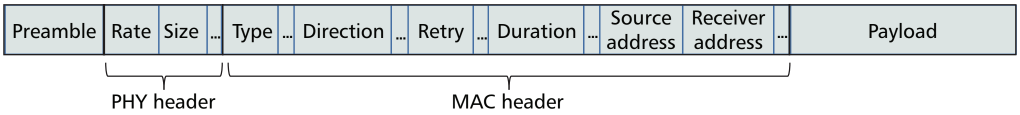


Figure 2. Typical 802.11 frame preamble, PHY header, and MAC header.

Active Attacks (Selective Jamming)

Besides breaching user privacy, SCI can be used by malicious users to disrupt wireless communications by selectively jamming transmissions and preventing correct decoding at the receiver (Rx). Jamming includes random attacks, persistent attacks (barrage jamming), and smart/selective attacks in which only certain packets or parts of a particular packet are jammed. In selective (reactive) jamming, a targeted packet (or part of it) is selected based on the amount of disruption caused by not delivering this packet to its intended Rx. For example, TCP acknowledgment (ACK) packets are much shorter in duration than TCP data packets, but are critical for maintaining high TCP throughout by preventing a significant reduction in the congestion window size. Jamming these packets requires less energy than jamming a data packet. At the same time, it can deceive the TCP sender into thinking that the last data packet was not successfully received due to network congestion. Consequently, the sender may unnecessarily reduce its packet transmission rate and retransmit the last packet, which was already received correctly. The attacker can identify the TCP ACK by analyzing the sequence of inter-arrival times and packet sizes. In the case of link-layer ACK packets, the packet type can also be identified by inspecting the unencrypted MAC header (Fig. 2).

Unencrypted PHY-layer header fields (Fig. 2) can be intercepted and used to detect and jam data packets transmitted at high rates. Wireless devices adapt their transmission rates based on channel conditions. A good channel prompts the Tx to use a higher-order modulation scheme, and hence a high data rate. When a packet is not successfully received, the Tx attributes that to channel conditions and accordingly retransmits the packet at a lower rate. This can be exploited by the attacker to jam only high-data-rate packets, prompting the Tx to reduce its rate and waste communication resources.

In addition to the PHY header, the modulation scheme of the frame payload may disclose the data rate. The frame preamble can also be exploited to detect the arrival of a packet and launch reactive attacks. This preamble is a publicly known signal, prepended to the beginning of a frame to help the Rx detect the frame and estimate various communication parameters (e.g., frequency offset, channel response). Correct decoding of a frame depends on correct estimation of these parameters. Once a frame is detected, an attacker can jam a vulnerable part of the preamble to disrupt the parameter estimation functions at the Rx [7].

SCI Extraction and Analysis Methods

Extraction of Traffic Attributes

Unencrypted PHY and MAC headers are the main sources for extracting important traffic parameters. At the frame level, the PHY header contains the packet size/duration and transmission rate/modulation scheme fields. Parameters such as source and destination MAC addresses, direction of the packet, and packet type (e.g., a retransmission) are specified in the MAC header (Fig. 2).

In addition to header fields, SCI can be used to extract certain packet parameters and radiometric features. For example, the modulation scheme used for the frame payload reveals the packet size and data rate. In digital communications, a bit sequence is modulated into symbols before transmission over the air. The number of possible symbols of a modulation scheme (known as modulation order) relates to the number of bits that can be represented by a single symbol. Because of channel noise, symbols must be sufficiently separated so that the Rx can distinguish them from one another. Consequently, a noisier channel can support fewer bits per symbol. Hence, the transmission of a fixed-size payload can take different durations under different channel conditions. By measuring the frame duration (in seconds) and detecting the modulation scheme, an adversary can estimate the payload size (in bytes).

Flow-level parameters and statistical distributions can be calculated based on packet-level parameters. These include (but not limited to) the total traffic volume and the number of unique packets. To reduce the effect of packet collisions on traffic statistics, retransmitted packets are identified and removed from the observation set. When normalized to the total session duration, the traffic volume may provide a reliable data rate statistic, despite variations in the instantaneous data rate.

Analysis of Traffic Attributes

Although acquiring traffic attributes is often sufficient to launch selective active attacks, classification-based passive attacks require further traffic analysis. Supervised machine learning is the main approach used for classification. In this approach, features of known traffic types (e.g., specific websites or activities) are used to train a classifier. These features include frame size, traffic direction, inter-packet times, and so on (Fig. 3).

Different classifiers can be employed to identify the class of a features set. These include (multinomial) naive Bayes, support vector machine (SVM), *k*-nearest neighbor algorithm, decision tree, neural networks, and hidden Markov models (HMMs). Samples belonging to different known classes are used to train the classifier. Once the attributes of unknown traffic have been extracted, the classifier tries to find the most similar traffic type to the observed features. SVM and HMM usually provide better classification accuracy than others [5].

Prevention and Remedies

Several defense mechanisms have been proposed to prevent SCI-enabled passive and active attacks. Some countermeasures obfuscate SCI to distort traffic statistics. Others employ a PHY-layer-specific approach, whereby potential eavesdroppers are deafened through friendly jamming (FJ) to prevent correct decoding of unencrypted fields. SCI obfuscation through PHY-level cryptography has also been considered. In this section, we first discuss the limitations of PHY-layer encryption and then present other solutions. Note that although spreading techniques, often used to combat narrowband and pulse jamming attacks, can be used against non-selective and random jamming, other preventive approaches are needed to counter SCI-enabled selective attacks. Moreover, spread spectrum techniques used in common wireless standards (e.g., 802.11b/g) rely on known spreading patterns, and hence cannot prevent the leakage of SCI.

Limitations of Header Encryption

Given that a significant amount of SCI is leaked through PHY/MAC headers, a natural question to ask is "Why not encrypt these headers?" However, this is usually not a viable option for the reasons stated below.

Transmitter Authentication: Encryption is based on a shared secret key. In a network of nodes, different pairs of nodes establish distinct keys for different sessions during the association process at the MAC layer. Session participants are identified by globally unique MAC addresses. Each node maintains a table of session keys that are associated with the MAC addresses of the participants of each session. This means that before decoding the MAC addresses in an incoming frame, a node does not know the sender and intended receiver of that frame; hence, it cannot immediately look up the corresponding decryption key.

Instead of the MAC address, the Tx-Rx channel or other radiometric features may be used as a PHY-level identifier to look up the key. However, mobility and inaccuracy of low-end RF receivers limit the applicability of these identifiers [8].

Broadcast Operation: Certain fields in the header are to be broadcast to every node in the vicinity of the transmitting node (e.g., the "duration" field in the 802.11 MAC header). In a multi-user environment, while a user is transmitting, other users should remain silent to avoid collision. Sensing the frequency carrier before a transmission is a common collision avoidance approach, which has been adopted in 802.11 schemes. Devices may also perform virtual carrier sense by overhearing the duration field and updating their network allocation vectors (NAVs) accordingly. Thus, if the MAC header is encrypted, other users cannot overhear the duration field.

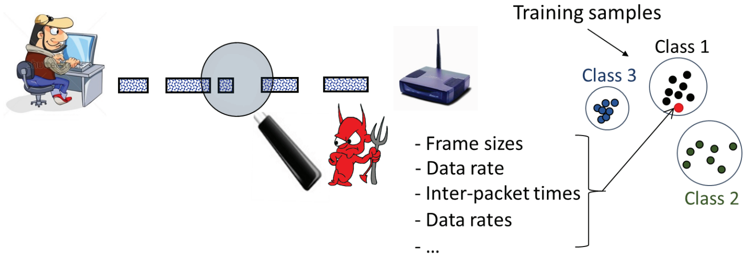


Figure 3. Traffic classification based on the training samples of three known classes. The closest class to the observed features set is likely the correct class.

Delay and Complexity: The decryption process of an encrypted header incurs additional delay and complexity, especially when block ciphering is employed. Specifically, the Rx needs to set its buffer timer and initiate its demodulator according to PHY header fields. Delay in decrypting the PHY header may prevent timely operation at the Rx.

Note also that header encryption (if feasible) cannot prevent the leakage of certain SCI, such as the modulation scheme.

Obfuscation of Traffic Features

The methods used to thwart traffic analysis attacks can be divided into two subcategories, based on whether or not the attacker can be prevented from tracking the user.

Identifier Concealment: To accurately extract traffic statistics pertaining to a given device, Eve needs to filter out packets of other devices from the set of captured packets. Packets belonging to a user or to the traffic between a user and an access point (AP) are identified by their MAC addresses. With *traffic reshaping* [9], several virtual interfaces that have different MAC addresses are configured for the same device. Generated packets are dynamically divided among these interfaces to create different traffic patterns on each interface. This prevents Eve from linking the packets to the same sender and estimating the true traffic statistics.

Another way of decoupling packets that belong to the same traffic flow is to change the MAC address based on a chain of secure identifiers. In this approach, Tx and Rx agree on a sequence of bogus identifiers before they start to exchange data packets. SlyFi [10] is one such method in which the Tx and Rx true addresses are encrypted together with the elapsed time of the session to generate a set of time-rolling identifiers. A chain of hash values can also be employed to change the MAC addresses on a per-packet basis. These obfuscation methods, however, cannot conceal PHY-layer identifiers, including device fingerprints, or even packet-level parameters.

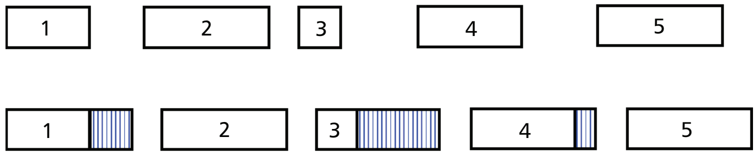


Figure 4. A simple padding scheme: The packets of the actual traffic (top) are padded to have the same, indistinguishable size and inter-arrival time (bottom).

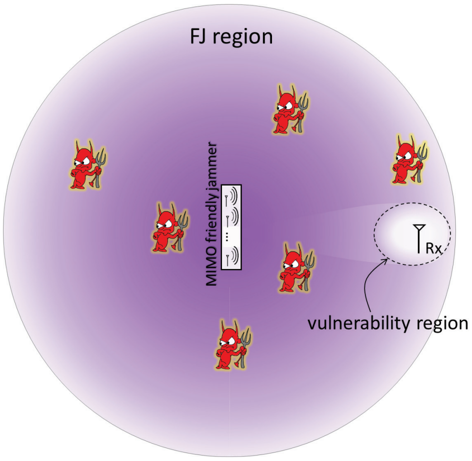


Figure 5. FJ region of a MIMO friendly jammer. The FJ signals are nullified at (and sometimes up to several wavelengths around) the legitimate Rx and along the LOS direction. In other places, eavesdroppers experience high jamming.

Padding: SCI can be distorted by appending bogus bits to packet payloads (padding) or adding dummy packets (Fig. 4). This obfuscates the inter-arrival times, packet sizes, and number of packets (hence total traffic volume). Different methods have been proposed to calculate the required amount of padding to efficiently hide the type of underlying traffic. For example, traffic morphing techniques modify the traffic pattern by altering packet sizes and inserting dummy packets. However, these techniques can be extremely inefficient, incurring up to 400 percent overhead (e.g., in defending against a website identification attack [4]).

Eavesdropper Deafening

The challenges of PHY and MAC header encryption along with the overhead and limitations of traffic feature obfuscation techniques necessitate a complementary PHY-layer approach based on friendly jamming. Jamming involves transmitting a noisy signal that interferes with the data signal, making it undecodable. It is typically considered as an adversarial act against a legitimate Rx; however, it can be employed to degrade the eavesdropper's channel without affecting the legitimate reception.

Friendly Jamming: The idea of friendly jamming (FJ) is that when two identical signals of opposite signs arrive simultaneously at the Rx, they cancel out (nullify) each other. If each friendly jammer knows the phase and amplitude of its signal at the Rx (i.e., by knowing its channel to the legitimate Rx), then several friendly jammers can cooperatively adjust their signals such that they are collectively nullified only at the Rx.. This idea can be generalized to multiple jammers or multiple antennas at a node, that is, a multiple-input multiple-output (MIMO) jammer (Fig. 5).

MIMO-capable friendly jammers can be placed in various locations with respect to the Tx and Rx. For example, when FJ is generated by the Tx of the information signal, it is called *Tx-based* FJ. Likewise, the placement of FJ close to the legitimate Rx results in *Rx-based* FJ. Using *full-duplex* radios, Rx-based FJ is possible even with a single antenna. In full-duplex communications, the self-interference caused by a node's own transmission (in this case, a jamming signal) is suppressed during the reception of the information signal. One use case of Rx-based FJ is to secure unencrypted communications of implantable medical devices (IMDs) [11]. Whenever an IMD sends a signal to the access point, the AP receives the IMD's transmission while simultaneously generating a jamming signal. Rx-based FJ can complement Tx-based FJ if the latter fails to deafen the eavesdropper, who may reside in a "vulnerability region" around the Rx. This region contains the set of locations the channel state information (CSI) of which is highly correlated with the Rx CSI. As a result, the Tx-based FJ signal is weak in this region. These points can be along the line of sight (LOS) direction and close to the Rx (Fig. 5)

As a result of superposing the FJ signal, eavesdroppers are unable to decode the PHY and MAC headers. However, in some practical scenarios, Eve may be able to estimate and remove the FJ signal from the received signal. Specifically, a MIMO-capable eavesdropper can estimate the Tx-based FJ signal by exploiting one of the known parts of the information signal (e.g., preamble) and then subtracting the estimated FJ from the received signal [12]. Furthermore, if the FJ signal is transmitted on antennas other than the antennas of the information signal, Eve can tune her antennas to receive the same FJ signal at different antennas and then subtract the received combined signal at one antenna from the received signal received at another to remove the FJ signal [13]. Even if the FJ signal is not removed at Eve, hypothesis-test cross-correlation attacks can reveal the content of a header field that takes one of a few known values [14]. The intuition is that an FJ signal that is produced via a pseudo-random noise generator averages out when it is cross-correlated with another independent signal. Thus, Eve may try to correlate the received composite signal against each possible value of the information signal and detect with high confidence the true value of the information field. In addition to capturing such header fields, Eve may extract SCI in the presence of random-noise FJ by using frame and modulation classification techniques designed for noisy channels. In the case of Rx-based FJ, Eve may use a directional antenna to suppress the FJ signal.

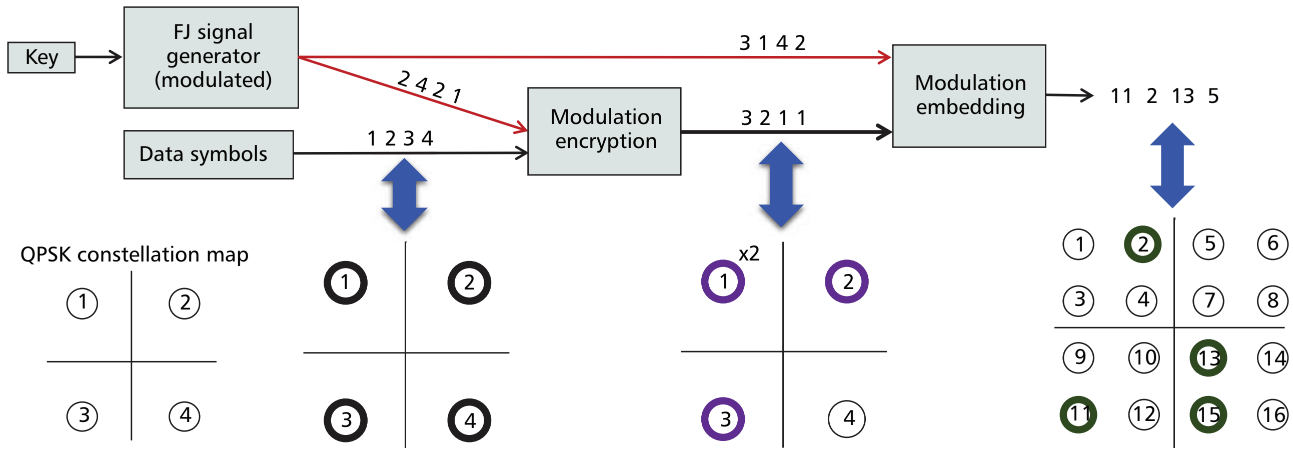


Figure 6. Example of applying FCJ on a quadrature phase shift keying (QPSK)-modulated signal. The FJ signal is divided into two parts: one for modulation encryption and another for stealthily embedding the modulated QPSK symbols into a 16-QAM modulation scheme. The numbers represent the decimal values of modulated symbols, shown with thick circles on the constellation maps.

Friendly CryptoJam: The robustness of FJ against the aforementioned attacks can be significantly improved if the FJ signal is transmitted from the same antenna as the information signal and is intermixed cryptographically with it (i.e., via stream ciphering). *Friendly CryptoJam* (FCJ) [14] does that by using a secret modulated FJ signal to encrypt the modulated headers of the frame. In contrast to classic FJ, this signal is known to the Rx and is not a function of the Tx-Rx channel. Furthermore, FCJ obfuscates the payload's modulation scheme (and hence packet size and data rate) by hiding it in the highest-order modulation scheme.

FCJ combines the jamming signal with the modulated data signal before transmission. Instead of nullifying the FJ signal at the Rx, in FCJ a secret sequence of modulated symbols is generated at both the Tx and Rx, based on a shared secret key. This signal is used for two purposes. First, it encrypts the modulated symbols by securely relocating each modulated symbol within the same constellation map (Fig. 6). Rx uses the same sequence to recover the original modulated symbols from the received encrypted symbols. Eve cannot decrypt the header because she cannot generate the same FCJ signal without knowing the secret key. Second, the FCJ signal is used to embed the payload's symbols (which have been modulated with any one of several available modulation schemes) into the constellation map of the highest-order modulation scheme. Different FCJ symbols map the same encrypted data symbol into different locations on the constellation map of the highest-order modulation scheme. In the example in Fig. 6, encrypted symbol 1 can be mapped to 16-quadrature amplitude modulation (QAM) symbols 13 and 5 when the corresponding FCJ symbols are 4 and 2, respectively. This embedding tries to preserve the original "distances" between symbols on the constellation map so as to maintain the same performance of the original modulation scheme. From Eve's standpoint, the frame payload will always seem to have been generated using one modulation scheme (16-QAM in the example). As a result, Eve cannot identify the true modulation scheme.

The generation of the FJ signal is location-independent and is robust to node mobility, time synchronization errors, and packet losses. This is achieved by generating this signal on a per-frame basis, using a per-frame identifier that is embedded in the preamble (without disrupting normal preamble functions). This PHY-layer identifier can also be used for sender identification in place of the encrypted Tx MAC address. Changing the jamming signal on a per-frame basis also prevents Eve from detecting retransmitted frames.

Conclusions and Future Directions

SCI leaked from encrypted wireless communications can be exploited to violate user privacy using various traffic analysis techniques. Moreover, software-defined radios (SDRs) have been used to experimentally demonstrate SCI-enabled reactive jamming attacks. In particular, it has recently been shown that field programmable gate array (FPGA) implementation of a reactive jammer can achieve extremely fast reaction time [15]. On the other hand, despite recent standardization activities for MAC address randomization in 802.11 systems [16], emerging (multi-user, MU-)MIMO and 5G systems and upcoming 802.11 standards, such as 802.11ai/aq/ax, have not yet incorporated PHY-layer security in their designs. This leaves the header fields, modulation schemes, spreading patterns, and certain types of management frames exposed to eavesdroppers. To prevent the leakage of SCI, PHY-layer technologies (including full-duplex MIMO Tx/Rx) and joint design of FJ and cryptography (e.g., FCJ) have been proposed. SDRs have also been used to implement Rx/Tx-based FJ schemes, and to demonstrate the limitations of FJ. However, so far these techniques cannot completely prevent the leakage of SCI without incurring high overhead, which calls for further research in this area.

SDRs have been used to implement Rx/Tx-based FJ schemes, and to demonstrate the limitations of FJ. However, so far these techniques cannot completely prevent the leakage of SCI without incurring high overhead, which calls for further research in this area.

Acknowledgments

This research was supported in part by the Army Research Office (grant # W911NF-13-1-0302) and in part by the National Science Foundation (grants # IIP-1265960 and CNS-1409172). Any opinions, findings, conclusions, or recommendations expressed in this article are those of the author(s) and do not necessarily reflect the views of ARO or NSF.

References

[1] T. Stöber *et al.*, "Who Do You Sync You Are? Smartphone Fingerprinting via Application Behaviour," *Proc. Sixth ACM Conf. Security and Privacy in Wireless andMobile Networks*, Budapest, Hungary, 2013, pp. 7–12.

[2] C. Neumann, O. Heen, and S. Onno, "An Empirical Study of Passive 802.11 Device Fingerprinting," *Proc. 32nd IEEE Int'l. Conf. Distributed Computing Sys. Wksps.*, June 2012, pp. 593–602.

[3] V. Brik *et al.*, "Wireless Device Identification with Radiometric Signatures," *Proc. ACM MobiCom'08*, San Francisco, CA, 2008, pp. 116–27.

[4] K. Dyer *et al.*, "Peek-a-Boo, I Still See You: Why Efficient Traffic Analysis Countermeasures Fail," *Proc. IEEE Symp. Security and Privacy*, May 2012, pp. 332–46.

[5] F. Zhang *et al.*, "Inferring Users' Online Activities through Traffic Analysis," *Proc. 4th ACM Conf. Wireless Network Security*, Hamburg, Germany, 2011, pp. 59–70.

[6] S. Chen *et al.*, "Side-Channel Leaks in Web Applications: A Reality Today, a Challenge Tomorrow," *Proc. IEEE Symp. Security and Privacy*, May 2010, pp. 191–206.

[7] H. Rahbari, M. Krunz, and L. Lazos, "Swift Jamming Attack on Frequency Offset Estimation: The Achilles Heel of OFDM Systems," *IEEE Trans. Mobile Computing*, vol. PP, no. 99, pp.1–1, DOI: 10.1109/TMC.2015.2456916.

[8] S. U. Rehman, K. W. Sowerby, and C. Coghill, "Analysis of Impersonation Attacks on Systems Using Rf Fingerprinting and Low-End Receivers," *J. Computer and System Sciences*, vol. 80, no. 3, 2014, Special Issue on Wireless Network Intrusion, pp. 591–601.

[9] F. Zhang *et al.*, "Thwarting Wi-Fi Side-Channel Analysis through Traffic Demultiplexing," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, Jan. 2014, pp. 86–98.

[10] B. Greenstein *et al.*, "Improving Wireless Privacy with an Identifier-Free Link Layer Protocol," *Proc. 6th Int'l. Conf. Mobile Sys., Appl., and Services*, Breckenridge, CO, 2008, pp. 40–53.

[11] S. Gollakota *et al.*, "They Can Hear Your Heartbeats: Non-Invasive Security for Implantable Medical Devices," *Proc. ACM SIGCOMM 2011 Conf.*, Toronto, Ontario, Canada, Aug. 2011, pp. 2–13.

[12] M. Schulz, A. Loch, and M. Hollick, "Practical Known-Plaintext Attacks against Physical Layer Security in Wireless MIMO Systems," *Proc. Network and Distrib. Sys. Security Symp.*, Feb. 2014.

[13] N. Tippenhauer *et al.*, "On Limitations of Friendly Jamming for Confidentiality," *Proc. IEEE Symp. Security and Privacy*, May 2013, pp. 160–73.

[14] H. Rahbari and M. Krunz, "Friendly CryptoJam: A Mechanism for Securing Physical-Layer Attributes," *Proc. ACM Conf. Security and Privacy in Wireless and Mobile Networks*, Oxford, United Kingdom, July 2014, pp. 129–40.

[15] D. Nguyen *et al.*, "A Real-Time and Protocol-Aware Reactive Jamming Framework Built on Software-Defined Radios," *Proc. 2014 ACM Wksp. Software Radio Implementation Forum*, Chicago, IL, 2014, pp. 15–22.

[16] "IEEE Group Recommends Random MAC Addresses for Wi-Fi Security," http://goo.gl/m6rOCE.

Biographies

Hanif Rahbari (rahbari@email.arizona.edu) is currently an electrical and computer engineering Ph.D. candidate at the University of Arizona. He received his B.Sc. in information technology from Sharif University of Technology and his M.Sc. in computer networks from AmirKabir University of Technology, Tehran, Iran. His research interests include wireless communications and networking, PHY-layer security, hardware implementation, dynamic spectrum access networks, and multimedia networking.

Marwan Krunz (krunz@email.arizona.edu) [S'93, M'95, SM'04, F'10] received his Ph.D. degree in electrical engineering from Michigan State University in 1995. He is the Kenneth VonBehren Endowed Professor of electrical and computer engineering and the site co-director of the NSF Broadband Wireless Access and Applications Center. His research interests are in wireless communications and networking, with emphasis on resource management, adaptive protocols, and security issues. He has published more than 225 journal articles and peer-reviewed conference papers. He received numerous awards, including the 2012 IEEE TCCC Outstanding Service Award and the NSF CAREER Award. He was an Arizona Engineering Faculty Fellow (2011–2014) and an IEEE Communications Society Distinguished lecturer (2013 and 2014). He has served on the editorial boards of several IEEE journals. He has been General and Program Chair for numerous conferences, including INFOCOM '04, SECON '05, WoWMoM '06, and WiSec '12.

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.

Guest Editorial

Vehicular Networking for Autonomous Driving

Alexey Vinel

Henrik Pettersson

Lan Lin

Onur Altintas

Oleg Gusikhin

The research area of vehicle-to-vehicle and vehicle-to-infrastructure (V2X) networking and respective cooperative driving applications has been growing in recent decades. Now it is clear that V2X wireless technology will be a communication baseline for many promising cooperative automotive applications, which will make driving safer, more energy-efficient, and more comfortable. Autonomous driving is the next step, which is considered a strategic direction by many vehicle manufacturers. Although there is still a long way before fully autonomous vehicles are introduced massively in ubiquitous city environments, it is already practically feasible to consider fully automatic operation of vehicles in restricted areas (harbors, parking lots, dedicated public transport lanes). Autonomous cooperative driving enabled by V2X communications has highly demanding operating conditions and generates delay-sensitive data traffic with requirements for high reliability

To discuss the above aspects of V2X vehicular networking, this Feature Topic comprises four original articles with contributions to some aspects of highly automated and fully autonomous vehicles.

The first article, "Enhancements of V2X Communication in Support of Cooperative Autonomous Driving" by L. Hobert *et al.*, analyzes cooperative sensing and cooperative maneuvering as two key functions that can greatly enhance autonomous driving. Their communication system relies on the current set of communication standards for WiFi-based V2X communication, more specifically on the European variant of communication protocols based on ITS-G5. The article proposes adaptations and extensions for V2X-specific messages and networking. It provides directions for starting standardization activities on communication support for cooperative†autonomous driving. The extensions of protocols and standards allow for a gradual deployment of the next generation of V2X communication and enable an increasing level of vehicle automation.

The second article is "Reliable and Efficient Autonomous Driving: the Need for Heterogeneous Vehicular Networks" by K. Zheng *et al.* The authors propose the concept of heterogeneous vehicular networks (HetVNETs) together with an improved protocol stack and new messages in order to satisfy the various communication requirements of autonomous driving under certain highway and urban intersection scenarios. In particular, the development of HetVNETs will realize reliable and highly efficient V2X communications that are essential to making commercial autonomous driving vehicles a reality on the road before 2020 toward the 5G era.

Support for seamless connectivity in highly mobile environments is now a main requirement to enable the development of cooperative automotive applications. However, in these environments, traditional handover mechanisms are inadequate, and proactive mechanisms must be considered. The third article, "Enabling Seamless V2I Communications: Toward Developing Cooperative Automotive Applications in VANET Systems" by A. Ghosh *et al.*, uses two concepts from the Y-Comm framework to show how a new probabilistic handover approach will allow the development of seamless handover mechanisms for VANET systems. In addition, this work shows how proactive handover techniques can also be used to improve channel allocation, leading to enhanced network performance and better user experience. These results are being deployed on a new experimental VANET testbed.

Finally, the last article, "Cooperation Strategies for Vehicular Delay-Tolerant Networks" by J. Dias *et al.*, overviews and studies the concept of cooperation and how it may influence the performance of vehicular delay-tolerant networks (VDTNs), especially on how this architecture deals with the presence of misbehaving nodes. The authors propose two strategies, and their performance is compared and analyzed with available solutions. Conducted studies have shown the effectiveness of both approaches in improving the network performance when non-cooperative nodes are allowed on the network. They not only increase the probability of bundles reaching their final destination, but also manage to decrease the amount of wasted resources, resulting in power and energy savings, which are very important in networks with resource constraints like VDTNs.

BIOGRAPHIES

ALEXEY VINEL (alexey.vinel@hh.se) is a full professor of data communications with the School of Information Technology, Halmstad University, Sweden. He received his Bachelor's (Hons.) and Master's (Hons.) degrees in information systems from Saint-Petersburg State University of Aerospace Instrumentation, Russia, in 2003 and 2005, respectively, and his Ph.D. degrees in technology from the Institute for Information Transmission Problems, Russia, in 2007 and Tampere University of Technology, Finland, in 2013. His research area is vehicular networking.

HENRIK PETTERSSON is a senior engineer at Scania CV, Sodertalje. He received his Master's degree in mechanical engineering from Linkoping University, Sweden, in 1995 and a Ph.D. degree in technology from the Division of Fluid and Mechanical Engineering Systems at Linkoping University in 2002. Since 2002 he has been working on control strategies for commercial vehicles, and is currently involved in research projects on control and coordination of heavy vehicle platoons and cooperative driving.

LAN LIN joined Hitachi in Sophia Antipolis, France, in 2005. She is a senior researcher of Hitachi Europe, Center of Social Innovation, Automotive & Industry Lab, actively involved in R&D activities on intelligent transport systems, big data, and digital manufacturing. She is currently Chair of ETSI Technical Committee ITS WG1 (applications) and Rapporteur of several standards Work Items, as well as Vice-Chair of the European Car to Car Communication Consortium (C2C-CC) application working group.

ONUR ALTINTAS is a Fellow at the R&D Group of Toyota InfoTechnology Center, Co. Ltd, in Tokyo. He has been co-founder and General Co-Chair of the IEEE Vehicular Networking Conference since 2009. He serves as an Associate Editor for *IEEE ITS Magazine* and is on the Editorial Board of the Connected Vehicles Series of *IEEE Transactions on Vehicular Technology*. He is an IEEE VTS Distinguished Lecturer.

OLEG GUSIKHIN is a technical leader at Ford Research and Advanced Engineering. For over 20 years, he has been working at Ford Motor Company in different functional areas including information technology, advanced electronics manufacturing, and research and advanced engineering. During his tenure at Ford, he has been involved in the design and implementation of advanced information technology and intelligent controls for manufacturing and vehicle systems.

Enhancements of V2X Communication in Support of Cooperative Autonomous Driving

Laurens Hobert, Andreas Festag, Ignacio Llatser, Luciano Altomare, Filippo Visintainer, and Andras Kovacs

Laurens Hobert is with Hitachi Europe.

Andreas Festag and Ignacio Llatser are with Technische Universität Dresden.

Luciano Altomare and Filippo Visintainer are with Centro Ricerche Fiat.

Andras Kovacs is with Broadbit.

ABSTRACT

Two emerging technologies in the automotive domain are autonomous vehicles and V2X communication. Even though these technologies are usually considered separately, their combination enables two key cooperative features: sensing and maneuvering. Cooperative sensing allows vehicles to exchange information gathered from local sensors. Cooperative maneuvering permits inter-vehicle coordination of maneuvers. These features enable the creation of cooperative autonomous vehicles, which may greatly improve traffic safety, efficiency, and driver comfort. The first generation V2X communication systems with the corresponding standards, such as Release 1 from ETSI, have been designed mainly for driver warning applications in the context of road safety and traffic efficiency, and do not target use cases for autonomous driving. This article presents the design of core functionalities for cooperative autonomous driving and addresses the required evolution of communication standards in order to support a selected number of autonomous driving use cases. The article describes the targeted use cases, identifies their communication requirements, and analyzes the current V2X communication standards from ETSI for missing features. The result is a set of specifications for the amendment and extension of the standards in support of cooperative autonomous driving.

INTRODUCTION

In the last years, there has been tremendous interest in the development of vehicles capable of driving autonomously, from both the research community and industry. Autonomous vehicles promise highly increased traffic safety and fuel efficiency, better use of the infrastructure, and the liberation of drivers to perform other tasks. For these reasons, autonomous driving may create a paradigm shift in the way people and goods are transported.

Most autonomous vehicles currently in development are based on a perception subsystem consisting of onboard sensors, which build a map of the vehicle's environment, and a control subsystem that governs the longitudinal and lateral motion of the vehicle [1–3]. Even though this approach has already been demonstrated in field tests, it presents some drawbacks: first, the limited perception range of onboard sensors only allows for detecting adjacent vehicles; and second, the vehicles are unable to cooperate in order to efficiently perform maneuvers with a high complexity.

These limitations may be overcome by means of vehicle-to-vehicle/infrastructure (V2X) communication, which enables two key features in autonomous vehicles: *cooperative sensing* increases the sensing range by means of the mutual exchange of sensed data, and *cooperative maneuvering* enables a group of autonomous vehicles to drive coordinatedly according to a common centralized or decentralized decision-making strategy. The integration of onboard sensors and V2X communication also results in a solution that is more cost-effective than an approach based on high-quality sensors only.

The application of V2X communication to autonomous driving has been a research topic for many years, such as in the pioneering implementations of the PROMETHEUS initiative in Europe and the PATH Automated Highway System in the United States. More recently, several research activities [4, 5] and successful field trials of V2X communication for safety and traffic efficiency [6] have triggered manifold ongoing activities to bring V2X communication for autonomous driving closer to reality. Cooperative autonomous driving is currently being further developed by the European R&D projects AutoNet2030 [7], i-GAME,[1] AdaptIVe,[2] and COMPANION.[3]

We regard V2X communication in support of autonomous driving as a natural evolution of the communication system for cooperative vehi-

[1] http://www.gcdc.net/i-game

[2] http://www.adaptive-ip.eu

[3] http://www.companion-project.eu

0163-6804/15/$25.00 © 2015 IEEE

cles. The latter, here referred to as first generation V2X communication systems (1G-V2X), has been designed to provide driver assistance, which corresponds to level 1 in the definition of automation levels in SAE J3016 [8]. Higher levels of automation introduce new requirements that are not covered by 1G-V2X; therefore, the definition of new or enhanced messages, communication protocols, and their standardization is needed for cooperative autonomous driving.

The next section outlines some important use cases of autonomous driving where V2X communication plays a key role. The main V2X requirements for the implementation of the considered use cases are then identified, and an overview of the state-of-the-art V2X standards in Europe is given following that. Based on the presented requirements and standards, the message extensions required to support autonomous driving use cases are then explained. Finally, we conclude the article.

Autonomous Driving Use Cases

Use cases for autonomous driving can be grouped in three categories: close-distance, urban, and freeway use cases. Whereas close-distance use cases typically cover autonomous vehicles with the lowest operating velocities — an example is a vehicle able to park autonomously — urban and freeway use cases focus on common traffic situations. The latter two categories have the highest potential to improve traffic safety and efficiency. For this reason, we present the following four urban and freeway use cases for autonomous driving:

Convoy Driving

One of the autonomous driving applications that has gained strong attention from research and industry in recent decades is platooning. In a platoon, vehicles in the same lane are grouped together in a stable formation with small inter-vehicle distances to increase road capacity, driver safety, and comfort. A platoon typically consists of one master, usually the leading vehicle, and multiple following vehicles.

However, a platoon is not the only approach to group vehicles on freeways. In a multi-lane convoy use case, as studied in the AutoNet2030 project, a master, centralized controller, or supervisor does not exist. Instead, the vehicle control, in both lateral and longitudinal directions, is distributed over all members of the convoy (Fig. 1). The result of this approach is that vehicle disturbances, such as a braking vehicle, affect all members of the convoy to a greater or lesser extent, resulting in a stable formation.

In order to maintain small inter-vehicle distances, convoy members rely on the high-frequency exchange of up-to-date and high-quality vehicle dynamics data among vehicles in the convoy. The convoy control algorithm presented in [9] requires just the vehicle dynamics information of neighbor vehicles, instead of the information of all convoy members. As such, the algorithm scales well to large convoys and converges easily to a desired formation when vehicles join and leave the convoy.

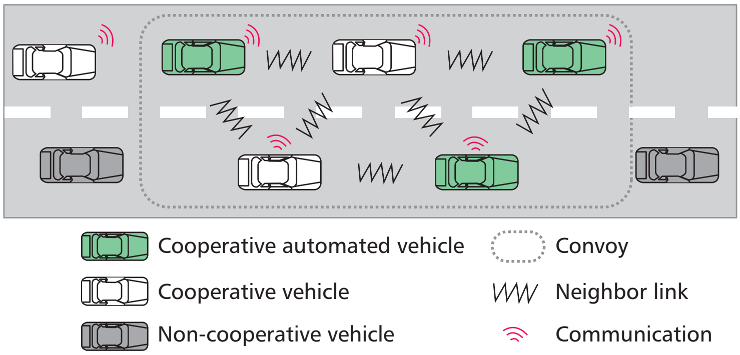


Figure 1. Exchange of vehicle dynamics data for multi-lane convoy driving.

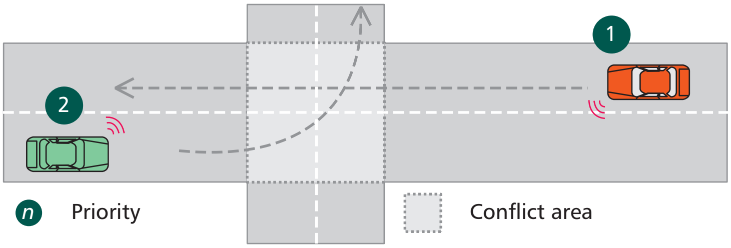


Figure 2. Priority-based coordination of incoming vehicles at an intersection with V2X communication.

Cooperative Lane Change

In the cooperative lane change use cases, cooperative vehicles (both autonomous and manually driven) collaborate to perform a lane change of one or a group of cooperative vehicles (e.g., a convoy) in a safe and efficient manner. Unlike in a traditional lane change situation, cooperative vehicles share their planned trajectories in order to negotiate and align their maneuvers.

The cooperative lane change may be aided by a roadside unit, which supports the communication among the interacting vehicles. However, when this infrastructure is not available, vehicles are forced to coordinate the lane change in an ad hoc fashion.

Cooperative Intersection Management

A cooperative intersection allows cooperative vehicles to traverse an intersection without the need for traffic lights [10]. This scenario requires a coordination mechanism in case their planned trajectories overlap.

A possible solution is shown in Fig. 2, where a roadside unit coordinates the traffic flow through the intersection by assigning relative priorities to incoming vehicles in real time. Then vehicles are able to cross the intersection efficiently following the order of their assigned priority.

Cooperative Sensing

All of the above presented use cases, as well as autonomous driving in general, depend on an adequate and reliable perception of the vehicle surroundings in order to navigate through traffic and ensure safety with a high level of automation. Broken sensors, blind spots, and low level of trust in sensor data may degrade the perfor-

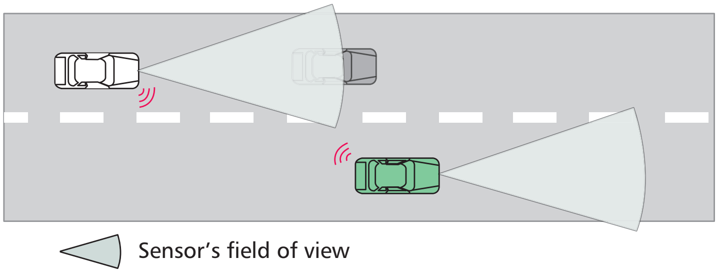


Figure 3. Exchange of detected objects for cooperative sensing.

mance or even disable automated functions of the vehicle.

In the cooperative sensing use case, shown in Fig. 3, neighbor vehicles share information gathered from local perception sensors in order to improve the quality and reliability of individual detections.

Communication Requirements

1G-V2X mainly addresses road safety and traffic efficiency for manually driven vehicles. Typical applications include obstacle warning, road works information, in-vehicle signage, traffic light phase assistance, and others [6]. The use cases for autonomous driving presented above demand new requirements. New *functional requirements* are below.

Additional Vehicle Status Data: In 1G-V2X, every vehicle broadcasts periodic safety messages to inform neighbors of its position, speed, heading, and other parameters. Autonomous vehicles need to include additional data in the periodic messages for the convoy driving and cooperative lane change use cases, such as their predicted path over the next few seconds.

Convoy Management: In 1G-V2X, a vehicle communicates with vehicles and roadside stations in its neighborhood, or located in a specific geographical region, also called a relevance area for safety information. Opposed to this "open group" concept without an explicit membership, a convoy represents a "closed group" where a vehicle needs to become a group member to participate. In order to create and maintain convoys, as well as to coordinate decentralized maneuvering negotiations, new fault-tolerant mechanisms for group management are needed.

Maneuver Negotiation: In autonomous driving, vehicles may actively need to reserve road space for lane change maneuvers. Unlike the distribution of periodic or event-driven safety messages for 1G-V2X, a reservation requires a negotiation among the involved vehicles to request and acknowledge the maneuver. This exchange enables optimal and safe trajectories for the cooperative vehicles and minimizes their collision risk.

Intersection Management: 1G-V2X is limited to the periodic broadcast of static and dynamic information of intersections, that is, to the distribution of the intersection topology and traffic light information, enabling use cases such as green light optimal speed advisory. It also allows requesting and changing the status of traffic light control systems for priority control and preemption of road traffic. With autonomous driving, communication for intersection management is extended to allow for more detailed information of the intersection geometry and to assign priorities to incoming vehicles, potentially displacing traffic lights.

Cooperative Sensing: Communication allows the exchange of locally acquired sensor data from the radar, camera, and other sensors. The captured data from the local sensors is aggregated into a list of detected objects along the road, such as obstacles, vehicles, and pedestrians, that can be exchanged with neighboring vehicles. Cooperative sensing increases the sensors' field of view to the V2X communication range and enables cooperative perception among vehicles. In 1G-V2X, the aggregation level of sensor data is much higher, and messages only carry a coarse event classifier and relevance area. Instead, the cooperative sensing use case requires the exchange of highly detailed information about the detected objects.

In addition to the functional requirements, specific qualitative *performance requirements* for cooperative autonomous driving include the following.

High Message Rate: In 1G-V2X, vehicles periodically broadcast safety messages with an interval between 100 ms and 1 s, where the rate within these limits is controlled by the dynamics of the generating vehicle and the load on the wireless channel. In contrast, the small inter-vehicle distance among autonomous vehicles requires the use of a high and fixed broadcast frequency with a timeliness guarantee on the information that autonomous vehicles possess about their neighbors. These requirements demand that autonomous vehicles have a complete and up-to-date environmental model, which allows them to coordinate maneuvers in a safe manner.

Data Load Control: The small inter-vehicle distance and corresponding high vehicle density lead to a higher data load in the network. This is even amplified by the high message rate and by additional data load for the exchange of control messages. In order to control the amount of data traffic in the network, efficient utilization of the available frequency spectrum, effective prioritization of messages by the decentralized congestion control (DCC) function, and strict control of the forwarding operations are required.

Low End-to-End Latency: The end-to-end latency is mainly composed of the delay to gather data from local sensors, the processing delay in the protocol stack, and the transmission delay over the wireless link. The end-to-end delay also includes the delay induced by the security mechanisms (generation and verification of signature and certificate, respectively) and by queuing delays in the DCC function. In 1G-V2X, the latency requirements for critical road safety applications are set to 300 ms (ETSI TS 102 539-1). In autonomous driving use cases such as convoy driving, the latency requirement is more stringent due to the smaller inter-vehicle distance between vehicles and also to ensure the string stability of large convoys.

Highly Reliable Packet Delivery: The requirement for reliable exchange of information is more critical than in 1G-V2X, since a lost or erroneous message might cause a malfunction of the vehicle control algorithms and create a safety risk.

Both functional and performance requirements impose demanding challenges on the V2X communication system. This article proposes enhancements of 1G-V2X to meet some of these challenges.

Current Standards for V2X Communication

The R&D efforts on V2X communication over the last years were accompanied by standardization efforts in the European Committee for Standardization (CEN), European Telecommunications Standards Institute (ETSI), IEEE, and International Standards Organization (ISO) in the context of cooperative intelligent transport systems (C-ITS). These activities have led to a consistent set of standards in Europe [11] and the United States [12]. We summarize the core standards for the European Release 1 defined by ETSI, which builds the basis for extensions for communication support toward autonomous vehicles, presented later in this article.[4]

The bottom layer of the reference model in Fig. 4 comprises access technologies: for V2X communication, ITS-G5[5] [EN 302 663] is the most relevant access technology in the context of this work. It has similar features as IEEE 802.11a (e.g., orthogonal frequency-division multiplexing, OFDM), but operates in the 5.9 GHz frequency band, enables a basic ad hoc mode, and disables management procedures. The medium access scheme relies on the well-known enhanced distributed channel access (EDCA) from IEEE 802.11 with carrier sense multiple access with collision avoidance (CSMA/CA) and quality of service (QoS) support. At the ITS network and transport layer, the GeoNetworking protocol (EN 302 636-4) provides single-hop and multihop packet delivery in an ad hoc network of vehicles and roadside stations. Specifically, it utilizes geographical positions carried in the packet headers for geographical addressing and forwarding of packets on the fly. On top of GeoNetworking, the Basic Transport Protocol, BTP (EN 302 636-5-1) provides a UDP-like connectionless transport protocol service.

Facilities layer standards specify application-supporting functionality: the cooperative awareness message (CAM) standard (EN 302 637-2) conveys critical vehicle state information in support of safety and traffic efficiency applications, with which receiving vehicles can track other vehicles' positions and movements. While the CAM is a periodic message sent over a single wireless hop, the decentralized environmental notification message (DENM) standard (EN 302 637-3) specifies a protocol for dissemination of event-driven safety information in a geographical region, typically via multiple wireless hops. Facility-layer messages for vehicle-to-infrastructure communication are specified in TS 103 301, including for transmission of static information about intersection topologies (MAP) and dynamic information for traffic lights. The standards at the application layer specify requirements for road hazard signaling (RHS), intersection collision risk warning (ICRW), and longitudinal collision risk warning (LCRW) (TS 101 539-1,-2,-3). RHS comprises use cases for initial deployment, including emergency vehicle approaching, hazardous location warning, and emergency electronic brake lights. ICRW and LCRW address potential vehicle collisions at intersections and rear-end/head-on collisions. Standards at the security block enable cryptographic protection by digital signatures and certificates (TS 103 097); changing pseudonyms for support of anonymity impedes tracking. Finally, management standards mainly cover support for decentralized data congestion control (TS 103 175).

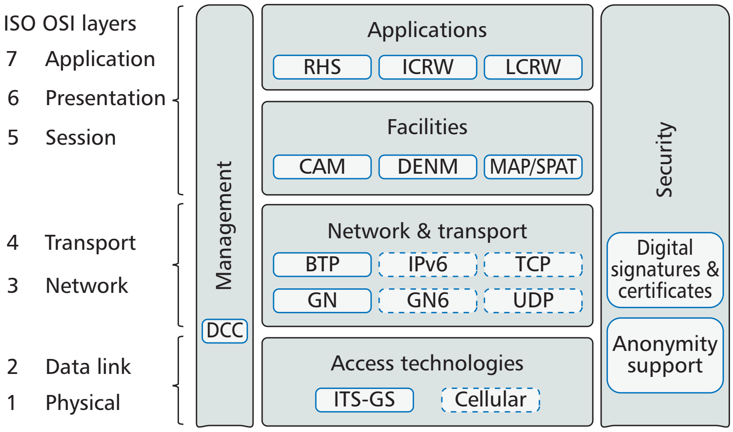


Figure 4. Reference model for 1G-V2X (functional components surrounded by solid lines are within the scope of this article).

Message Extensions for Cooperative Autonomous Driving

The specification of the European 1G-V2X system and its corresponding standards have been driven by application requirements of RHS, ICRW, and LCRW. Cooperative autonomous driving creates additional communication requirements as described above and justifies a new generation of V2X communication. Compared to 1G-V2X, the new generation still relies on ITS-G5 but modifies the upper protocol layers. We extend and amend the facilities layer to satisfy the functional and performance requirements, in particular the CAM standard (ETSI EN 302 637-2), and we introduce new facilities layer components as shown in Fig. 5. The figure also illustrates enhanced networking and transport protocols; we have already shown that the GeoNetworking protocol can be adapted to meet the network requirements for platooning use cases [13]. Also, we introduce a modification of BTP called Reliable BTP (RBTP). However, here the focus is on facility layer components, indicated by solid boxes in Fig. 5.

The vehicle state information conveyed in a CAM (ETSI EN 302 637-2) is insufficient for the convoy and cooperative intersection use

[4] Available at http://etsi.org/standards

[5] ITS-G5 can be regarded as the European variant of the former "p"-amendment to IEEE 802.11, which has been integrated into IEEE 802.11-2012.

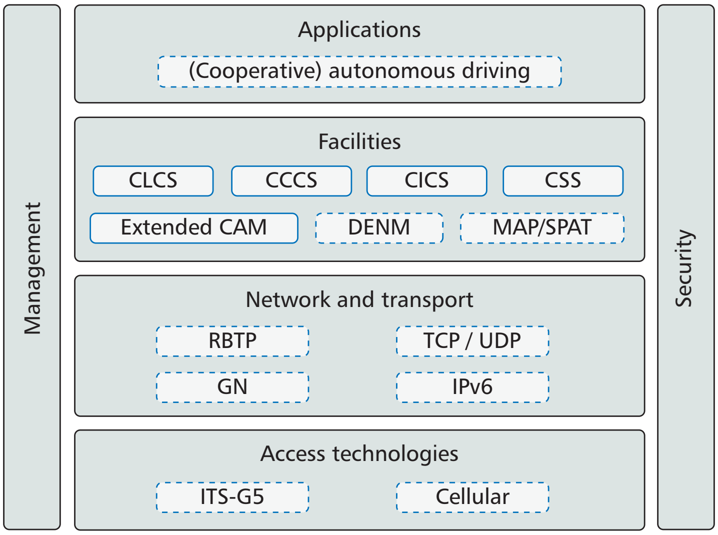


Figure 5. V2X communication architecture considered in the AutoNet2030 project.

cases. For planning maneuvers and avoiding safety-critical situations, both use cases require the exchange of periodic control-related data between neighbor vehicles, such as their predicted trajectory. This trajectory is calculated by the autonomous vehicle and cannot be measured with external sensors. Additionally, driving in a convoy requires the exchange of additional information, such as the distance to the preceding and following vehicles, target speed and acceleration, and convoy identifier.

In order to satisfy these data requirements, we propose to extend the CAM standard with additional *high* and *low* frequency containers that carry the control data specific to cooperative autonomous vehicles. The high frequency container includes only the minimum set of highly dynamic vehicle attributes for convoy driving to limit the total CAM size, including speed, heading, acceleration, and others. The low frequency container contains the less critical vehicle control data mentioned above.

In addition, two operating modes are introduced: *normal* mode and *high awareness* mode. In normal mode, CAMs are broadcast with variable frequency according to the standardized triggering conditions (i.e., between 1 and 10 Hz) depending on the vehicle dynamics. The high awareness mode augments the normal mode and increases the transmission frequency to a fixed value of 10 Hz. The newly introduced containers are only generated in high awareness mode and transmitted to single-hop neighbor vehicles using ITS-G5 on a separate service channel to relieve the heavily used control channel.

Convoy Control Communication Service

The *convoy control communication service* (CCCS) supports the exchange of information messages among cooperative vehicles in the convoy driving use case and satisfies the functional requirement for convoy management. The transmission frequency of convoy messages is dynamically adjusted depending on the convoy properties and traffic conditions. The messages exchanged among convoy vehicles via the CCCS enable each vehicle to maintain a local graph, whose nodes are the convoy members; the edges represent the dependence of the vehicle dynamics. A decentralized vehicle control algorithm performs the cooperative maneuvering, adjusting the vehicle lateral and longitudinal dynamics to keep a balanced formation and performing lane changes as required [9].

The message types offered by the CCCS to convoy members are the following.

Join/leave convoy: A join request is a single-hop broadcast message sent by an approaching vehicle, which detects a convoy and requests to become a convoy member. Similarly, a convoy vehicle that decides to abandon it (e.g., when it reaches its destination) will broadcast a leave request to inform its neighbors of its intention.

Lane change: A lane change message allows convoy vehicles to change their lane within the convoy. The message is broadcast by a convoy vehicle to inform its neighbors of a planned lane change. This way, the convoy members in the destination lane will adjust their positions to make space for the incoming vehicle.

Modify local graph: As a result of a lane change or a new vehicle entering the convoy, a vehicle may update its local graph. In this case, the new graph is broadcast to its neighbors by means of a modify local graph message. The neighbor vehicles then modify their own local graphs accordingly, thereby ensuring the consistency of the graphs among all the convoy members.

Cooperative Lane Change Service

The *cooperative lane change service* (CLCS) enables the communication for the cooperative lane change use case. CLCS supports maneuver negotiations among vehicles not belonging to the same convoy and relative space reservation by dedicated messages. The cooperative lane change is divided into three phases.

Search Phase: The planned lane change of a subject vehicle is announced in this phase, in search of a peer vehicle to start the lane change negotiation. This phase is optional and only executed when the subject vehicle has insufficient awareness of the traffic situation, and is unable to select the appropriate peer in advance. The planned lane change is described in a *lane change request* (LCR) message and is broadcast multihop around the lane change area. Any vehicle receiving the LCR will decide, based on its own planned trajectory, whether it is a suitable peer and will respond with a *lane change response*, which is unicast multihop to the subject vehicle. The subject vehicle eventually selects the most appropriate peer vehicle and informs all vehicles around the lane change area, including the selected peer vehicle, about this decision by broadcasting periodically an updated LCR message until the cooperative lane change has finished.

Preparation Phase: The selected peer vehicle opens the requested headway distance, and both vehicles adjust to the agreed speed and time of arrival. Once prepared, the peer vehicle informs the subject vehicle with a *lane change prepared* message that the next phase can start.

Execution Phase: The lane change maneuver is executed in this phase without communication support of the CLCS component. The maneuver safety is ensured by the autonomous vehicles, based on received CAMs and local sensor information.

During all cooperative lane change phases, unexpected events may occur, which require to abort the lane change. In this case, a dedicated *lane change abort* (LCA) message is exchanged between the subject and peer vehicle. The CLCS component uses a retransmission and acknowledgment mechanism in order to improve the reliable delivery of LCA messages.

Cooperative Intersection Control Service

The *cooperative intersection control service* (CICS) supports the traversing of an intersection by cooperative autonomous vehicles, that is, *intersection management* as the functional requirement. In order to allow for a collision-free and deadlock-free intersection crossing, a roadside unit acts as intersection controller to coordinate the maneuvers of the vehicles approaching the intersection [10]. The intersection controller sends on-demand messages to incoming vehicles in order to assign them priorities based on information about their current status and desired trajectories; these regulate the order in which they are allowed to cross the intersection.

The message types offered by CICS are below.

Intersection Entry Request: This unicast message is sent by approaching vehicles, which detect the presence of the intersection controller. In the intersection entry request, the vehicle specifies its desired entry and exit lanes, the predicted time to enter the intersection, and information about the vehicle dynamics.

Intersection Entry Cancellation: With this message, a vehicle is able to inform the intersection controller that it wants to cancel a previous intersection entry request, for instance, in order to send a new entry request with different parameters.

Intersection Entry Status: The calculated relative priorities by the intersection controller are broadcast to all cooperative vehicles near the intersection. With this information, the vehicles are able to maneuver cooperatively and traverse the intersection safely.

It is worth noting that CICS also supports non-cooperative vehicles crossing the intersection. Two cases can be considered: first, if a non-cooperative vehicle is driving on its own, the intersection controller communicates the assigned priority by means of traffic lights; second, if the non-cooperative vehicle belongs to a platoon led by a cooperative vehicle, all the platoon vehicles will cross the intersection according to the priority assigned to the platoon leader.

Cooperative Sensing Service

The *cooperative sensing service* (CSS) enables the sharing of detected objects, including vehicles, pedestrians, cyclists, and so on, by means of *cooperative sensing messages* (CSMs) and enables the cooperative sensing use case.

A CSM can describe up to 16 detected objects in terms of their main attributes, including position, heading, speed, acceleration, and respective confidence level. Compared to raw sensor data, such as video frames of a camera or point cloud of a lidar, object attributes are less sensor-dependent and result overall in smaller messages being transmitted.

The tendency in the design of future autonomous vehicles is to combine the data of multiple sensors in order to create more concise detections and improve the overall detection accuracy compared to individual detections. The CSS component can interface with such a sensor fusion process in two ways: as a consumer and as a producer of perception data. As a consumer, the CSS constructs new CSMs with the sensor fusion output. As a producer, the CSS component can provide the content of received CSMs and act as a virtual sensor.

The nature of many perception sensors is to measure and provide relative object attributes, such as the distance or relative speed of a detected vehicle. Even though these values are appropriate for the control of an autonomous vehicle, relative object attributes are not suitable for inter-vehicle sharing. For this reason, the CSM only contains absolute object attributes.

The CSS component constructs CSMs at a rate of 1 Hz and disseminates the message over a single wireless hop to the neighbor vehicles. In order to deal with a higher data load, CSMs are transmitted on the service channel (e.g., SCH1) rather than on the control channel on which packets are typically transmitted in 1G-V2X.

Conclusion

Autonomous driving is regarded as a major innovative step that has the potential to fundamentally transform the mobility of people and goods. Today, most developments target standalone autonomous vehicles, which are capable of sensing the surroundings and control the vehicle based on this perception, with limited or no driver intervention. The inherent drawback of this solution is the lack of coordination among vehicles and the limited range of sensors, which results in suboptimal performance. Vehicle-to-vehicle/infrastructure communication (V2X) overcomes these drawbacks by increasing the planning horizon of autonomous vehicles and enabling two key features for autonomous driving: *cooperative maneuvering* and *cooperative sensing*.

In this article, we have presented four use cases for cooperative autonomous driving, and analyzed their requirements for safe and efficient operation. Compared to the first generation of V2X communication systems (1G-V2X) and its corresponding Release 1 of communication standards, cooperative autonomous driving requires adaptations and extensions. We have presented an evolution of the V2X communication system as standardized by ETSI. In particular, we have shown how the CAM standard as the V2X core facility can be extended, and we have introduced new facilities layer components.

The proposed V2X communication system for cooperative autonomous driving uses an enhanced ITS-G5-based protocol stack. This

Vehicle-to-vehicle/infrastructure communication (V2X) increases the planning horizon of autonomous vehicles and enables two key features for autonomous driving: cooperative maneuvering and cooperative sensing.

approach allows a gradual deployment of the next generation of V2X communication for cooperative autonomous driving relying on 1G-V2X. While the introduction of 1G-V2X is expected in the next few years, AutoNet2030 and other projects contribute to the development of concepts, protocols, prototype implementations, evaluation, and standards for the next generation of V2X communication systems. AutoNet2030 in particular will focus on the analysis of quantitative performance requirements using simulations and demonstration of AutoNet2030 concepts in a prototypical implementation. All in all, these developments will demonstrate the level of automation that can be achieved by V2X communication toward the vision of fully automated driving.

AutoNet2030 in particular will focus on the analysis of quantitative performance requirements using simulations and demonstration of AutoNet2030 concepts in a prototypical implementation. All in all, these developments will demonstrate the level of automation that can be achieved by V2X communication toward the vision of fully automated driving.

Acknowledgment

This work was supported by the European Commission under AutoNet2030, a collaborative project part of the Seventh Framework Programme for research, technological development and demonstration (Grant Agreement NO. 610542). The authors would like to thank all partners within AutoNet2030, and in particular Dr. Angelos Amditis (ICCS) as project coordinator, for their cooperation and valuable contribution.

References

[1] S. Shladover, “Automated Vehicles for Highway Operations (Automated Highway Systems),” *Proc. Inst. Mechanical Engineers, Part I: J. Systems and Control Engineering*, vol. 219, 2005, pp. 53–75.

[2] C. Urmson *et al*., “Autonomous Driving in Urban Environments: Boss and the Urban Challenge,” *J. Field Robotics*, vol. 25, no. 8, 2008, pp. 425–66.

[3] J. Levinson *et al.*, “Towards Fully Autonomous Driving: Systems and Algorithms,” *Proc. IEEE Intelligent Vehicles Symp*., Baden-Baden, Germany, Jun. 2011, pp. 163–68.

[4] S. Kato *et al.*, “Vehicle Control Algorithms for Cooperative Driving with Automated Vehicles and Intervehicle Communications,” *IEEE Trans. Intelligent Transportation Systems*, vol. 3, no. 3, Sept. 2002, pp. 155–61.

[5] J. Baber *et al*., “Cooperative Autonomous Driving: Intelligent Vehicles Sharing City Roads,” *IEEE Robotics Automation Mag.*, vol. 12, no. 1, Mar. 2005, pp. 44–49.

[6] H. Stübing *et al*., “SIM-TD: A Car-to-X System Architecture for Field Operational Tests,” *IEEE Commun. Mag*., vol. 48, no. 5, 2010, pp. 148–54.

[7] A. de La Fortelle *et al*., “Network of Automated Vehicles: The AutoNet 2030 Vision,” *Proc. ITS World Congress*, Detroit, TX, Sept. 2014, http://www.autonet2030.eu.

[8] SAE, “Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems,” *Soc. Automotive Engineers*, tech. rep. SAE J3016, Jan. 2014.

[9] A. Marjovi *et al*., “Distributed Graph-Based Convoy Control for Networked Intelligent Vehicles,” *Proc. IEEE Intelligent Vehicles Symp*., Seoul, Korea, June 2015, pp. 138–43.

[10] X. Qian *et al*., “Priority-Based Coordination of Autonomous and Legacy Vehicles at Intersection,” *Proc. ITSC ’14*, Qingdao, China, Oct. 2014.

[11] A. Festag, “Cooperative Intelligent Transport Systems (C-ITS) Standards in Europe,” *IEEE Commun. Mag*., vol. 12, no. 52, Dec. 2014, pp. 166–72.

[12] J. Kenney, “Dedicated Short-Range Communications (DSRC) Standards in the United States,” *Proc. IEEE*, vol. 99, no. 7, July 2011, pp. 1162–82.

[13] I. Llatser, S. Kühlmorgen, A. Festag, and G. Fettweis, “Greedy Algorithms for Information Dissemination within Groups of Autonomous Vehicles,” *Proc. IEEE Intelligent Vehicles Symp.*, Seoul, Korea, June 2015, pp. 1322–27.

Biographies

Laurens Hobert has been a development engineer since 2010 for the Automotive and Industry Laboratory of Hitachi Europe, Sophia Antipolis, France. He received his Diploma (M.Sc.) in telematics from the University of Twente in 2012. His research interests include intelligent transport systems, wireless communication, automated driving, and software prototyping. He has been involved in various European projects, including CoVeL, DRIVE-C2X, eCo-FEV, and AutoNet2030, and actively contributes to standardization in ETSI Technical Committee ITS and the CAR-2-CAR Communication Consortium.

Andreas Festag is research group leader and lecturer at the Technical University Dresden, Vodafone Chair Mobile Communication Systems. He received a diploma degree (1996) and Ph.D. (2003) in electrical engineering from the Technical University of Berlin. His research is concerned with 5G cellular systems and vehicular communication. He actively contributes to the CAR-2-CAR Communication Consortium and ETSI Technical Committee ITS.

Ignacio Llatser received a double degree in telecommunication engineering and computer science (2008), a Master’s degree (2011), and a Ph.D. degree (2014) from the Technical University of Catalonia, Barcelona, Spain. He is currently a postdoctoral researcher at the Technical University of Dresden, Germany, in the framework of the FP7 project AutoNet2030. His research interests lie in the fields of nanoscale communication networks, vehicular networks, and autonomous driving.

Luciano Altomare is a researcher and developer at Centro Ricerche Fiat, FCA company, since 2007. He received a Master’s degree (2006) in physics from the University of Bologna, Italy. Currently his research activity is mainly focused on vehicle-to-vehicle applications and prototype integration.

Filippo Visintainer graduated in physics and is a senior researcher of Centro Ricerche FIAT, Trento Branch, part of FCA Vehicle Research and Innovation, where he has been working since 2001. Within the European Framework Programmes he has been Project Manager within several projects on intelligent mobility, preventive safety, e-Inclusion, and vehicular communication, including DRIVE C2X, TEAM IP, AutoNet2030, and AdaptIVe. Lately he has been leading an innovation project of his company on connected vehicle services.

Andras Kovacs has been working in the field of ITS for 10 years. His background is in telecommunication networks research. He has contributed to the work of the Car-2-Car Communication Consortium in the past, and currently participates in the work of ETSI ITS. His research interests relate to geonetworking, ITS applications, and their testing aspects. He is currently the director of R&D at BroadBit and the technology manager of AutoNet2030.

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

Reliable and Efficient Autonomous Driving: The Need for Heterogeneous Vehicular Networks

Kan Zheng, Qiang Zheng, Haojun Yang, Long Zhao, Lu Hou, and Periklis Chatzimisios

Kan Zheng, Qiang Zheng, Haojun Yang, Long Zhao, and Lu Hou are with Beijing University of Posts & Telecommunications

Periklis Chatzimisios is with Alexander Technological Educational Institute of Thessaloniki.

ABSTRACT

Autonomous driving technology has been regarded as a promising solution to reduce road accidents and traffic congestion, as well as to optimize the usage of fuel and lane. Reliable and highly efficient vehicle-to-vehicle and vehicle-to-infrastructure communications are essential for commercial autonomous driving vehicles to be on the road before 2020. The current article first presents the concept of heterogeneous vehicular networks (HetVNETs) for autonomous driving, in which an improved protocol stack is proposed to satisfy the communication requirements of not only safety but also non-safety services. We then consider and study in detail several typical scenarios for autonomous driving. In order to tackle the potential challenges raised by the autonomous driving vehicles in HetVNETs, new techniques from transmission to networking are proposed as potential solutions.

INTRODUCTION

The automotive industry has recently shifted from developing advanced vehicles to safe and comfortable ones, which stimulates the development of new intelligent vehicles with autonomous driving control [1]. The autonomous driving vehicle (ADV) is a multidisciplinary product that can integrate automotive control, information processing, communication capabilities, and so on. Governments and society can substantially benefit from autonomous driving, including prevention of road accidents, reduction of traffic congestion, as well as optimal usage of fuel and lane [2]. In order to realize autonomous driving, vehicles need to be capable of sensing the surrounding environment as well as performing control and path planning without any human intervention [3]. Global automakers and information technology companies, such as General Motors, Volkswagen, Toyota, and Google, expect to have ADVs on the market in 2020 and 25 percent of the vehicles out on the road to be ADVs by 2035 [4], which coincides with the timetable of fifth generation (5G) wireless communication systems.

Nevertheless, several challenges still need to be conquered for autonomous driving [5], such as:

- To have knowledge of the exact position of the vehicle and to decide how to reach the destination
- To sense the surrounding environment in order to avoid vehicle collision
- To detect the road signs as well as lanes, crosswalks, speed bumps, and so on

Currently, in order to face these challenges, sensor systems with cameras, radar, or laser range finders, and advanced autonomous driving algorithms are employed. However, it is still far from enough since the driving behavior of vehicles is significantly affected by the surrounding vehicles, and this is not well exploited due to the limited communication ability between vehicles. Moreover, the main approach to detect the surrounding environments is performed by utilizing sensor systems but is highly limited by the environment in which vehicles operate (e.g., obstacles, other vehicles, weather conditions).

Thanks to the rapid development of wireless communication technologies, vehicular networks are expected to boost the development of autonomous driving, and employ vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication techniques, which can be effectively used to detect surrounding conditions. The autonomous driving vehicle may become safer if it makes autonomous decisions with reliable information provided by vehicular networks. For example, every vehicle can periodically broadcast safety-related messages about its current condition to its neighboring vehicles, which is helpful for all vehicles in order to accurately know their surrounding environment. On the other hand, vehicular networks can significantly improve traffic efficiency. Moreover, passengers in autonomous driving vehicles are likely to enjoy infotainment content during their journey by accessing the Internet mainly via the use of V2I communication.

Due to high mobility and the dynamic change of network topology, it is difficult to provide sat-

0163-6804/15/$25.00 © 2015 IEEE

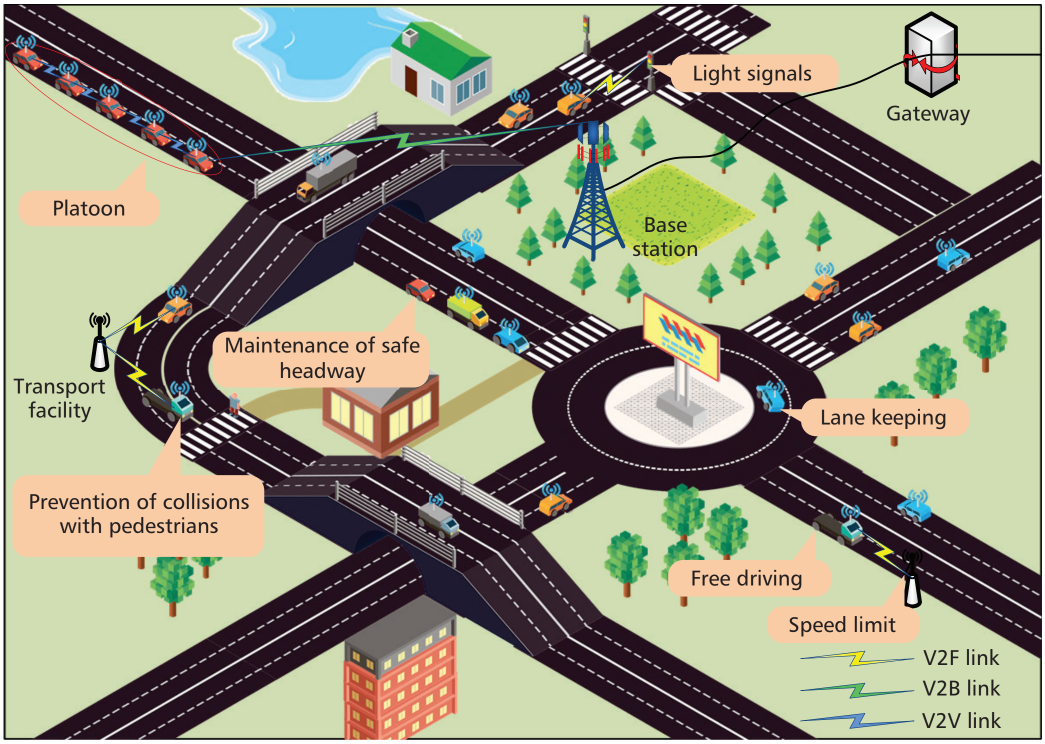


Figure 1. Illustration on HetVNET infrastructure for autonomous driving.

Based on the information obtained from perception and localization modules, the navigation behavior of an ADV can be determined via a planning module. The control module executes the desired command received from the planning module; and the system management module supervises the overall state of the autonomous driving system.

isfactory services through only a single wireless access network, such as dedicated short range communication (DSRC) or third generation (3G) Long Term Evolution (LTE) networks [6–8]. In particular:

- DSRC networks are mainly designed for short-range communication without considering the pervasive communication infrastructure.
- The delay incurred in LTE networks may significantly deteriorate with a large number of vehicles, while strict latency for delivering real-time information for autonomous driving is required.
- The huge volume of data generated by sensors in ADVs is beyond the capacity of the current vehicular networks.

Therefore, this article presents heterogeneous vehicular networks (HetVNETs), which integrate different types of wireless access networks, such as LTE and DSRC, in order to satisfy the various communication requirements of autonomous driving [9, 10]. We also improve the existing protocol stacks and define certain types of messages in HetVNETs that are essential to support autonomous driving. Taking advantage of these different message types, ADVs can achieve the desired traffic behavior such as overtaking, changing lanes, and so on. Moreover, several typical autonomous driving scenarios are discussed and thoroughly analyzed through studying their specific traffic and communication characteristics. Moreover, in order to ensure reliable low-latency message delivery and efficient provision of high data rate transmission for ADVs, several advanced techniques are correspondingly proposed for HetVNETs.

HetVNETs for Autonomous Driving

Network Infrastructure

As shown in Fig. 1, there are mainly three types of vehicles in HetVNETs for autonomous driving, as follows.

Manually Driven Vehicle with Communication Modules: A manually driven vehicle (MDV) is fully controlled by humans. Different from traditional vehicles, each MDV is equipped with an appropriate communication module, for example, having both DSRC and LTE communication, which support real-time information exchange among neighboring vehicles as well as between vehicles and base stations (BSs).

Autonomous Driving Vehicle: An ADV has the ability to cruise safely without human intervention. Apart from the communication module, five other basic modules are usually needed to support autonomous driving, that is, perception, localization, planning, control, and system management [1]. Perception is the process that obtains a clear view of the surrounding environment via various techniques such as radar, lidar, and vision detection. Localization can be implemented by using a global positioning system (GPS). Based on the information obtained from perception and localization modules, the navigation behavior of an ADV can be determined via a planning module. The control module executes the desired command received from the planning module, and the system management module supervises the overall state of the autonomous driving system.

Platoon: A platoon consists of a group of vehicles, that is, one head vehicle and several

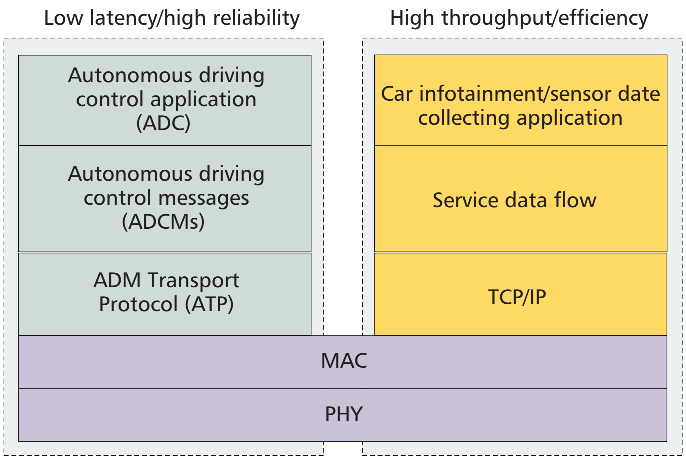

Low latency/high reliability
High throughput/efficiency
Autonomous driving control application (ADC)
Car infotainment/sensor date collecting application
Autonomous driving control messages (ADCMs)
Service data flow
ADM Transport Protocol (ATP)
TCP/IP
MAC
PHY

Figure 2. Protocol stack to support ADMs in HetVNETs.

Categories	Message contents	Examples
Periodic state messages	Position	Example 1: When ADVs are traveling on the road, they need to broadcast and report PSMs at appropriate intervals. Example 2: When any ADV is out of order, it needs to broadcast malfunction messages to warn nearby ADVs to keep some distance from it.
	Direction	
	Speed	
	Malfunction	
	Others	
Action-triggered messages	Change lanes	Example 1: ADV 1 broadcasts ATMs to warn surrounding ADVs before it starts to change lanes. Example 2: When an emergency vehicle enters into a road segment, it needs to broadcast ATMs to other ADVs to let them pull over.
	Overtake	
	Brake	
	Emergency vehicle avoidance	
	Others	

Table 1. Types of autonomous driving control messages in HetVNETs.

followers. The head vehicle can be either an MDV or an ADV, while the followers incorporate automatic longitudinal speed control, and their lateral movements are controlled by the header. In order to safely control the followers, the head vehicle can interrupt the platoon mode at any time. In addition, safe headaway between vehicles must be maintained as well as other safety requirements. The required communication capabilities of a platoon are to broadcast its active state using safety messages, to receive these messages, and to establish unicast sessions.

To achieve autonomous driving, it is very important to let the ADVs reliably identify the behavior of other vehicles. Hence, the communication for information exchange is critical, which may happen in the following different ways.

V2V Communication: It provides an efficient way for ADVs to share information with each other, which can help to enhance safety, reduce traffic congestion, and avoid vehicle collisions.

V2I Communication: According to different types of network infrastructure, V2I communication can be further divided into V2F for local communication and V2B for global communication as follows.

- **Vehicle-to-Facility:** The communication module can also be installed at road facilities (e.g., speed limit signs and traffic lights), which provide the capability to disseminate periodical and event-driven messages. The main functionality of V2F communications is to broadcast warning messages to specific road sections, speed limit notifications, and traffic light signals to nearby vehicles in order to avoid accidents. Another functionality is to act as a relay link to improve the communication reliability in given areas such as intersections with high buildings or obstacles. Also, V2F can be used by transport management departments to regulate the vehicles on the road.
- **Vehicle-to-Base Station:** V2B communication mainly refers to the wireless link between vehicles and BSs. Through V2B links, the vehicles can access the core network (CN) and Internet. It is mainly supported by cellular networks and plays an important role in autonomous driving. Most non-safety-related services are provided through V2B links. Based on the information obtained via V2B links, the service center can obtain a global view of the traffic network and give useful insights to vehicles such as optimal navigation paths.

By integrating HetVNETs into the autonomous driving system, the travel efficiency and safety of the vehicles is envisioned to be significantly improved.

Types of Autonomous Driving Messages

Figure 2 presents the proposed HetVNETs protocol stack for vehicular networks, which is expected to support the various types of messages for autonomous driving. Autonomous driving control (ADC) applications are in charge of the control and management of ADVs. The layer of ADC messages (ADCMs) is used in order to support these applications. Similar to wireless access in vehicular environments (WAVE) Short Message Protocol (WSMP) in DSRC [11], efficient ADCM Transport Protocol (ATP) is designed since it is critical to deliver ADCMs with low latency for high mobility scenarios. One of the basic functions of ATP is to provide broadcasting services without connection establishment, which facilitates messages dispensed between ADVs or from ADVs to network infrastructure. On the other hand, the passengers of ADVs want to use entertainment services while traveling. Meanwhile, ADVs are equipped with a large number of sensors resulting in the generation of massive volumes of sensor data. Thus, there is a great demand for HetVNET to support applications that have high throughput and efficiency requirements.

As shown in Table. 1, ADCMs can be roughly categorized into two types.

Periodic State Messages: PSMs are mainly employed to indicate the state information of vehicles, such as position and traveling directions. This information can be collected by the neighboring ADVs in order to estimate safety factors before taking any action and is also broadcast to infrastructure via V2B links. Based on the PSMs, the service center can make data analysis, gather statistics of traffic flow, and so on.

Action-Triggered Messages: ATMs include the action contents of the ADVs, which can be used for decision in the next moment. Only by utilizing these messages can an ADV accurately know the surroundings as well as the movement of other ADVs. Thus, it can take the appropriate reaction autonomously and then send its changing state to other vehicles.

For the sake of illustration, a few examples are also given in Table. 1.

Typical Scenarios of Autonomous Driving

In order to understand the requirements of ADVs in HetVNETs, it is necessary to study several typical application scenarios with autonomous driving. Thus, three scenarios — highway free flow, highway synchronized flow, and urban intersection — are discussed in this section, where both traffic and communication characters are analyzed.

Scenario 1: Highway Free Flow

Traffic Characteristics: As illustrated in Fig. 3a, the number (density) of ADVs is very small (low) in a highway free flow scenario. In general, four types of ADV traffic behaviors exist, as listed below.

- **Normal Driving Behavior:** ADVs can travel freely at the desired speed, which is not constrained by other vehicles on the road. This is the main behavior in the free flow scenario.
- **Overtaking Behavior:** In order to maintain the desired speed, ADVs sometimes need to overtake other vehicles, which travel at relatively low speed. There are two phases in the overtaking action. The first phase is lane changing with safety when the adjacent lanes are vacant. The second one is to accelerate and surpass the heading vehicle until it acquires a safe distance and then changes back to the original lane. As illustrated in Fig. 3a, ADV 3 is driving behind ADV 2, the speed of which is lower than that of ADV 3. In this case, ADV 3 may overtake ADV 2 for the sake of maintaining its desired speed. Therefore, ADV 3 needs to first inform ADV 2 about its action through V2V communication, and then checks whether the adjacent passing lane is safe or not. If the adjacent passing lane is clear, it can execute lane change action to overtake another ADV. Otherwise, it must wait for a while before the next attempt.
- **Avoidance Behavior:** For the purpose of safe driving, avoidance behavior is essential as a driving behavior of the highway free flow. For example, when ADV 4 and ADV 5 are too close in Fig. 3a, ADV 5 has to slow down to avoid bumping into ADV 4.
- **Emergency Avoidance Behavior:** When ADVs are traveling, they may encounter some emergency vehicles (e.g., ADV 6 in Fig.3a). At this moment, ADVs have to reduce their speed and pull over immediately in order to ensure that emergency vehicles quickly pass by. Thus, ADV 6 broadcasts the appropriate emergency messages to other vehicles all the time when it is moving. This behavior may happen in all three scenarios, and thus is not discussed in the next scenarios.

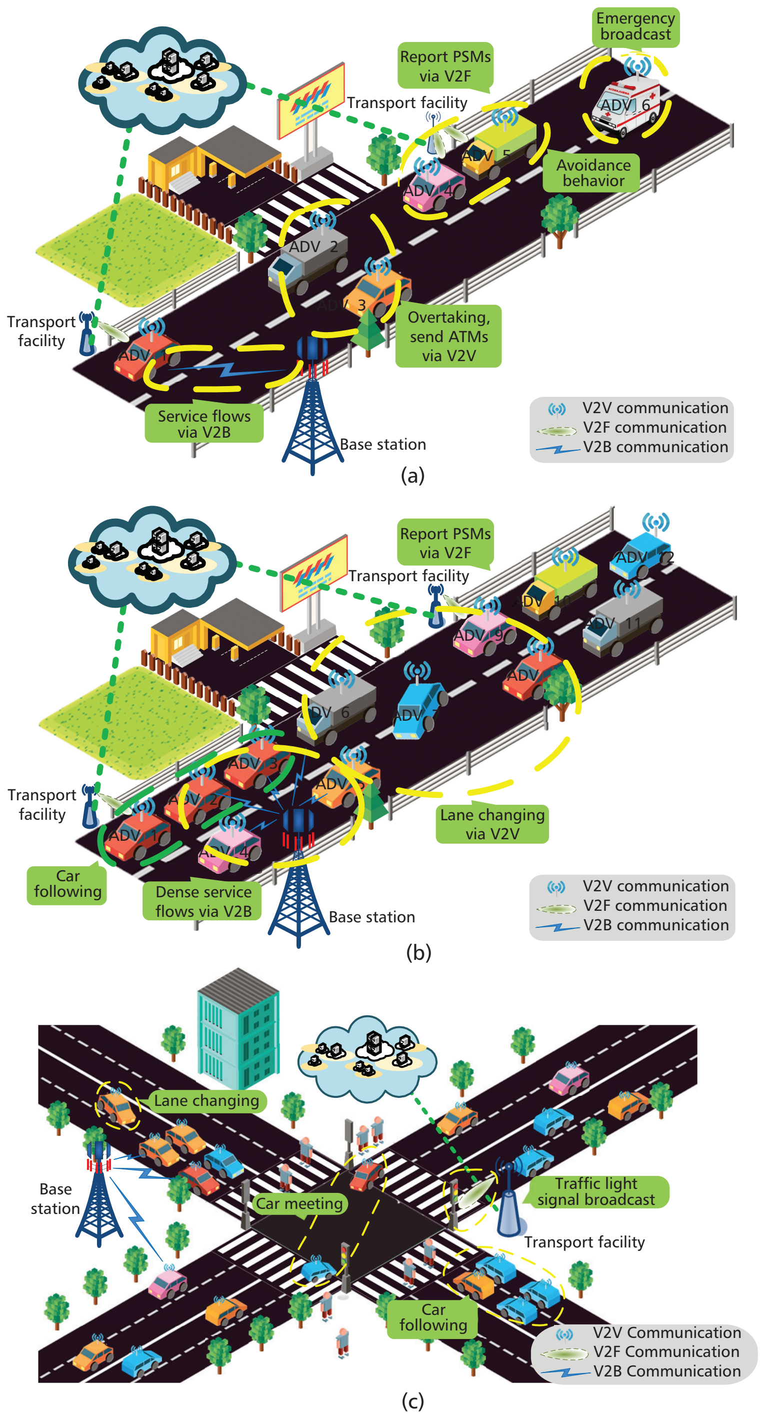


Figure 3. Illustration of typical scenarios of autonomous driving vehicles on the road: a) scenario 1: highway free flow; b) scenario 2: highway synchronized flow; c) scenario 3: urban intersection.

Communication Characteristics: Each traffic behavior has its specific communication requirements in order to achieve the corresponding action safely. Therefore, different communication ways with the corresponding message types need to be applied to guarantee the success of each traffic behavior. For example, an ADV with normal driving behavior needs to distribute its

Due to the high speed of the vehicles in the free flow, the fast fading propagation effects of the radio channels are quite serious, which significantly deteriorate the quality of communication links. Therefore, it becomes a challenge how to guarantee reliable communication under such scenario.

PSMs through both V2V and V2I links. Moreover, it is necessary for an ADV to send ATMs via V2V communication to ensure safety when it performs the overtaking.

Furthermore, services such as entertainment applications can be enjoyed mainly by the occupants through V2B communication. However, due to the high speed of the vehicles in free flow, the fast fading propagation effects of the radio channels are quite serious, which significantly deteriorate the quality of communication links. Therefore, guaranteeing reliable communication under such a scenario becomes a challenge.

Scenario 2: Highway Synchronized Flow

Traffic Characteristics: As illustrated in Fig. 3b, the number of vehicles in the highway synchronized flow scenario is larger than that in the free flow, which results in higher density and lower speed. Generally, the synchronized flow contains two traffic characteristics:

- Vehicles are in flux while their speed is relative high.
- The speed of all vehicles in different lanes tends to be synchronized.

Therefore, ADVs traveling in this scenario have very limited flexibility. Such characteristics give rise to three types of typical traffic behaviors, as listed below.

- **Car Following Behavior:** Due to the converging speed, ADVs have to follow the front vehicles, which is the most common behavior under this scenario. In Fig. 3b, there are a few ADVs with car following behavior, for example, ADV 1, ADV 2, and ADV 3. If ADV 1 encounters an unexpected event and thus has to slow down, ADV 2 and ADV 3 also have to reduce their speed correspondingly in order to maintain safe headway.
- **Lane Changing Behavior:** In the synchronized flow, overtaking ADVs is very difficult due to the high vehicle density. However, in order to improve traffic efficiency, some ADVs with relatively high speed are likely to change lanes (e.g., ADV 7 in Fig. 2b).
- **Avoidance Behavior:** This behavior is very important to avoid car crashes, especially in this scenario. Different from the free flow scenario, an ADV (e.g., ADV 8 in Fig. 3b) is likely to encounter a collision possibly due to lane changing as well as the car following.

Communication Characteristics: In order to guarantee traffic safety and efficiency, cooperative communication among vehicles is likely to happen, which is quite different from the highway free flow scenario. For instance, cooperative lane changing can decrease the time of direct lane changing to maximize traffic efficiency. Compared to the free flow scenario, besides the severe fast fading prorogation effects, the high density of vehicles makes network topology more complex. Meanwhile, a huge amount of communication requests is generated by the different ADVs on the road. All of these make the target of reliable and efficient communications extremely difficult. Novel communication techniques are desired to be applied, especially in this scenario.

Scenario 3: Urban Intersection

Traffic Characteristics: As illustrated in Fig. 3c, at the intersection ADVs not only interact with other vehicles, but also with transport facilities and pedestrians. Thus, the situation becomes much more complicated than those in highway scenarios. The behavior of vehicles and pedestrians is regulated by traffic lights. The driving speed (e.g., the maximum speed is 40 km/h) is much lower than in highway scenarios. Moreover, overtaking is seldom performed due to the high traffic load at the intersection. The following behaviors commonly happen.

- **Car Following Behavior:** This behavior usually occurs when vehicles are queuing in the lane to take a turn. The key message before executing this behavior is related to the accurate position of the car ahead.
- **Lane Changing Behavior:** When a vehicle drives on a lane that is not its target lane, it has to change lanes. For example, a vehicle has to go to the left lane if it wants to turn left.
- **Car Meeting Behavior:** Car meeting behavior is complicated, since two vehicles are involved and interact with each other. For example, turning right has higher priority than turning left when the target lanes are the same.

Communication Characteristics: The stop-and-go control at an intersection ensures safe crossing at the intersection. However, it may introduce the inconvenience of frequent stops and idling until achieving the right of way, which significantly reduces traffic efficiency. Through efficient communication via either V2V or V2I links in an intersection, each vehicle may have adequate maneuver commands in real time. Therefore, traffic operations at an intersection without stop-and-go-style traffic lights and signs become available when the road is full of ADVs.

Potential Challenges and Solutions

HetVNETs, utilizing improved protocols and messages, provide the feasibility of supporting the behavior of ADVs. However, new communication challenges arise when the quality of service (QoS) requirements of ADC messages and other services need to be supported in various scenarios. To tackle these challenges in HetVNETs, novel techniques from signal transmission to networking are then discussed as possible solutions.

Low-Latency and Highly Reliable Transmission Techniques

In autonomous driving systems, different types of messages or information have different QoS requirements, such as low latency and high reliability for safety messages, and high data rate for non-safety multimedia applications. It is hard to meet all these requirements with only the existing transmission techniques in either LTE or DSRC networks, which motivates the development of new transmission techniques in HetVNET for implementing autonomous driving.

Filtered-orthogonal frequency-division multiplexing (F-OFDM) applies a sub-band digital fil-

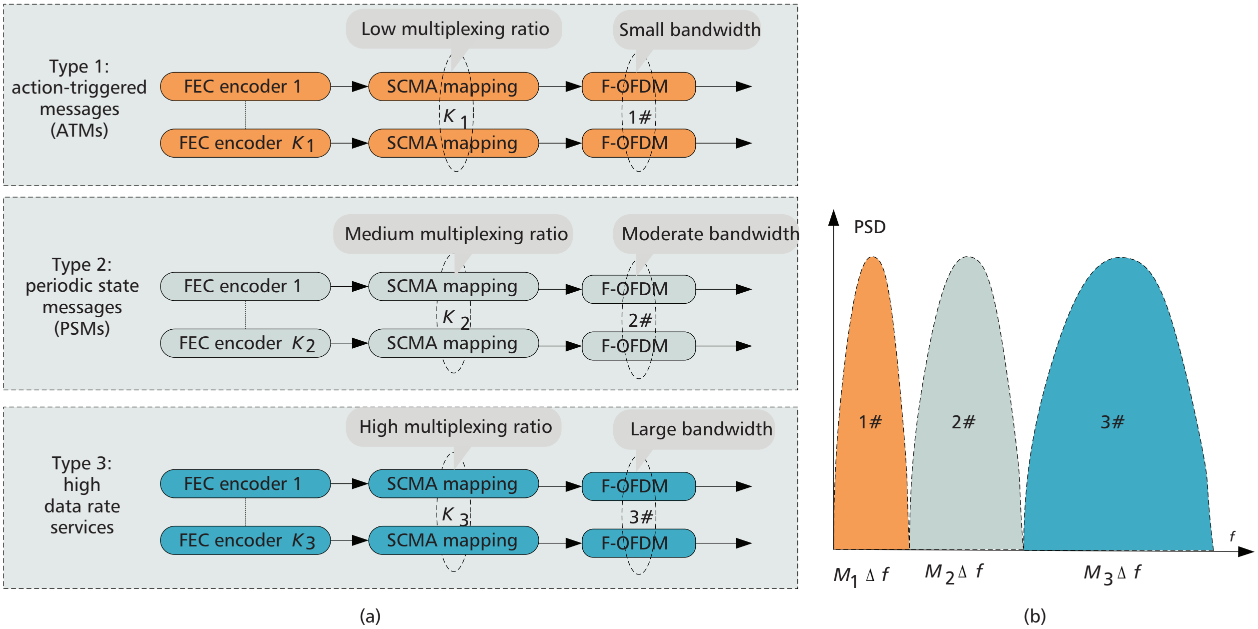


Figure 4. A paradigm of transmission design for HetVNETs: a) transmitter; b) bandwidth.

ter to shape the spectrum of a sub-band OFDM signal, which has good out-of-band leakage rejection and thus supports asynchronous orthogonal frequency-division multiple access (OFDMA) transmission without a timing advanced signal [12]. On the other hand, through joint optimization of multi-dimensional quadrature amplitude modulation (QAM) and non-orthogonal sparse codewords, sparse code multiple access (SCMA) is capable of multiplexing more users and improving system reliability [13]. Therefore, the combination of SCMA and F-OFDM is expected to be one of the possible candidate techniques to satisfy the communication requirements of HetVNETs in autonomous driving.

Figure 4a presents an example of a transmitter based on the combination of SCMA and F-OFDM, which supports three types of services. The three corresponding subbands are also shown in Fig. 4b, and are called *subband 1#* with bandwidth $M_1\Delta f$, *subband 2#* with bandwidth $M_2\Delta f$, and *subband 3#* with bandwidth $M_3\Delta f$ (where Δf is the subcarrier bandwidth and $M_1 \le M_2 \le M_3$). The ATMs and PSMs with low latency and high reliability can be transmitted in *subbands 1#* and *2#*, respectively. Meanwhile, the vehicles with non-safety multimedia applications of high date rate are served in *subband 3#*. Different number of vehicles (i.e., K_1, K_2, K_3) may be scheduled in different subbands. Usually the number of multiplexing vehicles in *subbands 1#* and *2#* are chosen to be small, while no such limitation exists in *subband 3#*.

This paradigm is capable of satisfying low-latency, high-reliability, and massive access for autonomous driving. First, the asynchronous character of F-OFDM and the grant-free access of SCMA, that is, the message or information transmission without signaling overhead between vehicles (e.g., grant request and acknowledgment), can significantly reduce delay of message delivery. Moreover, a small number of vehicles in a single subband may decrease the collision probability of SCMA codewords, which can improve the decoder performance of the Message Passing Algorithm (MPA) and thus increase transmission reliability. Moreover, F-OFDM is capable of utilizing the fragmented spectrum resources and shaping flexible bandwidth for different kinds of services, which maximizes spectrum efficiency. Thus, together with SCMA multiplexing more users, F-OFDM transmission is capable of supporting massive access of ADVs.

Besides, SCMA and F-OFDM do not change the essential transmission character of either downlink OFDM or uplink OFDMA. The combination of SCMA and F-OFDM keeps good backward compatibility with existing LTE systems. Therefore, SCMA/F-OFDM transmission provides a feasible evolutional roadmap from LTE to HetVNETs.

Cooperative Autonomous Driving Techniques

In complicated vehicular environments, ADVs need to have an in-depth understanding of their surrounding environment to make optimal cooperative driving decisions and path scheduling. However, the intrinsic limitations of traditional information perception such as cameras and radar often prevent cooperative decisions. Therefore, in this subsection, a cooperative autonomous driving framework based on HetVNETs is proposed, where each ADV can share information both locally for traffic safety and globally for traffic efficiency.

As illustrated in Fig. 5a, the proposed hierarchical cooperation can be divided into two layers: small-scale and large-scale cooperation. The former is executed only within the local area and at fine time resolution in order to ensure safety. The latter happens across a large geographical area with coarse time resolution in order to enhance traffic efficiency.

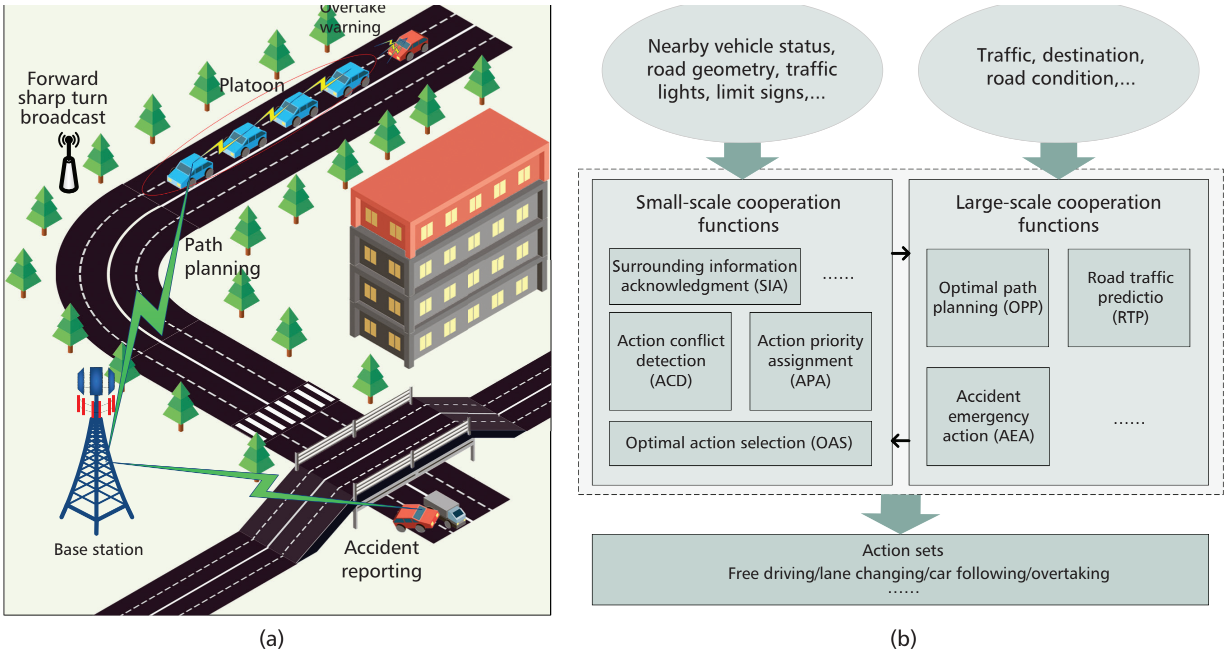

Overtake warning
Forward sharp turn broadcast
Platoon
Path planning
Base station
Accident reporting
(a)
Nearby vehicle status, road geometry, traffic lights, limit signs,...
Traffic, destination, road condition,...
Small-scale cooperation functions
Surrounding information acknowledgment (SIA)
......
Action conflict detection (ACD)
Action priority assignment (APA)
Optimal action selection (OAS)
Large-scale cooperation functions
Optimal path planning (OPP)
Road traffic predictio (RTP)
Accident emergency action (AEA)
......
Action sets
Free driving/lane changing/car following/overtaking
......
(b)

Figure 5. Illustration of hierarchical cooperative autonomous driving scheme: a) example scenario; b) typical functions.

Small-Scale Cooperation: The main objectives of small-scale cooperation are to guarantee traffic safety through cooperation between vehicles in the local area. Such cooperation is implemented in a distributed manner, which significantly reduces signal overhead between vehicles. As illustrated in Fig. 5b, there are several typical functions needed to support small-scale cooperation.

- **Surrounding Information Acknowledgment:** SIA can provide accurate environment information for the next step that can be classified into two types: dynamic and static information. The former includes nearby vehicle state information and event-driven messages, which are transmitted via V2V link. The latter mainly contains information such as lane, intersection, and speed limit, which is provided by V2I communication.
- **Optimal Action Selection:** OAS generates an optimal action for the next interval, which may need to solve an optimization problem based on obtained surrounding information and events. The action set mainly includes free driving, lane changing, lane keeping, car following, overtaking, platoon, and so on.
- **Action Conflict Detection:** A vehicle checks its required action with those of neighboring vehicles. Only when there is no conflict, the control command for this action can be executed.
- **Action Priority Assignment:** When a conflict happens between the required actions of the vehicles, only the action with the higher priority may be chosen.

Large-Scale Cooperation: It aims to disseminate information over a large geographical area to improve traffic efficiency. Furthermore, a few functionalities such as path prediction and scheduling capabilities of involved vehicles can be leveraged when the upcoming traffic congestion can be detected in advance via large-scale cooperation. Different from small-scale cooperation, it is executed in a centralized manner via V2B links. The framework of the large-scale cooperation is illustrated in Fig. 5b. First, the cloud server collects information such as road conditions, unexpected traffic congestion, adverse weather conditions, and traffic density via V2B links. Then it calculates the corresponding results for different applications. There are a few functions to support larger-scale cooperation, for example, optimal path planning, road traffic prediction, and accident emergency action.

Layered-Cloud Computing Techniques

It is estimated that an ADV generates around 1 GB/s, which mainly comes from sensors. To store such an amount of data in a vehicle during travel needs a huge local storage unit. Therefore, the remote cloud (RC) is proposed as a feasible solution by the aid of offloading techniques over high throughput wireless transmission. In particular, it provides abundant communication and computing resources in order to ensure the safety and traffic efficiency of ADVs. However, when ADVs become more and more popular, thousands of ADVs may be present on the road and simultaneously generate sensor data. Thus, it is impractical to transmit all the sensor information of each ADV over V2I links, which is a very important challenge even for 5G networks with the peak rate of 10 GB/s. Therefore, the wireless links between ADVs and the RC have to be efficiently utilized.

On one hand, since the data generated by ADVs has substantial correlation in the time domain, it is possible to process and compress data before transmission over V2I links. For example, when the sensor data change continuously in time, the ones with very small variation can be omitted. On the other hand, another obvious feature of the generated data is its local interests, which means that only ADVs in the vicinity are likely to enjoy common interests such as local traffic congestion and road condition messages. Therefore, the data of common interests can be kept local rather than uploaded to the RC, which may greatly reduce the capacity requirements of the V2I links.

Moreover, collaboration in the sharing and processing of sensor data between ADVs can significantly improve the location accuracy and safety of the driving. Vehicular cloud computing (VCC) is a promising new technology that takes advantage of cloud computing to serve vehicles [14]. The computing and storage resources in VCC can be utilized to enhance the abilities of ADVs. In other words, vehicular clouds (VCs) can provide a good platform for the coordinated deployment of the sensor aggregation, fusion, and database sharing applications required by ADVs. For example, ADVs can enlarge the sensing coverage by aggregating the data from geographically distributed ADVs.

Therefore, a layered cloud computing architecture for ADVs can be deployed as one feasible solution. It includes not only an RC but also the VCs. The ADVs can send a request of either driving or entertainment to any layer of the cloud.

A layered-cloud computing architecture for ADVs can be deployed as one of the feasible solutions. It includes not only a RC but also the VCs. The ADVs can send the request of either driving or entertainment to any layer of the cloud

Conclusion

Reliable and efficient communication is extremely important to guarantee the safety and comfortability of ADVs. Therefore, in this article, we propose HetVNETs together with an improved protocol stack and new messages that can support autonomous driving. Then three typical scenarios for the ADVs including highway and urban intersection are discussed. Their specific traffic behavior puts forward the various communication requirements, which raise various technical challenges for HetVNETs. Thus, in order to combat these challenges, new techniques such as F-OFDM/SCMA transmission, hierarchical cooperative driving, and layered cloud computing are presented for the development of HetVNETs toward the 5G era.

Acknowledgment

This work is funded in part by National Science Foundation of China (No.61331009), the National High-Tech R&D Program (863 Program 2015AA01A705), National Key Technology R&D Program of China (No.2015ZX03002009-004) and Fundamental Research Funds for the Central Universities (No.2014ZD03-02).

References

[1] K. Jo *et al*., "Development of Autonomous Car — Part I: Distributed System Architecture and Development Process," *IEEE Trans, Industrial Electronics*, vol. 61, no. 12, Dec. 2014, pp. 7131–40.

[2] J. M. Anderson *et al*., *Autonomous Vehicle Technology: A Guide for Policymakers*, RAND Corp., 2014, pp. 9–31.

[3] C. Urmson *et al*., "Autonomous Driving in Urban Environments: Boss and the Urban Challenge," *J. Field Robotic*, vol. 25, no. 8, Aug. 2008, pp. 425–68.

[4] J. Bierstedt *et al.*, "Effects of Next Generation Vehicles on Travel Demand and Highway Capacity," FP Think Working Group, Jan. 2014, pp. 10–11.

[5] J. Choi *et al*., "Environment-Detection-and-Mapping Algorithm for Autonomous Driving in Rural or Off-Road Environment," *IEEE Trans. Intelligent Transportation Syst*ems, vol. 13, no. 2, May 2012, pp. 974–82.

[6] A. Vinel, "3GPP LTE versus IEEE 802.11 p/WAVE: Which Technology is Able to Support Cooperative Vehicular Safety Applications?" *IEEE Wireless Commun. Lett.*, vol. 1, no. 2, 2012, pp. 125–28.

[7] G. Araniti *et al*., "LTE for Vehicular Networking: A Survey," *IEEE Commun. Mag.*, vol. 51, no. 5, May 2013, pp. 148–57.

[8] L. Lei *et al*., "Performance Analysis of Device-to-Device Communications with Dynamic Interference Uusing Stochastic Petri Nets," *IEEE Trans. Wireless Commun*., vol. 12, no. 12, Dec. 2013, pp. 6121–41.

[9] F. Dressler *et al*., "Intervehicle Communication: Quo Vadis," *IEEE Commun. Mag*., vol. 52, no. 6, June 2014, pp. 170–77.

[10] K. Zheng *et al*., "Heterogeneous Vehicular Networking: A Survey on Architecture, Challenges and Solutions," *IEEE Commun. Surveys & Tutorials*, vol. PP, no. 99, 2015, pp. 1–21.

[11] IEEE Std 1609.3-2010, "IEEE Standard for Wireless Access in Vehicular Environments (WAVE) — Networking Services," Dec., 2010.

[12] J. G. Andrews *et al*., "What Will 5G Be?" *IEEE JSAC*, vol. 32, no. 6, June 2014, pp. 1065–82.

[13] K. Hu *et al*., "Uplink Contention Based SCMA for 5G Radio Access," *Proc. IEEE GLOBECOM Wksps.*, Austin, TX, Dec. 2014, pp. 900–05.

[14] E. Lee, M. Gerla, and S. Oh, "Vehicular Cloud Networking: Architecture and Design Principles," *IEEE Commun. Mag.*, vol. 52, no. 25, Feb. 2014, pp. 148–15.

Biography

Kan Zheng (SM'09) is currently a full professor at Beijing University of Posts & Telecommunications (BUPT), China. He received his B.S., M.S., and Ph.D degrees from BUPT in 1996, 2000, and 2005, respectively. He has rich experiences in the research and standardization of new emerging technologies. He is the author of more than 200 journal articles and conference papers in the field of wireless networks, M2M networks, VANETs, and so on. He holds Editorial Board positions for several journals. He has organized several Special Issues in famous journals, including *IEEE Communications Surveys & Tutorials*, *IEEE Communication Magazine*, and *IEEE System Journal*.

Qiang Zheng received his B.S. degree from the College of Computer Science and Technology, Shandong University of Technology (SDUT), China, and M.S. degree in telecommunication and information system from BUPT in 2010 and 2012, respectively. He is currently a Ph.D. candidate in BUPT. His research interests focus on radio resource allocation, performance analysis, and optimization in heterogeneous vehicular networks.

Haojun Yang received his B.S. degree from BUPT in 2014, where he is currently a Ph.D. candidate. His research interests include vehicular networks and wireless communications.

Lu Hou received his B.S. degree from BUPT in 2014, where he is currently a Ph.D. candidate. His research interests include SDN and resource allocation in mobile cloud computing systems.

Long Zhao received a Ph.D. degree from BUPT in 2015, where he is currently a lecturer. From April 2014 to March 2015, he was a visiting student at the Department of Electrical Engineering, Columbia University. His research interests include wireless communications and signal processing.

Periklis Chatzimisios (SM'13) serves as an associate professor at the Computing Systems, Security and Networks (CSSN) Research Lab of the Department of Informatics at the Alexander TEI of Thessaloniki, Greece. Recently he was a visiting academic/researcher at the University of Toronto, Canada, and Massachusetts Institute of Technology. He is involved in several standardization activities serving as a member of the Standards Development Board for IEEE ComSoc (2010–present) and lately as an active member of the IEEE Research Groups on IoT Communications & Networking Infrastructure and Software Defined & Virtualized Wireless Access. He is the author/editor of 8 books and more than 100 peer-reviewed papers and book chapters on the topics of performance evaluation and standardization activities of mobile/wireless communications, quality of service/quality of experience, and vehicular networking. His published research work has received more than 1500 citations by other researchers. He received his Ph.D. from Bournemouth University, United Kingdom (2005) and his B.Sc. from Alexander TEI of Thessaloniki (2000).

Enabling Seamless V2I Communications: Toward Developing Cooperative Automotive Applications in VANET Systems

Arindam Ghosh, Vishnu Vardhan Paranthaman, Glenford Mapp, Orhan Gemikonakli, and Jonathan Loo

The authors are with Middlesex University London.

ABSTRACT

Cooperative applications for VANETs will require seamless communication between vehicle to infrastructure and vehicle to vehicle. IEEE 802.11p has been developed to facilitate this effort. However, in order to have seamless communication for these applications, it is necessary to look at handover as vehicles move between roadside units. Traditional models of handover used in normal mobile environments are unable to cope with the high velocity of the vehicle and the relatively small area of coverage with regard to vehicular environments. The Y-Comm framework has yielded techniques to calculate the time before vertical handover and the network dwell time for any given network topology. Furthermore, by knowing these two parameters, it is also possible to improve channel allocation and resource management in network infrastructure such as base stations, relays, and so on. In this article we explain our overall approach by describing the VANET Testbed and show that in vehicular environments it is necessary to consider a new handover model that is based on a probabilistic rather than a fixed coverage approach. Finally, we show a new performance model for proactive handover, which is then compared with traditional approaches.

INTRODUCTION

Seamless interoperability in highly mobile environments, such as vehicular ad hoc networks (VANETs), is vital in order to develop cooperative applications that can make full use of networking infrastructure. Traditional handover policies have been based on a reactive approach in which the mobile node (MN) reacts to signaling indicating changes in network connectivity as the MN moves around. However, in highly mobile environments with small cell coverage, such an approach can quickly lead to degradation of connections due to the small time there is to effect a handover.

Proactive handover in which the MN actively attempts to decide when and where to hand over can help to develop an efficient and reliable handover policy mechanism. By using proactive handover, it is possible to minimize packet loss and service disruption as an impending handover can be signaled to the higher layers of the network protocol stack [1]. Two key parameters are used to develop algorithms for proactive handover: time before vertical handover (TBVH), which is the time after which the handover should occur, and network dwell time (NDT), which is the time the MN spends in the coverage of the new network. The Y-Comm research effort has defined two types of proactive handover. The first is knowledge-based, which attempts to know, by measuring beforehand, the signal strengths of available wireless networks over a given area such as a city. This could involve physically driving around and taking these readings. The second, proactive, policy is based on a mathematical model that calculates the point when handover should occur and the time the MN would take to reach that point based on its velocity and direction [1].

In order to develop a useful model for real networks, it is necessary to accurately model the underlying communication mechanisms; hence, simulation based on the measurements from a real testbed is essential. Therefore, we have designed a new VANET Testbed at the Hendon campus of Middlesex University, London, which is currently being fully deployed. This testbed will also provide us with better physical layer and propagation models so that handover can be optimized. This is because, in highly mobile environments, it is necessary to have a more exact knowledge of the communication environment including knowing when a beacon can be reliably received when entering a new network. Such information will allow better management

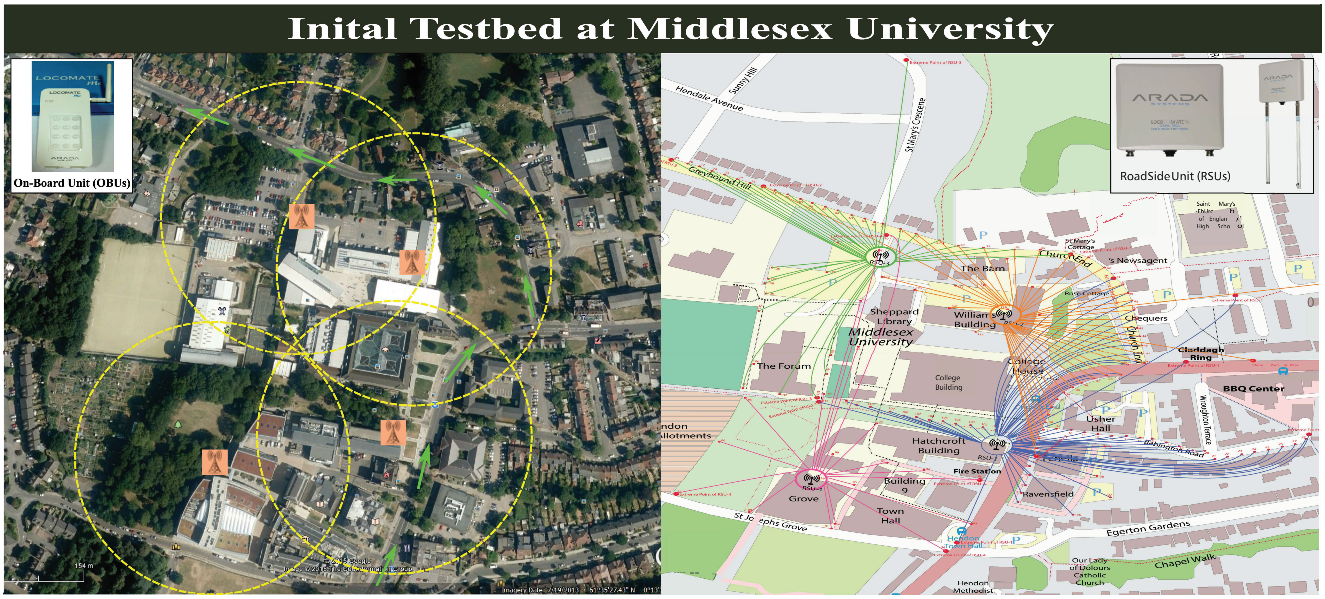


Figure 1. MDX-Vanet testbed: a) satellite view; b) map view.

of the handover process but will require a new probabilistic approach, which is outlined in this article.

A novel aspect in the provision of seamless proactive handover is the design and development of proactive resource allocation techniques. The concept of proactive channel allocation is introduced in this work using TBVH and NDT, but applied to the opportunity of the MN to acquire a channel. These two parameters allow us to determine the times when different nodes will need to acquire and release channels due to mobility. Hence, it is possible to explore periods of contention, which in turn will allow us to develop heuristic algorithms to optimize the use of the channel.

A major area of application of proactive resource allocation is in the area of intelligent transport systems (ITS) using VANETs. Characteristics of VANETs such as high velocity, smaller coverage range, and mobility patterns are serious challenges in providing seamless handover and resource allocation, and moving the services from the previous roadside unit (RSU) to the new RSU. Therefore, developing proactive handover and resource allocation models for VANET systems would be the best option to develop a reliable framework for cooperative applications. Another interesting area is in the management of heterogeneous networking (HetNet) environments using small cells, because their use is considered a promising strategy to cope with the explosion in mobile traffic. However, the signaling load on the network nodes might increase due to frequent handovers, and mobility robustness may be degraded due to increased handover failures and radio link failures [2]. This frequent handover failure can be addressed through proactive handover and resource allocation.

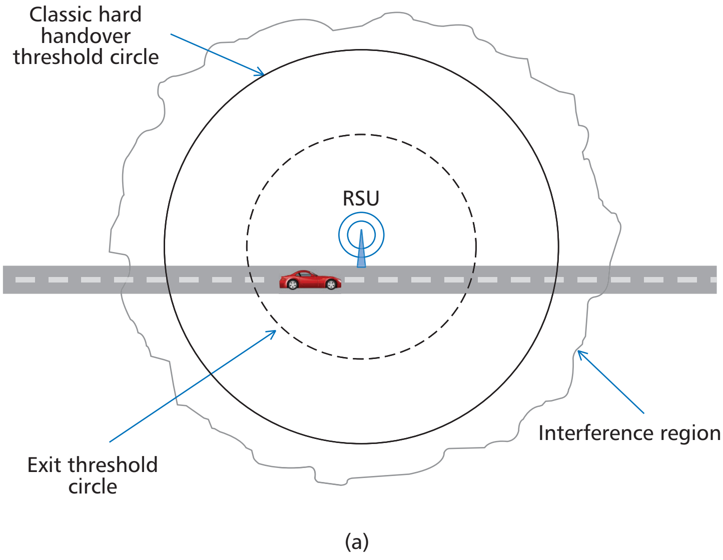
VANET Testbed at Middlesex University

This article presents a real-time VANET testbed that is being used to develop propagation models to gain a better understanding of the relationship between communication and mobility in a given physical space. This is necessary to accurately predict TBVH and NDT in order to develop proactive handover mechanisms. In addition, a probabilistic handover approach is presented based on cumulative probability (CP) using the Veins framework in OMNeT++, which is a discrete event simulation environment. Finally, we present preliminary results to show the benefits of proactive handover on overall system performance.

A VANET testbed is currently being fully deployed at the Hendon Campus, Middlesex University, London, with four RSUs, as shown in Fig. 1. The RSUs and onboard units (OBUs) were manufactured by ARADA Systems with the IEEE 802.11p (Wireless Access in Vehicular Environment — WAVE) standard specifications. Three RSUs were deployed on top of three buildings at varying heights, of which two were deployed to cover the roads around the campus, and the other is used to support the movement of pedestrians within the campus, hence enabling the development of vehicle-to-pedestrian (V2P) applications. The fourth was deployed around the car park area. Initial testing was carried out to measure the coverage of the deployment by using an OBU, moving around the university roads and inside of the campus. The power received was noted for every 10 m and represented as numbered dots as shown in Fig. 1b. In order to explore the path loss models with real test results, as initial work, the power received was compared with the free space path loss (FSPL). This effort was to understand the differences between the theoretical FSPL and measured values. These results indicate that more sophisticated propagation models such as terrain or finite element propagation models need to be developed. This detailed model will allow us to more accurately calculate TBVH and NDT for any given scenario.

Handover Policy Based on Cumulative Probability Approach

Handover in mobile environments can be depicted as shown in Fig. 2a. There is a hard handover threshold circle depicted by a hard barrier, and there is a dotted circle within the hard barrier representing the exit threshold. The exit threshold circle is the boundary to start the handover in order to finish it before reaching the hard barrier, which is needed for a successful soft handover. If the handover is not successful before the hard barrier, there is a break in the communication, which leads to a hard handover. Although this approach is currently being used for mobile communications, in highly mobile environments such as VANETs it presents two challenges: First, the exit radius is dependent on the velocity of the MN, so at high velocities there will be no time to do a soft handover. Second, the hard or fixed handover circle represents the area of coverage, but at this outer region, actual communication is difficult due to the probability of packets been received with error due to low signal-to-noise ratio (SNR). Hence, a more probabilistic approach is required that makes use of CP to provide a realistic boundary for handover.



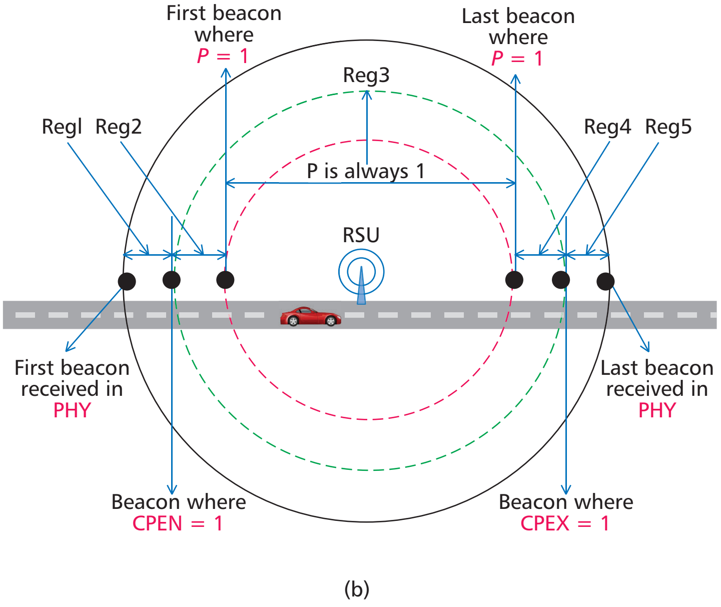

First beacon where P = 1
Last beacon where P = 1
Reg3
Regl Reg2
Reg4 Reg5
P is always 1
RSU
First beacon received in PHY
Last beacon received in PHY
Beacon where CPEN = 1
Beacon where CPEX = 1
(b)

Figure 2. a) Traditional; b) probabilistic segmentation.

Let the probability (P) represent the probability of a successful reception of beacon at the physical (PHY) layer. This probability can be calculated for each beacon with the knowledge of the SNR and the length of the beacon [3, 4]. In probability theory, P has a stationary distribution, that is, the possible outcomes are constant over time. Hence, we can define the CP as the probability of the event occurring — in this case, a successful beacon reception — before a given time or sequence number. In addition, when CP is 1, we are sure that the event has occurred. If P is constant, CP is normally 1 at infinity. In this case, however, P does not have a stationary distribution because as the MN moves toward the RSU, P increases significantly, and hence, CP will become 1 long before infinity and in fact may become 1 before P becomes 1. Thus, this shows that we can be certain of receiving a successful transmission due to CP before P becomes 1. This means it is necessary to use the CP approach to determine the regions of reliable communication. Therefore, we need to calculate CP for a sequence of N beacon receptions and compare it to when P is 1.

We define the CP as the vehicle enters a new network as the entrance CP (CP_{EN}). For exit scenarios, we consider the probability of not receiving the beacon P_n from the RSU as we drive away, that is, the exit CP (CP_{EX}). For the exit side, P the probability of the successful reception decreases as we move away from the RSU; hence, $1 - P$ is increasing. Our results therefore consider the effect of the cumulative frequencies on entrance and exit regions of RSU coverage.

Figure 2b presents the communication time between the segments or regions named Reg_1, Reg_2, Reg_3, Reg_4, and Reg_5. These regions are the communication times, that is, the time duration when beacons are received by the vehicle in a particular segment of RSU coverage as listed below:

- Reg_1: the region between the first beacon being heard in the PHY layer and the point when $CP_{EN} = 1$
- Reg_2: the region between $CP_{EN} = 1$ and the point where P is first equal to 1
- Reg_3: the region where P is always equal to 1
- Reg_4: the region between the last beacon where $P = 1$ and $CP_{EX} = 1$
- Reg_5: the region between $CP_{EX} = 1$ and the last beacon being heard at the PHY layer of that RSU

In order to explore these concepts, a simulation was carried out with one RSU and one vehicle moving along the road using the Veins Framework in OMNeT++. The Framework supports IEEE 802.11p, and the coverage radius of the RSU was 907 m with 20 mW transmission power and the minimum receiver gain set to –94 dBm [5]. For the simulation two different velocities were considered, 10 m/s (i.e., 36 km/h) for urban

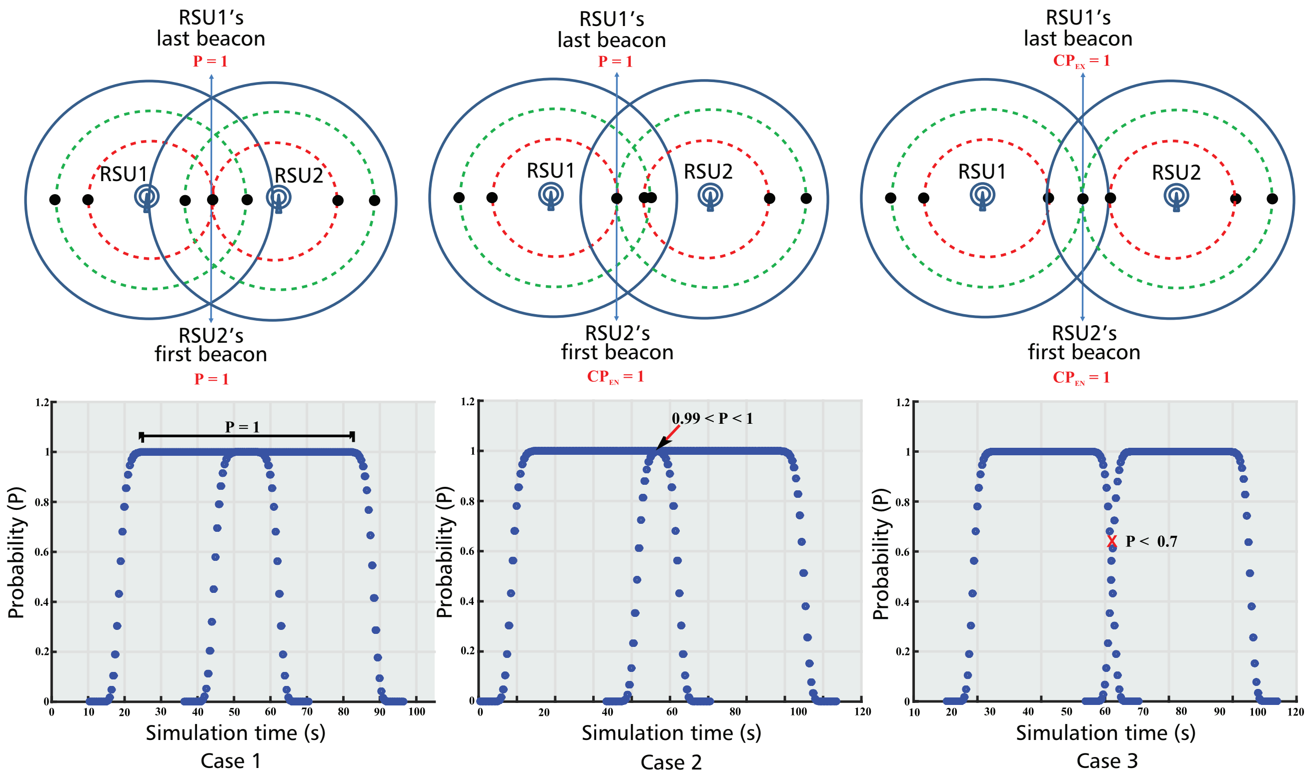


Figure 3. Overlapping scenarios.

speed and 30 m/s (i.e., 108 km/h) for motorway speed. The results in [6] also showed that for handover, a maximum beacon size between approximately 600 to 800 bytes could give the best chance for seamless communication. Hence, beacon sizes of 300, 500, and 723 bytes have been considered to conduct our study. In addition to this, the work in [6] also showed that an ideal range of beacon frequency for vehicular communication is between 10 to 20 Hz. Therefore, beacon frequencies of 10, 15, and 20 Hz are considered in this article. When there is an increase in beacon frequency, a considerable amount of communication time is achieved between $CP_{EN} = 1$ and $P = 1$ (i.e., Reg_2) and between $CP_{EX} = 1$ and $P = 0$ (i.e., Reg_5). This clearly indicates that a high beacon frequency should result in an increased NDT as the beacon is heard almost as soon the vehicle enters the coverage area.

Analysis of Overlapping Region

In order to verify our handover policy based on the CP approach, we have come up with three different scenarios of overlapping two RSUs, as shown in Fig. 3. A mobile node (in our case a vehicle) is made to travel over the coverage range of these two RSUs with velocities of 10 m/s and 30 m/s for collecting various values for our study. The same parameter settings were used as for the first RSU simulation experiment setup for calculating CP.

Case (i) — The two RSUs are overlapped such that RSU 1's last beacon received by the vehicle with $P = 1$ and RSU 2's first beacon with $P = 1$ are received one after another. The time difference between these two beacons is very small, and hence Fig. 3 shows these two beacons at the same point.

Case (ii) — The two RSUs are overlapped such that RSU 1's last beacon with $P = 1$ and RSU 2's first beacon reaching $CP_{EN} = 1$ are received one after another.

Case (iii) — The two RSUs are overlapped such that RSU 1's beacon reaching $CP_{EX} = 1$ and RSU 2's beacon reaching $CP_{EN} = 1$ are received one after another.

The simulation results for each case are illustrated as graphs in Fig. 3. In case (i), as mentioned earlier, the overlapping of two RSUs is set up such that P is 1 for both RSUs at the overlapping region. Thus, it is clearly evident from the graph that once the vehicle reaches the region where $P = 1$ of RSU1, there is no drop in P until the vehicle exits RSU2's $P = 1$ region, that is, P is always 1, as shown in Fig. 3. From this observation it is clear that this is the most reliable way of overlapping adjacent RSUs which ensures seamless handover. But this reliability comes at the cost of more overlapping distance, as shown in the graph in Fig. 3, and high interference issues, as indicated in [7], as both RSUs are in communication range of each other.

In case (ii) as the RSUs are set up such that RSU1's last beacon with $P = 1$ and CP_{EN} of RSU2 is 1 at the overlapping region. This way of overlapping yields less overlapping distance, as shown in Fig. 3, compared to case (i); however, there is a very negligible amount of drop in P

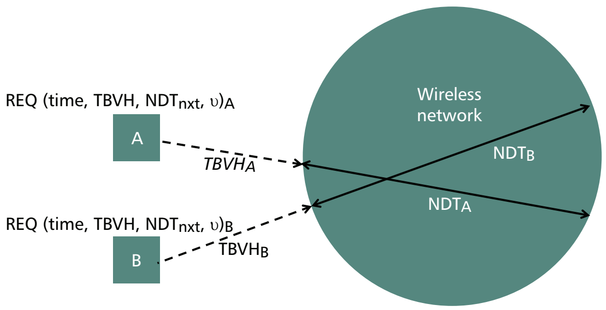


Figure 4. Request for channel allocation.

at the overlapping region that is, $0.99 < P < 1$, Fig. 3. According to [8] P should be greater than 0.99 for safety related applications. Hence, case (ii) is equally reliable and also ensures seamless handover.

In case (iii), the RSUs are set up considering CP_{EX} of RSU1 and CP_{EN} of RSU2 for overlapping. This way of overlapping gives an advantage of a much smaller overlapping distance compared to cases (i) and (ii). This also benefits the network with less interference as indicated in [7]. In the overlapping region, P reduces to less than 0.7, which is not suitable for seamless communication or safety-critical applications.

As shown above, case (ii) performs equally well as case (i); therefore, this approach can be adopted for a scenario where critical life-safety applications are given higher priority. By contrast, the case (iii) approach is more suitable for a scenario where optimal coverage is required and non-safety applications are used.

In addition, the CP approach can be used to improve handover since $CP_{EN} = 1$ tells us when we are certain to have received at least one beacon from the new RSU. Hence, we should ensure that handover can occur before $CP_{EN} = 1$. Similarly, $CP_{EX} = 1$ indicates when we are sure not to have heard a beacon from the current RSU and thus need to ensure that the MN has been handed over to the next RSU before this point. It is therefore no longer necessary to manage handover using the hard handover circle as this probabilistic approach based on CP should yield more reliable results. Therefore, the CP mechanism should be incorporated into the handover mechanism for MNs.

Proactive Channel Allocation

The probabilistic approach in the previous section allows us to calculate NDT and TBVH more accurately. In this section we use these parameters to explore proactive channel allocation. We first look into a simple scenario where a network uses a single channel, and two MNs are moving at a velocity (v) toward that network range as shown in Fig. 4. MN_A and MN_B can request the channel for communication. Assuming that v and TBVH are already known, t_c is the current time at the node and NDT_{nxt} is the estimated NDT

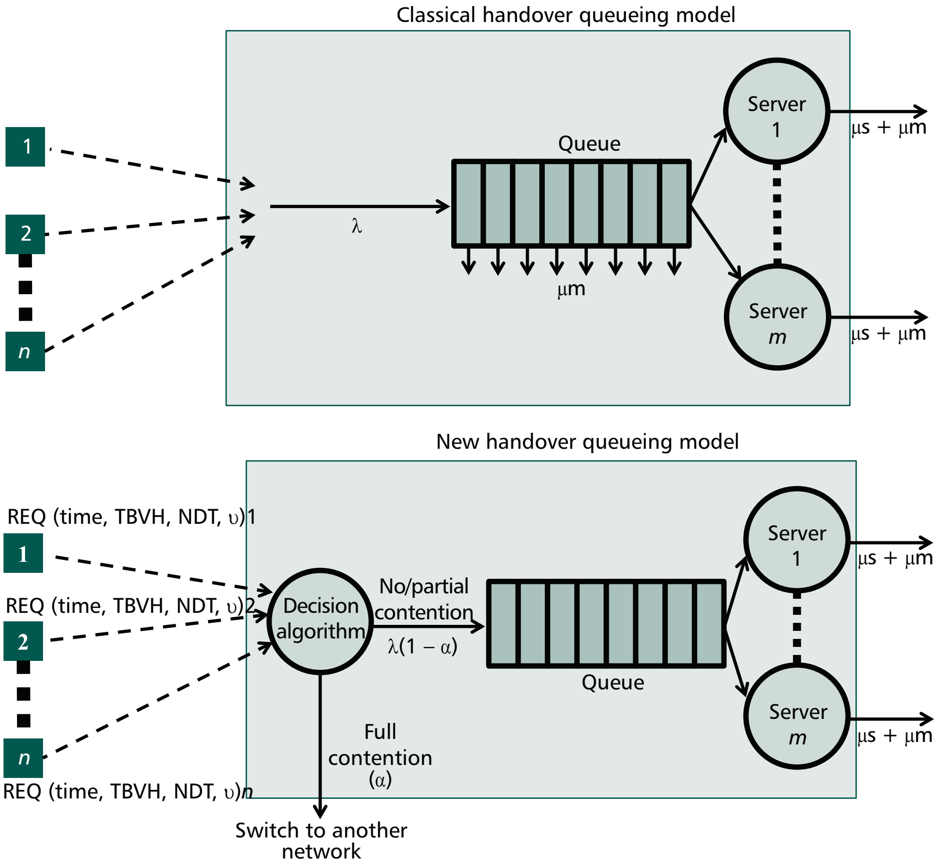

Classical handover queueing model
1
2
n
λ
Queue
μm
Server 1
Server m
μs + μm
μs + μm
New handover queueing model
REQ (time, TBVH, NDT, υ)1
1
REQ (time, TBVH, NDT, υ)2
2
n
REQ (time, TBVH, NDT, υ)n
Decision algorithm
No/partial contention
λ(1 – α)
Queue
Full contention (α)
Switch to another network
Server 1
Server m
μs + μm
μs + μm

Figure 5. Classical and proactive handover multi channel queueing system.

of the MN in the next network; hence, the time when the channel will be needed for communication and when a MN will release the channel due to mobility is as shown below:

- MN_A needs channel at $(t_c + TBVH)_A$.
- MN_A releases the channel at $(t_c + TBVH + NDT_{nxt})_A$.
- MN_B needs channel at $(t_c + TBVH)_B$.
- MN_B releases the channel at $(t_c + TBVH + NDT_{nxt})_B$.

Based on the channel request and holding time of MN_A, there are three possible contention possibilities that affect MN_B in this scenario.

No Contention — The channel release time of MN_A is less than the channel need time of MN_B. Hence, there is no contention as MN_B needs the channel after MN_A has finished using the channel.

Partial Contention — The channel release time of MN_A is less than the channel release time of MN_B. This means that MN_A uses the channel first. However, MN_A releases the channel while MN_B can still use the channel; hence, there is partial contention.

Full Contention—The channel release time of MN_A is greater than the channel release time of MN_B. In this scenario MN_A uses the channel and releases the channel after MN_B no longer needs the channel as MN_B has moved out of the range of the network. Hence, MN_B never gets access to the channel; this is called full contention.

Impact of Full Contention

In the event of full contention, MN_B will not get the channel from the next network range. If this total contention can be identified and reported before MN_B reaches the next network range, the contention can be signaled to MN_B, and MN_B can therefore use other available communication technology instead of waiting for a channel that is never available. For no or partial contention, MN_B can be signaled that it will get to use the channel and hence can queue for service. This approach should result in better network performance.

Proactive Queuing Approach for Handover in Mobile Environments

In this section we consider a simple scenario to explore the new proactive handover mechanism. In classic soft handover the MN will be placed in the queue to be served as shown in Fig. 5, that is, waiting for the channel to get an opportunity to communicate. This queuing model is commonly used to analyze mobile networks [9]. The server, in this case the channel mechanism, uses a first in first out (FIFO) service discipline, and requests are placed in the queue if the server is busy. Since the MNs are moving at a velocity and waiting for the channel, there is a probability that the MN will not get a channel due to mobility. For the MN that is being served, there is a possibility that it can also leave the network partially served due to mobility. Therefore, the rate at which the MN might leave the system due to mobility is denoted as μ_m. λ is the arrival rate of the request. μ_s is the rate at which the requests are being served. Thus, the overall service rate (i.e., the rate at which mobile nodes leave this network) varies as any MN may leave the queue without being served.

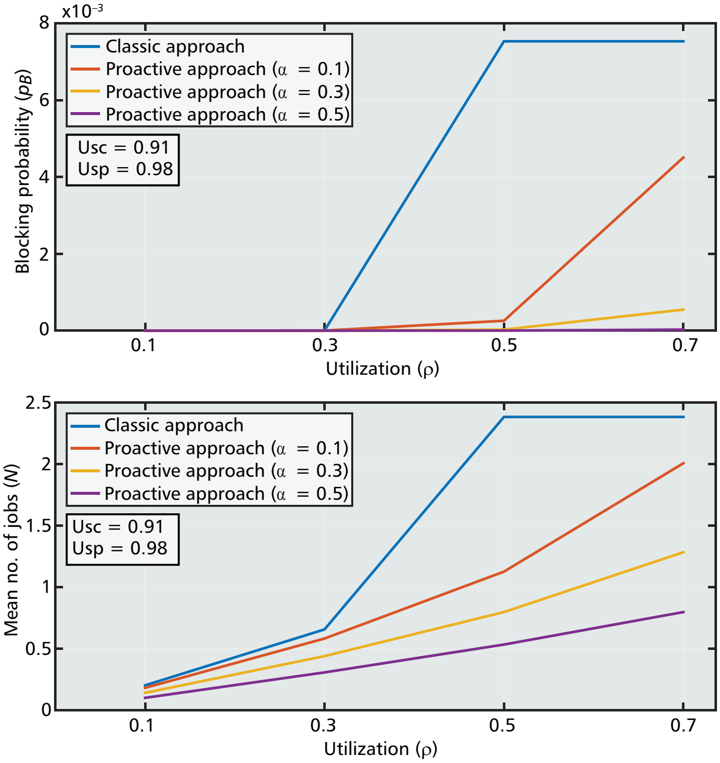


Figure 6. Two server: classical vs proactive approach (30m/s).

In our proactive approach, also shown in Fig. 5, the decision algorithm based on the contention analysis as described above will decide whether the node will be admitted to the queue. This ensures that nodes do not wait unnecessarily and leave the queue unserved because of mobility. Thus, all channel requests allowed into the queue will eventually be served, so the requests in the queue will not leave the queue due to mobility. However, only the request that is being served can leave the system due to mobility. Hence, the service rate is $\mu_s + \mu_m$.

We define α as the percentage of calls dropped due to contention. For the purpose of our analytical model we assume that α is constant. It is assumed that the rejected request time in the system is zero. Since requests are rejected from entering the queue due to contention, the arrival rate is $\lambda(1 - \alpha)$. If α is independent of $(\mu_s + \mu_m)$, the queue can be treated as a normal $M/M/1/K$ where K is the maximum number of packets in the system.

The key parameters used in comparing the two models were the use of two servers and a velocity of 30 m/s. We assume a mixed traffic pattern where on average a minimum of 2 slots of 0.5 ms as in Long Term Evolution (LTE) and described in [10]. Therefore, we use a conservative value of a service rate μ_s of 4000 packets/s.

The analytical results for both the classic and proactive handover approaches using a two-channel system are presented as graphs in Fig. 6 for the velocity 30 m/s. The graphs clearly show that the proactive approach works far better than the classic approach in terms of blocking probability (p_B) and mean number of jobs (N).

It has been shown how the cumulative probability approach is a better mechanism for estimating these values. Furthermore, based on realistic TBVH and NDT values it has been shown how these can be used for proactive channel allocation.

CONCLUSION

In this article we have presented a new VANET Testbed which is being deployed at Middlesex University, London. In addition, we have shown that to accurately calculate useful values of TBVH and NDT, a probabilistic approach based on accurate propagation models from a real testbed is required. It has been shown how the cumulative probability approach is a better mechanism for estimating these values. Furthermore, based on realistic TBVH and NDT values, we have shown how these can be used for proactive channel allocation.

REFERENCES

[1] G. E. Mapp *et al.*, "Y-Comm: A Global Architecture for Heterogeneous Networking," *Proc. 3rd Int'l. Conf. Wireless Internet*, ICST, Brussels, 2007, pp. 22:1–22:5.

[2] T. Yamamoto and S. Konishi, "Impact of Small Cell Deployments on Mobility Performance in LTE-Advanced Systems," *Proc. 24th IEEE Int'l. Symp. Personal, Indoor and Mobile Radio Commun.Wksps.*, 2013, Sept 2013, pp. 189–93.

[3] K. Sjöberg *et al.*, "Measuring and Using the RSSI of IEEE 802.11p," 2010.

[4] P. Fuxjager *et al*.,, "IEEE 802.11 p Transmission Using Gnuradio," *Proc. IEEE 6th Karlsruhe Wksp. Software Radios*, 2010, pp. 83–86.

[5] C. Sommer, R. German, and F. Dressler, "Bidirectionally Coupled Network and Road Traffic Simulation for Improved IVC Analysis," *IEEE Trans. Mobile Computing*, vol. 10, no. 1, Jan. 2011, pp. 3–15.

[6] A. Ghosh *et al*., "Exploring Efficient Seamless Handover in Vanet Systems using Network Dwell Time," *EURASIP J. Wireless Commun. and Networking*, vol. 2014, no. 1, 2014, p. 227.

[7] C. Ganan *et al*., "Analysis of Inter-RSU Beaconing Interference in VANETS," *Proc. Multiple Access Commun.*, ser. Lecture Notes in Computer Science, B. Bellalta *et al.*, Eds. Springer Berlin Heidelberg, 2012, vol. 7642, pp. 49–59.

[8] A. Vinel, D. Staehle, and A. Turlikov, "Study of Beaconing for Car-to-Car Communication in Vehicular Ad-Hoc Networks," *Proc. IEEE ICC Wksps.*, June 2009, pp. 1–5.

[9] W. Li, H. Chen, and D. Agrawal, "Performance Analysis of Handoff Schemes with Preemptive and Nonpreemptive Channel Borrowing in Integrated Wireless Cellular Networks," *IEEE Trans. Wireless Commun*., vol. 4, no. 3, May 2005, pp. 1222–33.

[10] E. Gemikonakli, *Multi-Service Traffic Modelling for Wireless Communication Systems*, Ph.D. dissertation, School of Sci. and Tech., Middlesex Univ., May 2014.

BIOGRAPHIES

ARINDAM GHOSH (A.Ghosh@mdx.ac.uk) received his B.Sc. (First Class Honours) degree in computer science from the University of Kent Canterbury (UKC), United Kingdom, in 2007. He was with Cisco Systems, Reading, United Kingdom (2007–2008) for a year in industry placement as an intern. He received his M.Sc. degree in computer networks management (with Distinction) from Middlesex University, London, in 2009. Currently, he is pursuing his Ph.D. degree in vehicular communication networks at the Department of Computer Science, Middlesex University. His research interests include prediction techniques for handover in VANETs for ubiquitous communication along with simulation and real-time testing and proactive handover approaches for DSRC radios in V2I communications.

VISHNU VARDHAN PARANTHAMAN received his B.Tech degree in information technology from Rajalakshmi Engineering College, Chennai, India, in 2010 and his M.Sc. degree in computer networks management (with Distinction) from Middlesex University in 2013. Currently, he is pursuing his Ph.D. degree in VANETs at the Department of Computer Science, Middlesex University. His research interests include in channel estimation and resource allocation for vehicular communications for DSRC radios in V2I communications.

GLENFORD MAPP received his B.Sc. (First Class Honours) from the University of the West Indies in 1982, his M.Eng. (Distinction in Thesis) from Carleton University, Ottawa, Canada, in 1985, and his Ph.D. from the Computer Laboratory, University of Cambridge, United Kingdom, in 1992. He then worked for AT&T Cambridge Laboratories for 10 years before joining Middlesex University as a principal lecturer. He is now an associate professor in the Department of Computer Science, School of Science and Technology, Middlesex University. His primary expertise is in the development of new technologies for mobile and distributed systems. He does research on Y-Comm, an architecture for future mobile communications systems. He also works on service platforms, cloud computing, network addressing, and transport protocols for local environments. He is currently focusing on the development of fast portable services that can migrate or replicate to support mobile users.

ORHAN GEMIKONAKLI received his B.Eng. degree in electrical engineering from the Eastern Mediterranean University, Famagusta, Cyprus, in 1984, and his M.Sc. degree in electronics and Ph.D. degree in electronic communications from King's College, University of London, in 1985 and 1990, respectively. He was a post-doctoral research associate working on trellis-coded high-level digital modulation schemes for satellite communications at King's College from 1989 to 1990. In 1990, he joined Middlesex University, where he is now a full professor in telecommunications. He has been an active committee member of the UKRI section Communication Chapter of the IEEE. His current research is in telecommunications, security, computer systems engineering, and performance evaluation of complex multi-server systems.

JONATHAN LOO received an M.Sc. degree in electronics (with Distinction) and a Ph.D. degree in electronics and communications from the University of Hertfordshire, United Kingdom, in 1998 and 2003, respectively. Between August 2003 and May 2010, he was a lecturer in multimedia communications at the School of Engineering and Design, Brunel University, United Kingdom. During this period, he was also the course director for the M.Sc. in digital signal processing. He is currently an associate professor in the Department of Computer Science, School of Science and Technology, Middlesex University. His research interests are in the area of multimedia communications, including visual media processing, video coding and transmission, wireless communications, digital signal processing, embedded systems and wireless network, security, cognitive radio, and software defined radio.

Call for Papers
IEEE Communications Magazine
Communications, Caching, and Computing for Content-Centric Mobile Networks

Background

The driving forces behind the exponential growth in mobile cellular network traffic have fundamentally shifted from being the steady increase in demand for "connection-centric" communications, such as phone calls and text messages, to the explosion of "content-centric" communications, such as video streaming and content sharing. The mobile cellular network architectures of today are, however, still designed with a connection-centric communication mindset. Moreover, the myriad technological advances proposed for beyond 4G and 5G mobile networks still mostly focus on capacity increase, which is fundamentally constrained by the limited radio spectrum resources as well as the diminishing investment efficiency for operators, and therefore will always lag behind the growth rate of mobile traffic. It can be argued that the logjam in cellular networks cannot be addressed by improving connection capability alone, but instead must be tackled by fundamentally addressing the underlying ineffectiveness of the current communication architecture for massive content delivery.

To cope with the shift to content-centric mobile cellular networks, a new design paradigm beyond the current connection-centric communication architecture is needed. In today's mobile networks, caching and computing capabilities are already ubiquitous, both at the base stations and on user devices themselves. How to effectively utilize these existing capabilities to address the needs for massive content distribution is a fundamental question the current and future research must address.

This Feature Topic aims to consolidate the timely and comprehensive overviews of the current state of the art in terms of fundamental research ideas and network engineering geared toward exploiting communications, caching, and computing (3C) for future content-centric mobile networks. The themes of interest within the scope of this Feature Topic include (but are not limited to):

- Theoretical Foundations: Fundamental limits of caching; coding for distributed storage; cooperative communications for content distribution
- Architectural Advances: Broadcast and cellular network convergence; content-centric resource virtualization, mobile edge computing, and cloud RAN
- Protocol Engineering: Distributed caching at wireless edge; pricing/incentive-based content distribution
- Modeling/Analytics: Content-centric traffic modeling
- Prototyping, testbeds, and field trials

Submissions

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words. 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 manuscript are posted at http://www.comsoc.org/commag/paper-submission-guidelines. Please send a pdf (preferred) or MSWORD formatted paper via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to Author Center. Follow the instructions there. Select "August 2016/3C for Content-Centric Mobile Networks" as the Feature Topic category for your submission.

Schedule for Submissions

- Manuscript Submission Due: January 1, 2016
- Acceptance Notification: April 1, 2016
- Final Manuscript Due: June 1, 2016
- Publication date: August 2016

Guest Editors

Prof. Meixia Tao (leading guest editor)
Shanghai Jiao Tong University, China
mxtao@sjtu.edu.cn

Prof. Wei Yu
University of Toronto, Canada
weiyu@ece.utoronto.ca

Dr. Wei (Andrew) Tan
Huawei Technologies, Shanghai, China
andrew.tan@huawei.com

Prof. Sumit Roy
University of Washington,
Seattle, United States
sroy@u.washington.edu

Prof. Tarik Taleb
Aalto University, Finland
talebtarik@gmail.com

Cooperation Strategies for Vehicular Delay-Tolerant Networks

João A. F. F. Dias, Joel J. P. C. Rodrigues, Neeraj Kumar, and Kashif Saleem

ABSTRACT

Vehicular communications are emerging as a promising technology to enable communications using vehicles as network nodes. VDTNs appear as a novel approach to enable services and applications where it is not possible to define an end-to-end path. To allow communications in such challenging environments, VDTNs rely for their operation on cooperation between network nodes, which contributes to increasing network connectivity and improving overall network performance. To accomplish such a task, nodes may be asked to share their constrained storage, bandwidth, and energy resources with one another. However, nodes may be unwilling to cooperate in order to save resources or due to selfish behavior. This kind of node severely affects the network functionality. This article gives an overview of the field, providing motivations, challenges, and an evaluation of the impact of cooperative measures on the performance of VDTN networks.

INTRODUCTION

In the last two decades, delay-tolerant networks [1, 2] have been exhaustively studied in order to allow communications in a wide range of extreme scenarios where a conventional end-to-end path may not be available. The vehicular delay-tolerant network (VDTN) [1] has emerged as a new delay-tolerant architecture that aims to provide innovative solutions for challenged vehicular communications. VDTNs exploit opportunistic contacts between vehicles (e.g., cars, buses, trams) and fixed infrastructures to provide connectivity to either urban or remote regions. However, they may also be used for other purposes, such as road safety, driver assistance, or even road traffic optimization.

The VDTN architecture (Fig. 1) follows the Open Systems Interconnection (OSI) reference model and the TCP/IP architecture, which allows the VDTN protocol stack to support services and functionalities performed by the two lower layers. Contrary to other delay-tolerant architectures, VDTNs define an IP over DTN approach by placing a data aggregation and de-aggregation layer, called a bundle layer, below the network layer. This allows the creation of large-size packets which aggregate IP datagrams that share common characteristics. These large packets are called bundles and define the VDTN protocol data unit. This architecture also gathers contributions from the optical burst switching (OBS) paradigm [3] since it considers an out-of-band signaling approach with separation of the control and data planes. At the control plane, nodes exchange signaling information (i.e., setup messages) using the bundle signaling control (BSC) layer. These messages may be used to make several routing decisions and to set up several parameters of the data plane. For example, according to the information collected at the control plane, nodes may decide to accept/ignore a contact opportunity. The control plane functions are always active, allowing node discovery. At the data plane, bundles are assembled and processed using the bundle aggregation and de-aggregation (BAD) layer. The data plane link is only activated when both nodes in contact are in condition to exchange data bundles. The use of different planes allows nodes to perform independently using their own layers, protocols, and even technologies.

The VDTN architecture considers three different types of nodes: mobile, relay, and terminal nodes. Mobile nodes store-carry-and-forward capabilities are exploited to forward bundles between network nodes (fixed or mobile). Terminal nodes are responsible for generating bundles (traffic sources) and processing incoming bundles from others (traffic skin). They are also responsible for providing connectivity to end users. In several scenarios where it is not possible to have terminal nodes, vehicles may act as terminal nodes. Relay nodes aim to increase the number of contact opportunities, reducing the impact of the high mobility of vehicles that leads to short contact durations. This situation limits the already limited amount of data to be transferred due to bandwidth and transmission ranges [4].

In spite of the improvements already achieved by the VDTN architecture, some of its particular features (e.g., sparse and sporadic contacts between network nodes, huge and inconstant

João A. F. F. Dias and Joel J. P. C. Rodrigues are with Instituto de Telecomunicações, University of Beira Interior.

Joel J. P. C. Rodrigues is with the University of Fortaleza.

Neeraj Kumar is with Thapar University.

Kashif Saleem and Joel J. P. C. Rodrigues are with King Saud University.

0163-6804/15/$25.00 © 2015 IEEE

message delivery delays, frequent network partitioning, and the possible absence of end-to-end connectivity) may compromise overall network performance. Due to these unique features, cooperation between network nodes assumes an important role. It is through this cooperation and their store-carry-and-forward capabilities that bundles travel from a source to their final destination.

This work's main goal is to overview and study the concept of cooperation and how it may influence the performance of VDTNs. The remainder of the article is organized as follows. The next section addresses the cooperation problem and discusses how cooperation functions can be implemented at the control and data planes; then we give an overview of the VDTN cooperation strategies to deal with cooperative/uncooperative nodes. After that, we analyze the impact of cooperative measures on the performance of VDTNs, and the final section concludes the article, providing a summary on cooperation in VDTNs.

Cooperation in Vehicular Delay-Tolerant Networks

In a fully cooperative scenario all network nodes are willing to cooperate. However, in a real scenario this is not a realistic assumption because nodes may reject communications with others in order to save resources or due to selfish/malicious behavior. In fact, one of the major issues faced by vehicular networks is related to resource limitations and their influence on the performance and capacity of the network. To ensure the success of data communications in such networks, it is very important to ensure cooperation between network nodes. Nodes use their resources to perform network functionalities and exchange bundles with each other. In cooperative environments, when a contact opportunity is available, nodes store and forward bundles from others. Such behavior allows DTNs to be resilient to network or individual node failures. Furthermore, when nodes are in a non-cooperative scenario, and no action is performed to motivate them to cooperate, they tend to exhibit selfish behavior. A node with selfish behavior may not be interested in store-and-forward bundles from others. Such behavior can be determined by resource limitations (e.g., storage, energy) or malicious conduct. A network that holds selfish nodes is a network with severely compromised performance.

Following the delay-tolerant paradigm, VDTNs also take advantage of cooperation between network nodes to improve overall network performance. Assuming the separation of the control and data planes, cooperation between network nodes may also be performed separately on these two planes in order to perform all network functions at each contact opportunity (Fig. 2). When network nodes encounter each other, they start to perform control plane functions, which results in the cooperative exchange of crucial information (e.g., location, speed, vehicle destination, energy, and buffer state). Such cooperative exchange allows node discovery and resource reservation to configure the data plane.

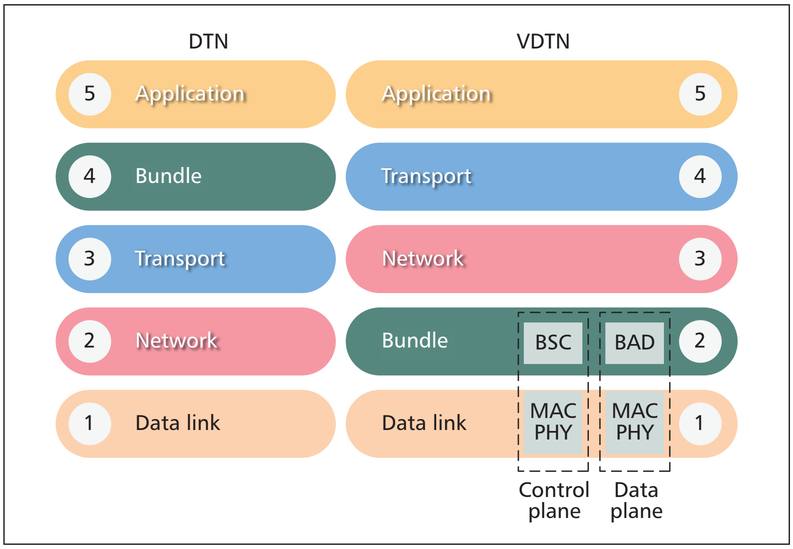


Figure 1. Illustration of the VDTN network architecture in comparison with DTN.

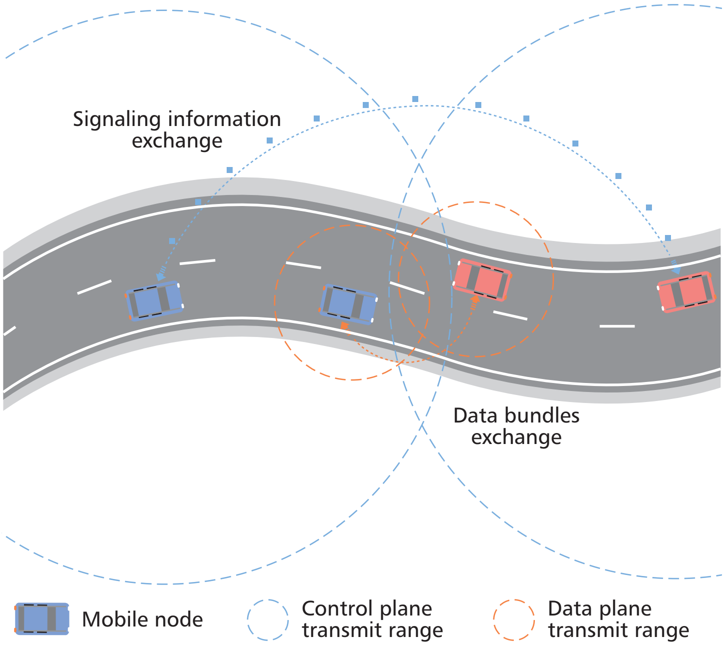

Signaling information exchange
Data bundles exchange
Mobile node
Control plane transmit range
Data plane transmit range

Figure 2. Illustration of the VDTN architecture out-of-band signaling.

For example, sharing geographical information and current speed allows nodes to predict the period of time they will be performing data plane functions or the period of time they will be in range of each other [5]. Afterward, the same information may also be used to configure data plane links to be active only during that period of time and to calculate the maximum number of bytes that can be transferred avoiding bundle fragmentation. Such an approach has a huge impact on power savings, which is very important in a network with energy constraints. The information cooperatively exchanged at the control plane also allows nodes to decide if a contact opportunity is suitable to accept or ignore. In other words, a node might decide to ignore a contact opportunity if the other one presents

To improve cooperation between network nodes, other information can also be exchanged, such as bundle delivery notifications, routing state information, or even nodes reputation scores. This can be a starting point to propose cooperative mechanisms and reputation systems that optimize network nodes performance.

energy or buffer space constraints. This approach prevents bundle dropping after a transmission process, which results in unnecessary energy consumption. To improve network performance, other kinds of information can be cooperatively exchanged at the control plane.

At the data plane, network nodes must cooperate in order to forward data bundles between source and destination. However, it is not possible to take for granted fully cooperative behavior, since nodes, in spite of their cooperative behavior, may prefer to schedule their own bundles first. Nodes may also be unwilling to unconditionally store all bundles sent by others in order to save buffer resources or maintain data integrity. Both situations may limit network performance.

To improve cooperation between network nodes, other information can also be exchanged, such as bundle delivery notifications, routing state information, or even nodes' reputation scores. This can be a starting point to propose cooperative mechanisms and reputation systems that optimize network nodes' performance.

Cooperation Strategies for Vehicular Delay-Tolerant Networks

As described above, VDTN operations depend on how nodes cooperate with each other. This dependence may raise many challenges if nodes diverge from the protocol for a given reason. The presence of non-cooperative nodes in VDTNs leads to degradation of the performance of the entire network. In the literature, there are several studies [6–8] that deal with this problem. Nevertheless, they cannot be directly applied to VDTNs. This section summarizes the advances already achieved on cooperation for VDTNs. It starts with the presentation of a reputation system that considers several incentive mechanisms, which reward or punish nodes for their cooperative behavior. Thus, a cooperative watchdog developed for VDTNs is presented and described in detail.

Reputation System

When it is not possible to create a fully cooperative scenario, it is important to afford the network strategies to deal with non-cooperative nodes. The VDTN reputation system [9] is a tool that is able to identify non-cooperative nodes by performing two distinctive steps (Fig. 3):

- Isolate and avoid contact with non-cooperative nodes.
- Monitor and motivate these nodes to cooperate.

To allow the reputation system operation, each node builds a reputation table containing all encountered nodes and their reputation score. Each time a contact opportunity is available, nodes exchange signaling information (e.g., location, destination, current speed) and statistics (e.g., energy or buffer status, reputation score). This signaling information is exchanged using the control plane at the connection setup phase. If nodes' reputation scores have changed since the last meeting, both scores are updated in the nodes reputation table.

After exchanging and processing the control information, the reputation system will perform its operation based on the chosen mode. If the reputation system is set up to isolate and exclude non-cooperative nodes from the network (mode 1), the reputation score of nodes is used to decide whether to accept or reject a contact opportunity. If the reputation score is higher than the network reputation threshold (Δ), the contact is accepted and nodes are able to forward bundles between each other, performing all data plane functionalities. At the end of data plane operations, nodes' reputation scores are recalculated taking into account their performance during the contact opportunity. In the second mode, the reputation system also classifies network nodes according to their reputation scores. However, contrary to the first mode, in this mode, misbehaving nodes start to be monitored for a predefined period of time (C_t). During this period, a monitored node is encouraged to share its own resources in order to continue consuming network resources. After the monitoring time, if it stills exhibiting non-cooperative behavior, it is excluded from the network. On the other hand, if the node starts sharing its own resources, it is rewarded by the reputation system with an increase of its reputation score to a value equal to the network cooperative threshold, and it is marked as a partially cooperative node. Being a partially cooperative node means that the reputation system will continue monitoring this node in order to verify that it continues to share its own resources. The reputation system workflow considering the two performing modes is illustrated in Fig. 3.

The reputation system considers several incentive strategies [9] to update the reputation score of each node. All strategies base their performance on two metrics: the number of deliveries and the number of dropped bundles. The reputation score of a node increases each time it successfully delivers a bundle to its final destination. On the other hand, when a node drops a bundle without exchanging it at least once, its reputation score decreases. *Simple increment simple decrement* (SISD), *double increment simple decrement* (DISD), and *simple increment double decrement* (SIDD) heuristics [9] share the same reasons to reward/punish nodes. The main difference between these schemes is how many nodes are rewarded/punished. The DISD scheme distinguishes itself from the SISD scheme by increasing nodes' reputation score by $2k$ units. The SIDD differs from the SISD scheme by punishing nodes in $2k$ units each time a bundle is dropped without having sent at least once. The *simple increment message hop decrement* (SIMHD) scheme [9] takes into account a bundle's path (i.e., hops between source and current node) to punish nodes when they drop a bundle without sending it at least once. This approach manages to punish selfish nodes in a more aggressive way than the previous schemes. A node's reputation score decreases $2k + h^*k$ units each time a bundle is dropped without being sent at least once. In this equation, h represents the number of bundle hops until the current node. This scheme intends to punish misbehaving nodes taking into account other nodes' efforts in forwarding bundles.

Conducted studies considering all four schemes [9] show that schemes that punish selfish nodes in a more aggressive way (SIMHD and SIDD) contribute to increasing overall network performance. This happens since such schemes manage to detect and avoid contacts with selfish nodes sooner.

Cooperative Watchdog System

The cooperative watchdog system (CWS) [10] aims to afford nodes with the capability to detect non-cooperative nodes without need for a centralized system. To perform such a task, each network node has a reputation score (α) which is used to determine the percentage of resources that may be shared with others. At the beginning of network operations all network nodes have a reputation score equal to 50, and over time it may change between 0 and 100. Nodes' reputation scores are updated and spread through the network, taking advantage of VDTN out-of-band signaling. While performing the control plane, nodes share information (e.g., number of relayed, dropped, and delivered bundles) about their performance in previous contact opportunities. This exchange allows nodes to evaluate each other, which will be stored in a neighbor reputation table that all network nodes maintain. The same information is also collected by the CWS in order to keep a record of the performance of each node in the network.

At the end of the data plane, the system updates the reputation score of all nodes that participated in the contact opportunity, taking into account three different kinds of scores: a node reputation score observed by its neighbors, a cooperative value assigned by the watchdog, and a node reputation score observed by the node itself [10]. Each of these scores is calculated independently using a different module. The neighbor's evaluation module aims to collect a node's neighbors' opinions. To perform this task, at the end of each contact opportunity, this module sends a request to *N* neighbors of a node asking them to share their opinion about their neighbor. When all the neighbors' answers are collected, the average of these scores is calculated, resulting in the node reputation score observed by its neighbors. To assign the cooperative value of a node, the CWS considers the impact of a node on the network. To calculate this impact, the number of relayed, delivered, and dropped bundles (information exchanged at the control plane) is considered. A node that has a normalized ratio between the number of delivered and dropped bundles closer to one is a cooperative node that contributes to increasing the overall performance of the network. For this reason, the CWS classifies this node as a cooperative node. On the other hand, a node that has the same ratio closer to zero is a possible non-cooperative node and is classified as a misbehaviing node.

Finally, the decision module uses an interface that communicates with the reputation system [9] which is deployed locally in each node. This approach aims to give nodes the ability to generate an opinion on their own performance. Having collected this score, the decision module calculates the new node reputation score based on all the scores calculated by the other modules and the importance imputed to each score. If the CWS is programmed to give more importance to the node opinion about itself, the score collected by the decision module will have a greater weight on the final equation. For example, if the system gives the node opinion a weight of 60 percent in the equation, 60 percent of the new reputation score will be given by the score calculated by the node itself and 40 percent will be given by the neighbors' score. The cooperative value assigned by the CWS will be used to reward or punish nodes for their cooperative behavior. A node's reputation score dropping below 20 causes the system to generate an alarm that will be spread by the network, informing all nodes to avoid contact with this specific node.

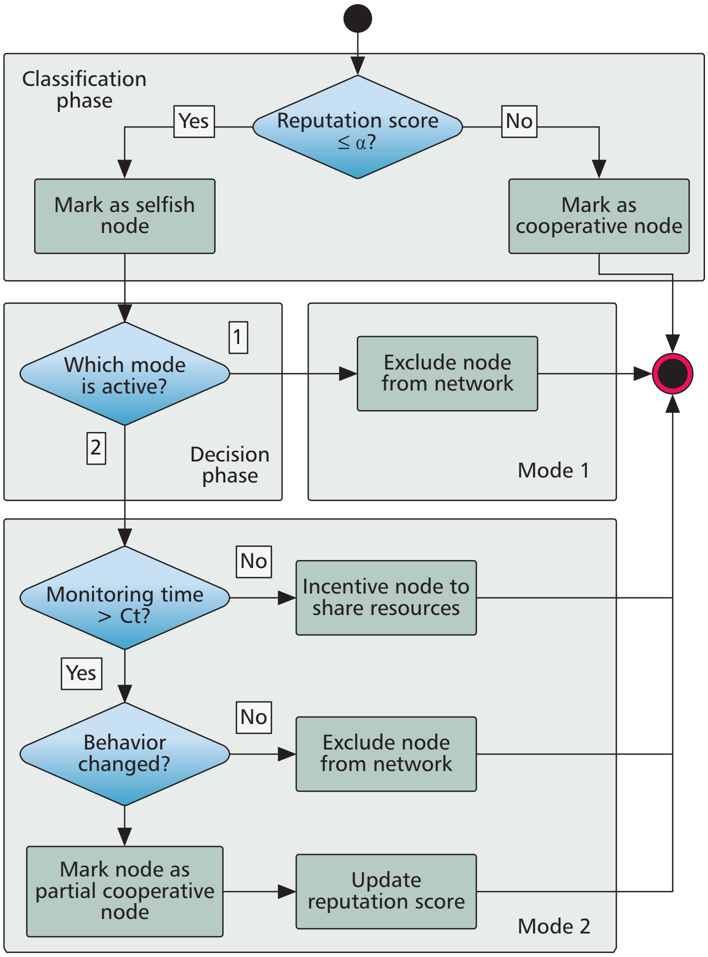


Figure 3. Illustration of the reputation system workflow deployed in VDTNs.

Performance Evaluation of Cooperative Strategies on VDTNs

In this section the impact of the above cooperation strategies on the performance of VDTNs are compared and analyzed. To perform this task, a simulation scenario was created in the

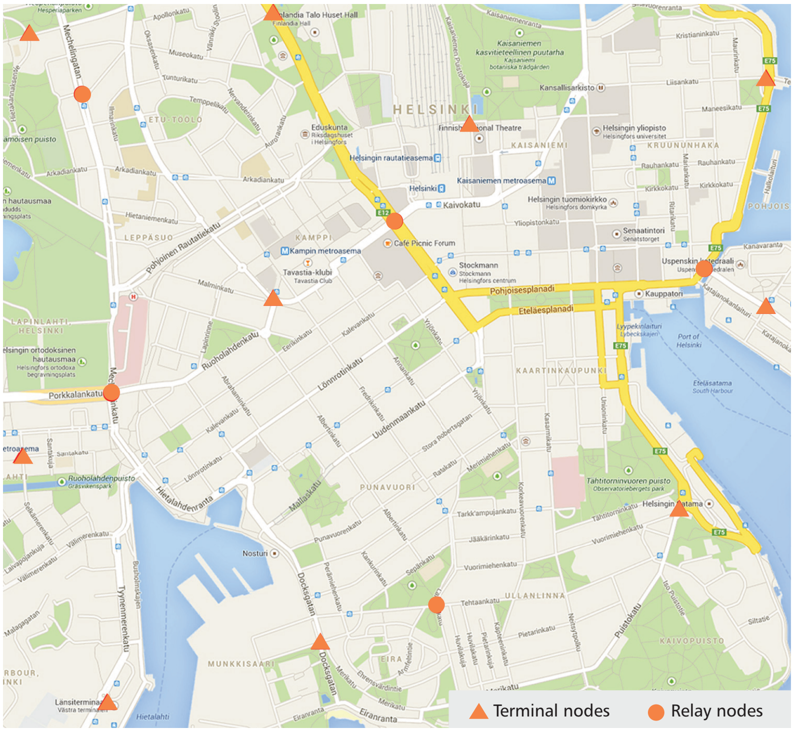


Figure 4. Simulation scenario representing a part of Helsinki, Finland, and the location of terminal and relay nodes.

VDTN simulator (VDTNsim) [11], representing a part of the city of Helsinki, Finland (Fig. 4). This tool was used because it was developed to emulate the most important principles of the VDTN architecture. In this section, the considered scenario is described, and a detailed analysis of the conducted performance studies is presented focusing on the impact of the above two strategies on the overall network performance when non-cooperative nodes are present in the network.

Network Scenario

Considering the scenario illustrated in Fig. 4 and a simulation period of 24 h, five relay nodes (each one with 200 MB of buffer capacity) were placed at the most important intersections. Ten terminal nodes (each one with 100 MB of buffer capacity) act as traffic source and traffic sinks. The number of mobile nodes (including cooperative and non-cooperative nodes) changes in the range of [30, 100]. All mobile nodes move along map roads with an average speed of 50 km/h and are equipped with a buffer with 50 MB capacity. To perform the out-of-band signaling, all network nodes are equipped with omnidirectional antennas, allowing a transmission range of 30 m and a transmission data rate of 6 Mb/s.

Bundles are uniformly generated in the range of [25, 35] s with a size uniformly distributed in the range of [50 kB, 750 kB]. Bundles have 360 min of time to live (TTL) to reach their final destination (i.e., another terminal node selected at creation time) or they will be dropped. Bundles may also be dropped due to buffer congestion. The binary version of the Spray and Wait routing protocol [12] (assuming eight bundle copies) was chosen as the underlying routing scheme.

To measure and evaluate the strategy's performance, the number of unique bundles that have been successfully delivered to destination nodes (bundle delivery probability) and the protocol overhead ratio were considered.

This network setup is intended to measure the efficiency of the two proposed mechanisms when deployed in an urban scenario with different levels of traffic (30 to 100 mobile nodes moving at an average speed of 50 km/h), to evaluate how both systems deal with bundle storage and TTL constraints (mobile nodes have a buffer capacity of 50 MB and have to deliver bundles in less than 360 min), and also to demonstrate the importance of the approach introduced by both of the systems in dealing with the presence of selfish nodes in the network.

Impact on Bundle Delivery Ratio

To analyze the performance of the two strategies, this study starts with analysis of the bundle delivery ratio. To do this, the percentage of non-cooperative nodes changes between 10 and 50 across all the experiments (30 for each point). This kind of node is discarded from the network when its reputation score is below the network reputation threshold (Δ), which was set to 20 at the beginning of network operations. As may be observed in Fig. 5, both strategies contribute to increase the number of unique delivered bundles when compared to an approach where no action is taken against non-cooperative nodes. Performing an analysis of the same figure allows one to conclude that the reputation system outperforms the CWS. This can be confirmed by comparing the two approaches performing with the same percentage of non-cooperative nodes. For example, when 50 percent of the mobile nodes are non-cooperative, the reputation system manages to increase the bundle delivery probability in approximately 5, 4, 6, 3, 3, 5, 3, and 2 percent (for a number of mobile nodes equals to 30, 40, 50, 60, 70, 80, 90, and 100, respectively). The reason this happens is because the reputation system manages to detect non-cooperative nodes more quickly, limiting their negative impact on the network. This particular feature is very important in scenarios with low node density. In such scenarios forwarding a bundle to a non-cooperative node could mean that this bundle may be dropped without a sufficient number of copies in the network to reach its final destination.

Impact on Protocol Overhead

The protocol overhead ratio represents how many "extra" copies of a bundle are needed to deliver it to its final destination. As may be seen in Fig. 6, both approaches manage to decrease the amount of "extra" bundles when compared to the approach where no action is taken against misbehaiving nodes. This is achieved by avoiding contact with such nodes, ensuring that bundle copies are transmitted only to nodes that are willing to forward them to others or to their final destination. This approach also decreases nodes' resource consumption since fewer transmissions are needed to deliver bundles to their final destination. Although both strategies contribute to decrease the protocol overhead ratio, the reputation system achieves better results. For example, considering the worst evaluate scenario (50 percent of non-cooperative nodes), the reputation system decreases the protocol overhead ratio in 2, 1, 3, 2, 2, 2, 2, and 3 bundles compared to the CWS. This happens due to the fact that the reputation system manages to decrease the number of bundle copies that are forwarded to misbehaving nodes, reducing the effort that cooperative nodes have to make to deliver bundles to their final destination (e.g., storing and sending more bundle copies). This situation also allows bundles to reach their final destination sooner.

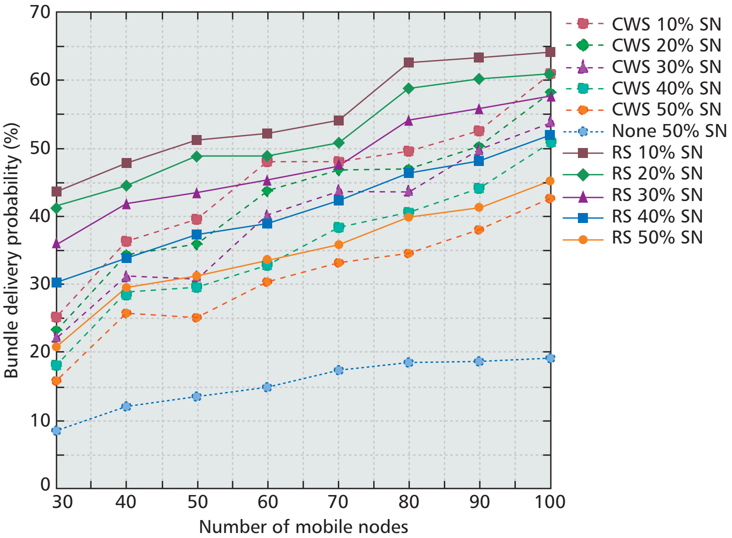


Figure 5. Bundle delivery probability as function of number of vehicles, from 30 to 100, considering the two strategies to deal with non-cooperative nodes in VDTN networks.

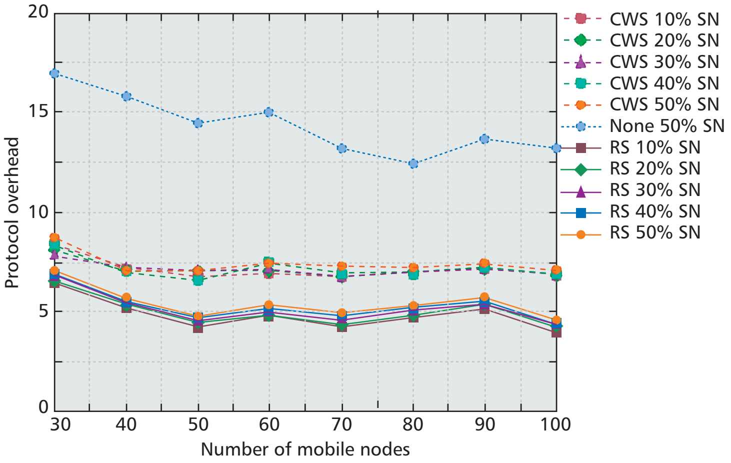


Figure 6. Protocol overhead as function of number of vehicles, from 30 to 100, considering the two strategies to deal with non-cooperative nodes in VDTN networks.

Conclusion

This article is focused on cooperation among nodes in VDTNs, especially on how this architecture deals with the presence of misbehaving nodes. Allowing misbehaving nodes in the network contributes to a huge decrease of network efficiency. To deal with these nodes, two strategies are presented, and their performance compared and analyzed. Both approaches try to balance the amount of resources shared with other nodes in order to not compromise nodes' integrity. The reputation system considers different incentive mechanisms to calculate nodes' reputation scores. Nodes' reputation scores are compared to the network reputation threshold (α) in order to classify nodes as cooperative or non-cooperative. Then the reputation system may decide whether to exclude or monitor non-cooperative nodes. The cooperative watchdog follows a similar approach, but contrary to the reputation system, which considers incentive mechanisms, it bases its operation on cooperative exchange of nodes' performance metrics and neighbor evaluation to classify nodes.

Conducted studies have shown the effectiveness of both approaches in improving the network performance when non-cooperative nodes are allowed in the network. They not only increase the probability of bundles reaching their final destination, but also manage to decrease the amount of wasted resources, resulting in power and energy savings, which is very important in networks with resource constraints like VDTNs.

For future research work, in order to obtain network performance improvements, several monitoring and management schemes may be developed.

Acknowledgments

This work has been partially supported by the *Instituto de Telecomunicaões*, Next Generation Networks and Applications Group (NetGNA), Portugal, by the Visiting Professor Program and King Saud University, and by National Funding from the FCT — *Fundação para a Ciência e a Tecnologia* through the UID/EEA/500008/2013 Project and the SFRH/BD/86444/2012 PhD grant.

References

[1] J. N. G. Isento *et al.*, "Vehicular Delay-Tolerant Networks — A Novel Solution for Vehicular Communications," *IEEE Intelligent Transportation Sys. Mag.*, vol. 5, no. 4, 2013, pp. 10–19; DOI: 10.1109/mits.2013.2267625.

[2] S. Burleigh *et al.*, "Delay-Tolerant Networking: An Approach to Interplanetary Internet," *IEEE Commun. Mag.*, vol. 41, no. 6, June 2003, pp. 128–36.

[3] J. J. P. C. Rodrigues, *Optical Burst Switching Networks: Architectures and Protocols*, Série Estudos de Engenharia (no 3): Fundação Nova Europa, Universidade da Beira Interior, Portugal, 2008; DOI: 10.1109/MCOM.2003.1204759.

[4] V. N. G. J. Soares, F. Farahmand, and J. J. P. C. Rodrigues, "Improving Vehicular Delay-Tolerant Network Performance with Relay Nodes," *Proc. 5th EuroNGI Conf. Next Generation Internet Networks*, Aveiro, Portugal, July 1–3, 2009.

[5] V. N. G. J. Soares *et al.*, "Exploiting Node Localization for Performance Improvement of Vehicular Delay-Tolerant Networks," *Proc. IEEE ICC 2010 General Symp. Sel. Areas in Commun.*, Cape Town, South Africa, May 23–27, 2010.

[6] L. Zhengming, L. Congyi, and C. Chunxiao, "On Secure VANET-Based Ad Dissemination with Pragmatic Cost and Effect Control," *IEEE Trans. Intelligent Transportation Sys.*, vol. 14, no. 1, 2013, pp. 124–35.

[7] T. Anantvalee and J. Wu, "Reputation-Based System for Encouraging the Cooperation of Nodes in Mobile Ad Hoc Networks," *Proc. IEEE ICC 2007*, Glasgow, Scotland, June 24–28, 2007, pp. 3383–88.

[8] L. Suk-Bok *et al.*, "Secure Incentives for Commercial

Ad Dissemination in Vehicular Networks," *IEEE Trans. Vehic. Tech.*, vol. 61, no. 6, 2012, pp. 2715–28; DOI: 10.1109/TVT.2012.2197031.

[9] J. A. F. F. Dias *et al.*, "A Reputation System to Identify and Isolate Selfish Nodes in Vehicular Delay-Tolerant Networks," *Proc. 13th Int'l Conf. ITS Telecommun.*, Tampere, Finland, Nov. 5–7, 2013, pp. 133–38.

[10] J. A. F. F. Dias *et al.*, "A Cooperative Watchdog System to Detect Misbehavior Nodes in Vehicular Delay-Tolerant Networks," Joint Special Section on Connected Vehicles — Advancements in Vehicular Technologies and Informatics, *IEEE Trans. Industrial Informatics* and *IEEE Trans. Ind. Electron*., 2015, DOI: 10.1109/TIE.2015.2425357.

[11] V. N. G. J. Soares, F. Farahmand, and J. J. P. C. Rodrigues, "VDTNsim: A Simulation Tool for Vehicular Delay-Tolerant Networks," *Proc. IEEE Int'l. Wksp. Computer-Aided Modeling Analysis and Design of Commun. Links and Networks*, Miami, FL, Dec. 3–4, 2010, pp. 101–05.

[12] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and Wait: An Efficient Routing Scheme for Intermittently Connected Mobile Networks," *Proc. ACM SIGCOMM 2005 — Wksp. Delay Tolerant Networking and Related Networks)*, Philadelphia, PA, Aug. 22–26, 2005, pp. 252–59.

Biographies

João A. F. F. Dias received a B.Sc. degree in informatics engineering from the University of Beira Interior, Portugal, in 2009. In 2011 he received his M.Sc. degree in informatics engineering from University of Beira Interior. Currently, he is a Ph.D. student in informatics engineering at the University of Beira Interior under the supervision of Prof. Joel J. P. C. Rodrigues. He is also a Ph.D. student member of the Instituto de Telecomunicações, Portugal. His current research topics include vehicular networks, delay-tolerant networks, and vehicular delay-tolerant networks. He has been an author or co-author of 21 conference papers, 10 international journal publications, and 2 technical reports.

Joel J.P.C. Rodrigues (joeljr@ieee.org) [S'01, M'06, SM'06] is a professor in the Department of Informatics of the University of Beira Interior, Covilhã, Portugal, and a senior researcher at the Instituto de Telecomunicações, Portugal. He received the Academic Title of Aggregated Professor from the University of Beira Interior, the Habilitation in computer science and engineering from the University of Haute Alsace, France, a Ph.D. degree in informatics engineering and an M.Sc. degree from the University of Beira Interior, and a five-year B.Sc. degree (licentiate) in informatics engineering from the University of Coimbra, Portugal. He is the leader of the NetGNA Research Group (http://netgna.it.ubi.pt), President of the scientific council at ParkUrbis Covilhã Science and Technology Park, Chair of the IEEE ComSoc Technical Committee on eHealth, Past Chair of the IEEE ComSoc Technical Committee on Communications Software, a Steering Committee member of the IEEE Life Sciences Technical Community, a Member Representative of the IEEE Communications Society on the IEEE Biometrics Council, and an officer of the IEEE 1907.1 standard. He is Editor-in-Chief of the *International Journal on E-Health and Medical Communications*, Editor-in-Chief of *Recent Advances on Communications and Networking Technology*, Editor-in-Chief of the *Journal of Multimedia Information Systems*, and an Editorial Board member of several journals. He has been General Chair and TPC Chair of many international conferences, including IEEE ICC, IEEE GLOBECOM, and HEALTHCOM. He is a member of many international TPCs and has participated in several international conferences' organization. He has authored or coauthored over 450 papers in refereed international journals and conferences, a book, and 2 patents. He had been awarded several Outstanding Leadership and Outstanding Service Awards by IEEE Communications Society and several best paper awards. He is a licensed professional engineer (as a Senior Member), a member of the Internet Society, an IARIA fellow, and a Senior Member ACM.

Neeraj Kumar received his Ph.D. in computer science engineering from Shri Mata Vaishno Devi University, Katra, India, and was a postdoctoral research fellow at Coventry University, United Kingdom. He is an associate professor in the Department of Computer Science and Engineering, Thapar University, Patiala, India. He has published more than 100 scholarly research papers in leading journals and conferences of IEEE, Elsevier, Springer, Wiley, and so on. Some of his research findings are published in top-cited journals such as *IEEE Transactions on Industrial Electronics*, *IEEE Transactions on Dependable and Secure Computing*, *IEEE Transactions on Intelligent Transportation Systems*, *IEEE Transactions on Consumer Electronics*, *IEEE Network*, *IEEE Communications Magazine*, *IEEE Wireless Communications*, *IEEE Internet of Things Journal*, *IEEE Systems Journal*, *Future Generation Computer Systems*, *Journal of Network and Computer Applications*, and *Computer Communications*. He has guided many research scholars to Ph.D. and M.E./M.Tech. degrees. His research is supported by funding from TCS and UGC.

Kashif Saleem is currently working as an assistant professor at the Center of Excellence in Information Assurance (CoEIA), King Saud University, Kingdom of Saudi Arabia. He received his Ph.D. (electrical engineering) and M.E. (electrical engineering — electronics & telecommunication) from Universiti Teknologi Malaysia in 2007 and 2011, respectively. His research interests include ubiquitous computing, biologically inspired algorithms, the Internet of Things, machine-to-machine communications, wireless mobile networks, wireless sensor networks, and mobile ad hoc networks. He has authored several research publications and handles ICT related funded research projects in the Middle East and European Union.

Call for Papers

IEEE Communications Magazine

Social and Mobile Solutions in Ad Hoc and Sensor Networking

Background

IEEE Communications Magazine announces, for the 10th Anniversary of the Series on Ad Hoc and Sensor Networks, a Feature Topic on Social and Mobile Solutions in Ad Hoc and Sensor Networking.

The Series on Ad Hoc and Sensor Networks of IEEE Communications Magazine intends to provide the latest developments in this very rich and exciting domain. The Series explores in depth the concept of ad hoc and sensor networking, highlighting the recent research achievements in the field, and also providing insight into theoretical and practical issues related to the development of these networks from different perspectives. This series offers a relevant forum for both academic and industrial research, at the same time covering the theory, practice, and state of the art of ad hoc and sensor networking.

The special focus of this FT aims to explore the ad hoc and sensor networks research and development related to social and mobile solutions: communications, applications, algorithms, and systems as well as services and computing. Thus, we are interested in articles that involve people, networked sensors, mobile phones, and/or sensing context, how they interact, and how they perform.

Both original research and review papers are welcome. Possible topics of social and mobile solutions in ad hoc and sensor networking include but are not limited to:

- Architectures and design
- Social and mobile systems
- Communication issues
- Pervasive computing
- Human mobility
- Analysis, simulation, and measurement
- Real experiment campaigns
- Social and mobile applications
- Mobile phone sensing
- Social and economic aspects

This list is not exhaustive; submissions related to new and interesting ideas relating broadly to social and mobile solutions in ad hoc and sensor networking are encouraged.

IEEE Communications Magazine is read by tens of thousands of Communications Society members. The papers will also be available on the Internet through Communications Magazine Interactive, the WWW edition of the magazine. Details about IEEE Communications Magazine can be found at http://www.comsoc.org/commag.

Submissions

Manuscripts must be submitted through the magazine's submissions web site at: http://commag-ieee.manuscriptcentral.com/. You will need to register and then proceed to the Author Center. On the manuscript details page, please select Ad Hoc and Sensor Networks Series from the drop-down menu.

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. Figures and tables should be limited to a combined total of six. The number of archivable references is recommended to not 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 manuscript are posted at http://www.comsoc.org/commag/paper-submission-guidelines. Please submit a pdf (preferred) or MS WORD formatted paper via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to Author Center. Follow the instructions there.

In your submission, please indicate in the sub-title that the paper is to be considered for this Feature Topic, and not as a regular submission to the Series on Ad Hoc and Sensor Networks, as follows:

PAPER TITLE
Paper submitted to the Feature Topic on Social and Mobile Solutions in Ad Hoc and Sensor Networking
Series on Ad Hoc and Sensor Networks

Schedule for Submissions

- Manuscript Submission Date: January 11, 2016
- Decision Notification: March 28, 2016
- Final Manuscript Due Date: April 25, 2016
- Publication Date: July 2016

Guest Editors

Edoardo Biagioni
University of Hawaii
esb@hawaii.edu

Xun Luo
Tianjin University of Technology
luo@tjut.edu.cn

Tao Tian
Qualcomm Incorporated
tao.tian@gmail.com

Silvia Giordano
University of Applied Science - SUPSI
silvia.giordano@supsi.ch

Tracy Camp
Colorado School of Mines
tcamp@ mines.edu

ADVERTISERS' INDEX

COMPANY	PAGE
IEEE Member Digital Library	Cover 4
Keysight	Cover 2
National Instruments	5
Rohde & Schwarz	3
Tutorial	Cover 3
Wiley	1

ADVERTISING SALES OFFICES

Closing date for space reservation: 15th of the month prior to date of issue

NATIONAL SALES OFFICE
James A. Vick
Sr. Director Advertising Business, IEEE Media
EMAIL: jv.ieeemedia@ieee.org

Marion Delaney
Sales Director, IEEE Media
EMAIL: md.ieeemedia@ieee.org

Mark David
Sr. Manager Advertising & Business Development
EMAIL: m.david@ieee.org

Mindy Belfer
Advertising Sales Coordinator
EMAIL: m.belfer@ieee.org

NORTHERN CALIFORNIA
George Roman
TEL: (702) 515-7247
FAX: (702) 515-7248
EMAIL: George@George.RomanMedia.com

SOUTHERN CALIFORNIA
Marshall Rubin
TEL: (818) 888 2407
FAX:(818) 888-4907
EMAIL: mr.ieeemedia@ieee.org

MID-ATLANTIC
Dawn Becker
TEL: (732) 772-0160
FAX: (732) 772-0164
EMAIL: db.ieeemedia@ieee.org

NORTHEAST
Merrie Lynch
TEL: (617) 357-8190
FAX: (617) 357-8194
EMAIL: Merrie.Lynch@celassociates2.com

Jody Estabrook
TEL: (77) 283-4528
FAX: (774) 283-4527
EMAIL: je.ieeemedia@ieee.org

SOUTHEAST
Scott Rickles
TEL: (770) 664-4567
FAX: (770) 740-1399
EMAIL: srickles@aol.com

MIDWEST/CENTRAL CANADA
Dave Jones
TEL: (708) 442-5633
FAX: (708) 442-7620
EMAIL: dj.ieeemedia@ieee.org

MIDWEST/ONTARIO, CANADA
Will Hamilton
TEL: (269) 381-2156
FAX: (269) 381-2556
EMAIL: wh.ieeemedia@ieee.org

TEXAS
Ben Skidmore
TEL: (972) 587-9064
FAX: (972) 692-8138
EMAIL: ben@partnerspr.com

EUROPE
Christian Hoelscher
TEL: +49 (0) 89 95002778
FAX: +49 (0) 89 95002779
EMAIL: Christian.Hoelscher@husonmedia.com

CURRENTLY SCHEDULED TOPICS

	PUBLICATION DATE	MANUSCRIPT DUE DATE
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
SOCIAL AND MOBILE SOLUTIONS IN AD HOC AND SENSOR NETWORKING	JULY 2016	JANUARY 11, 2016
SDN USE CASES FOR SERVICE PROVIDER NETWORKS	OCTOBER 2016	JANUARY 31, 2016

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

DECEMBER 2015

IEEE COMSOC live WEBINAR

Realizing 5G: Device-Centric Design in a New Spectral Landscape

Thursday, 3 December 2015 • 12:00 PM EST, 9:00 AM PST, 17:00 UTC/GMT

Fifth generation (5G) mobile networks will increase throughput from 150 megabits/second to 10 gigabits/second, expand the available spectrum into the millimeter-wave domain between 6 GHz and 300 GHz, and reduce latency from 10 milliseconds to 1 millisecond—fast enough to control real-world devices like automobiles in real time.This webinar will look at device-centric approaches to network design, which move the focus from the base station toward the edge.

IEEE ComSoc content sponsored by:

KEYSIGHT TECHNOLOGIES GL Communications Inc. NATIONAL INSTRUMENTS WILEY

Bringing Connected Vehicle Technology to the Next Level (Part 1)

AVAILABLE ON DEMAND

The Connected Vehicle market is on the verge of major changes with ubiquitous connectivity. To reach that potential, a number of challenges still need to be addressed. As the first part of a two-part series, a panel of experts will examine a number of these challenges, including mobility, security, V2V and V2I.

Bringing Connected Vehicle Technology to the Next Level (Part 2)

Tuesday, 8 December 2015 • 2:00 PM EST, 11:00 AM PST, 19:00 UTC/GMT

In this the second part of a two-part series, a panel of experts will examine the technology building blocks to support ADAS and Autonomous Driving, including high fidelity maps, sensor and data fusion, multi-modal human machine interaction and V2I and V2V communications. After the first part of this webinar series, the potential and challenges of a connected vehicle market will become clearer through an analysis of the key technologies.

Sponsor content provided by:

INTERDIGITAL.

Limited time only at >> www.comsoc.org/webinars

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

IEEE COMMUNICATIONS SOCIETY® IEEE®

COMMUNICATIONS STANDARDS

A Supplement to IEEE Communications Magazine

DECEMBER 2015

www.comsoc.org

- IoT/M2M from Research to Standards: The Next Steps
- Research and Standards: Advanced Cloud and Virtualization Techniques for 5G Networks

IEEE
IEEE COMMUNICATIONS SOCIETY
A Publication of the IEEE Communications Society

We're here to help you write your 5G future.

Unprecedented experience in wideband mmWave, 5G waveforms, and Massive MIMO design.

The world of wireless communications is about to change. Again. The fifth generation—5G—will mean "everything, everywhere, and always connected." If you're on the cutting edge of this emerging technology, we can help you. We have expertise in all areas of 5G research and development, including wideband mmWave, radio spectrum, ASIC, antenna technologies, and network architecture. So from design simulation and verification to wideband signal generation and analysis, from component characterization to optical solutions, we've got you covered.

HARDWARE + SOFTWARE + PEOPLE = 5G INSIGHTS

Keysight engineers are active in the leading 5G forums and consortia

Keysight engineers are keynote speakers at 5G conferences and key contributors in top technical journals

Applications engineers are in more than 100 countries around the world

KEYSIGHT
TECHNOLOGIES

Unlocking Measurement Insights

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

Keysight Technologies
Implementing a Flexible Testbed for 5G Waveform Generation and Analysis

White Paper

Keysight Authors:
Greg Jue, Applications Development Engineer/Scientist
Sangkyo Shin, Product Marketing/App Development

KEYSIGHT
TECHNOLOGIES

USA: 800 829 4444 CAN: 877 894 4414

COMMUNICATIONS STANDARDS

December 2015

Supplement Editor-in-Chief
Glenn Parsons

Managing Editor
Jack Howell

Standards News Contributors
Doug Arnold • Peter Ashwood-Smith • Alfred Asterjadhi • John C. Eidson • Tim Godfrey
Robin Mersh • Christopher Price • Yongho Seok • Michael Johas Teener

IoT/M2M from Research to Standards: The Next Steps (Part II)

Research & Standards: Advanced Cloud and Virtualization Techniques for 5G Networks (Part II)

5G Internet of Things

Glenn Parsons

Many of the applications and use cases that drive the requirements and capabilities of 5G are about end-to-end communication between devices. In order to distinguish them from the more human-centric wireless applications such as mobile telephony and mobile broadband, these applications are often labeled machine-type communication (MTC) or more broadly as the Internet of Things (IoT). Improved provisioning, device management, and service enablement will bring a wider range of potential IoT applications. Some predictions forecast that there will be 15 billion connected MTC devices by 2021, a nearly 40 fold increase over the number of currently deployed MTC devices. Although spanning a wide range of different applications, IoT can be divided into two main categories, massive MTC and critical MTC, depending on their characteristics and requirements. Each brings its own challenges and requirements for standardization. To create a truly networked society requires market driven standardization and the deployment of standardized technology for 5G IoT devices, networks, and software.

The importance of standards to the work and careers of communications practitioners is the basis of this publication. It is a platform for presenting and discussing standards related topics in the areas of communications, networking, research, and related disciplines. This supplement on Communications Standards contains continuations of two related feature topics on 5G and IoT. JaeSeung Song and his editorial team provide the second part of the feature topic "IoT/M2M from Research to Standards: The Next Steps" that draws the connection between 5G and IoT. Tarek Taleb and his editorial team provide the second part of the feature topic on "Research & Standards: Advanced Cloud & Virtualization Techniques for 5G Networks" exploring how these techniques can benefit 5G IoT. Each editorial team will introduce their feature topic papers in more detail.

Readers will notice the ongoing Commentary section with a recurring view from the IEEE-SA President and the IEEE Standards Education activity. This time we are happy to also include a message from the IEC Standards Management Board. The Standards News section offers the current status of standards work in various SDOs relevant to 5G and IoT, as well as pointers to SDO material. I trust that the reader will find these informative and illustrative of the fundamental role standards play in the communications networking ecosystem.

Looking forward to 2016, the *IEEE Communications Magazine* Standards Supplement will continue quarterly publication and each issue will be "anchored" around a topic of current market relevance to drive focus. The next issue will contain a feature topic on "Semantics for Anything-as-a-Service" and give us insight into information and data modelling that support the semantics needed for end-to-end service management. Proposals for future standards feature topics are welcome as we look forward to the Standards Supplement evolving into a stand-alone magazine in 2017.

Biography

Glenn Parsons [SM] (glenn.parsons@ericsson.com) is an internationally known expert in mobile backhaul and Ethernet technology. He is a standards advisor with Ericsson Canada, where he coordinates standards strategy and policy for Ericsson, including network architecture for LTE mobile backhaul. Previously, he has held positions in development, product management and standards architecture in the ICT industry. Over the past number of years, he has held several management and editor positions in various standards activities including IETF, IEEE, and ITU-T. He has been an active participant in the IEEE-SA Board of Governors, Standards Board and its Committees since 2004. He is currently involved with mobile backhaul standardization in MEF, IEEE and ITU-T and is chair of IEEE 802.1. He is a Technical Editor for IEEE Communications Magazine and has been co-editor of several IEEE Communications Society Magazine feature topics. He graduated in 1992 with a B.Eng. degree in electrical engineering from Memorial University of Newfoundland.

Introduction to IEEE Standards

By Bruce Kraemer, President, IEEE Standards Association

http://standards.ieee.org/develop/index.html

Some readers of these standards articles are intimately familiar with the procedures used by the Standards Association. Some, in fact, are responsible for not only applying them but for writing them. But I suspect there are some readers who might be interested in some brief operational insight.

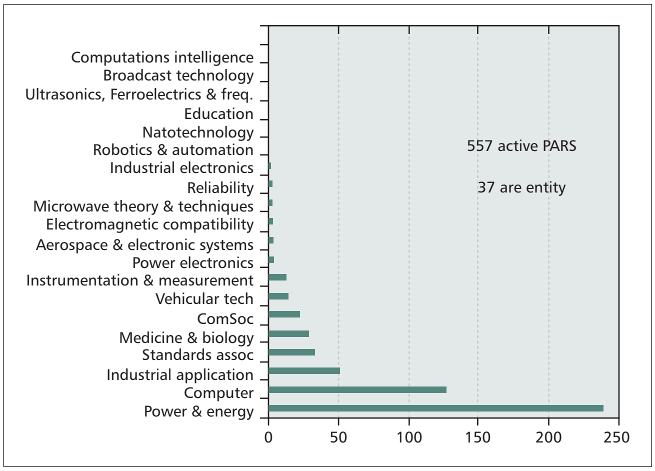
What does IEEE call a Standard? While the term Standard is used throughout the world, within IEEE it is a term that requires some further disambiguation. A Project Authorization Request (PAR) can indicate the published standard document will be one of four types:

- Standards: documents with mandatory requirements.
- Recommended practices: documents in which procedures and positions preferred by the IEEE are presented.
- Guides: documents in which alternative approaches to good practice are suggested, but no clear-cut recommendations are made.
- Trial-use documents: publications in effect for not more than two years. They can be any of the categories of standards publications listed above.

Starting and Completing Projects: A PAR is a standardized form used to describe the type of standards project being proposed by a sponsor (https://mentor.ieee.org/etools_documentation/dcn/12/etools_documentation-12-0006-MYPR-new-par-form-blank.docx). A PAR that has been submitted is reviewed by the members of the New Standards Committee (NesCom). When NesCom and the sponsor agree on the contents, it is recommended for approval and forwarded to the 35 members of the Standards Association Standards Board (SASB) for final review and approval.

Similarly, when the sponsor believes a draft standard document is ready for publication, it is submitted for review by the members of the Review Committee (RevCom). When RevCom confirms that the sponsor has fulfilled all requirements of the standard development process, the draft is recommended for publication and forwarded to the Standards Association Standards Board (SASB) for final approval.

When development of a draft standard is completed, it is published. IEEE currently lists more than 1000 standards as being published and active. By active, we refer to the standard as being currently relevant for use. There have been, and will continue to be, standards that were previously developed but subsequently deemed inappropriate for further use and converted to an in-active form. In the database these are labeled as "Superseded" or "Withdrawn."



Active PARs by sponsor, as of February 2015. (Source: IEEE Standards Association).

Standards Projects Statistics: IEEE often recites the fact that it contains 45 Societies and Technical Councils. We know that each of these has a designated engineering discipline such as computers, communications, electron devices, industry applications, etc. Some, but not all, of the societies sponsor standards development activities. The accompanying chart indicates that 19 societies (plus the Standards Association itself) were sponsoring 557 standard projects as of February 2015.

A couple items of note from the chart:

1. About 50 percent of all active projects are sponsored by the Power and Energy Society.

2. The rate of new project submissions has been very close to the rate of project completions and publications, so the "Active Project" quantity has remained stable at just over 550 projects for the past few years.

As stated above, the IEEE has written more than 1000 standards. Conformance tests have been left up to third party organization testing. This historical approach is changing and IEEE conformance tests are being created. In the next article I will highlight the IEEE Conformity Assessment Program (ICAP).

IEEE Standards Education

By Yatin Trivedi

In late 2014 the IEEE Board of Directors approved New Initiatives Committee (NIC) funding to create the IEEE Standards University, and the IEEE Foundation approved a related grant. The IEEE Standards University is a multi-track initiative intended to greatly expand IEEE's standards education content and resources for educators, students, and professionals focusing on the development and use of standards; the impact of standards on business; an understanding of patents and standards; the role of conformity assessment; and more. This joint initiative of IEEE Educational Activities and the IEEE Standards Association includes innovative elements, reimagined versions of existing elements that leverage content and experience, and coordinating elements that together present a comprehensive program that is greater than the sum of its parts.

We are happy to announce the availability of the new IEEE Standards University web platform beginning on 16 November 2015 at StandardsUniversity.org. The IEEE Standards University aims to expand the influence of IEEE Standards and benefit humanity and IEEE membership by (1) making standards education a viable reality at the university level by providing a critical mass of materials, and (2) modernizing the delivery method for educators, students, and professionals.

To align with the goals of the critical mass of teaching materials and new delivery methods, the IEEE Standards University consists of the following tracks:

- On-Line University Experience Track: new website developed with many improvements to greatly improve navigation and the user experience by compiling information and content from related standards education sites under one umbrella.
- Standards Education Video and eLearning Track: includes technical videos, interviews, a guest lecture series, and new standards-related eLearning courses.

•Publication Track: redesign of the IEEE *Standards Education e-Magazine.*

•Standards Simulation Game Track: the only standards development simulation game developed by standards experts.

•Workshop Track: expansion of successful face-to-face one-day training workshops for students, faculty, and/or professionals aimed at greatly increasing attendees' knowledge of standards and consensus-building.

•Massive Open Online Course (MOOC) Track: the MOOC, entitled "Innovation and Competition: Succeeding through Global Standards," is intended to help students develop a skillset around technical standards that enables them to compete more effectively in the global economy.

Creating the IEEE Standards University is a three-year project. Milestones for year one have been met. The new website has been built and includes the latest issue of the newly designed *Standards Education e-Magazine*. New standards videos were created focusing on the IEEE 802® family of standards and the National Electrical Safety Code (NESC®), and four new eLearning courses will also be released before the end of 2015. Titles of the courses include:

•How to Read a Standard.
•Ethics in Standards.
•Introduction to Conformity Assessment.
•IEEE Standard 1588™, IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems.

The work will continue in 2016 to include additional video and eLearning content and to create a new standards simulation game and a new type of workshop. Stay tuned for the updates and look for the MOOC in the spring of March-April 2016 on IEEEx.org!

IEC: The Home for Industry and Collaboration

By James E. Matthews III, Director, Technical Standards & Standards Policy, Corning Incorporated

Many IEEE members are aware of projects and activities in the IEC (International Electrotechnical Commission), but that awareness is usually confined to one area or activity directly related to their own work. The IEC covers a broad range of technical areas, working to provide International Standards and other deliverables in support of safety, efficiency, and compatibility of electrical, electronic, and communications devices and systems.

The IEC uses a process to produce Standards that are aligned with the WTO requirements through national representation in 167 countries across the world. The consensus based process that is the foundation for IEC work is a careful balance between speed in addressing market needs and consensus to build meaningful and useful results.

IEC technical work is carried out in working groups that are managed by technical committees (TCs) or subcommittees (SCs) using experts from stakeholder groups appointed by their National Committees. Each of the 177 TCs and SCs has a defined scope and field of work that is managed by the IEC Standardization Management Board (SMB).

Technology convergence and innovation are continuously resulting in new approaches and applications that were previously not possible. This also directly impacts and blurs traditional boundaries between individual working groups. For this reason the SMB continuously reviews work programs and areas of activities and redefines boundaries as needed. IEC experts also frequently interact with or even participate as experts in other organizations, including ISO, ITU, IEEE, and many others. The fact is, in any given technical area, the pool of experts is finite. It is important to recognize that the industries and organizations that participate in standardization have a limited number of people and resources at their disposal. For this very reason, the IEC and IEEE are in constant communication and have engaged on a path of collaboration.

In 2002 the IEEE and IEC agreed to a Dual Logo program, in which the publication of one organization could be adopted by the other and subsequently published jointly with the logo of both organizations. This was a great step forward to avoiding duplication and divergence in markets and technologies. It let both organizations build on some of the great work that was already achieved. In 2008 the effort was expanded to include the possibility of joint development of a single document bringing together experts from both organizations. A guide to these activities can be found at **http://www.iec.ch/about/brochures/pdf/tools/IEC_IEEE_Cooperation.pdf**

Recently, teams from both organizations met and fine-tuned the procedures based on experience to date. This has resulted in a new updated guide, which will be published shortly. Many IEC and IEEE groups are using this approach to build on established technical work as well as blending the expertise and best know-how of both organizations across a number of technical sectors.

Although the IEC is more than 100 years old, it is continuing to evolve in response to the needs of our stakeholders. We have clearly recognized, as expressed in our Masterplan, that we need to extensively collaborate with other groups and organizations as well as find new ways of addressing gaps in our work, if we want to remain the home of industry. While the overall approach of technical committees works well, their shortcoming is that they are focused on addressing specific product and technology areas. When faced with situations that involve the interaction of multiple technologies or product areas, a different approach is needed. For exactly this reason a new type of structure was put in place that is able to address work at the systems level. At first a Systems Evaluation Group (SEG), which welcomes participants from many different organizations, is tasked with defining system boundaries, identifying and mapping existing work, needs and gaps in a specific area, as well as developing a plan to address those needs. Thereafter a Systems Committee (SyC) may be established. A SyC is generally tasked with developing use cases, work plans, and standards to complement and support the work of a number of existing technical committees. Often, complex areas such as smart grids/smart energy, smart cities, and many more involve not only the work of the IEC, but they also require alignment with work that is accomplished by other groups. Representatives of relevant IEEE groups participate and collaborate regularly in all IEC SEGs and SyCs.

It is notable that over the last year, communication between the two organizations has grown at the management level in addition to the working levels. The two organizations are now working to inform each other about new projects and initiatives. An IEC representative has been a regular participant in the IEEE Standards Association Standards Board, and an IEEE representative has been a regular participant in the IEC SMB. This has also promoted better and more open communication at and between all levels of the two organizations.

It is important to note that these options for collaboration are available to all IEEE and IEC groups to use. Experts and industries involved can have the confidence that their hard work and resources are not duplicated or wasted on divergent efforts. IEEE and IEC experts now have more tools at their disposal to approach complex and challenging standardization projects.

Biography

Jim Matthews III is a Vice President of the IEC and Chairman of the IEC Standardization Management Board. He is the Past President of the U.S. National Committee of the IEC, and has worked as an expert in groups in IEC, ITU, IEEE, and many other standards organizations. He has worked for Corning Incorporated for more than 34 years, and is the Director of Technical Standards and Standards Policy. He has been an IEEE member since joining as a student and holds a BEE and MSEE from Georgia Tech.

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.

IEEE P1588 Working Group

Doug Arnold (Meinberg) and
John C. Eidson (Calnex), Co-Chairs

The second edition of IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems was published in 2008. Since that time the standard has been adopted by several industries as the approved technology for distributing time within systems.

For example, ITU-T Standards G8265.1, 8275.1, and 8275.2 all specify the use of IEEE 1588 for use in distributing time and/or frequency within telecom networks. Telecommunication operators in Europe and China have large deployments, several million nodes supporting IEEE 1588. Certain segments of the industrial automation, data acquisition, and military communities also have major deployments. In addition, the technology is being used or considered in applications in the finance and television and motion picture industries, among others.

Since 2008 there have been many suggestions for improvements and additional functionality. These suggestions will be incorporated into the next edition currently under preparation by the P1588 Working Group, with publication expected in late 2017. The principal items under consideration are:

- Introducing a layered model enabling IEEE 1588 to more easily integrate other time distribution technologies into systems deploying IEEE 1588. For example, provisions for incorporating links using IEEE 802.11 are under consideration.
- Adding time transfer using the techniques developed in the White Rabbit project at CERN. This technology is a combination of IEEE 1588 and layer 1 frequency transfer (sometimes called synchronous Ethernet) and should extend the accuracy and precision capabilities of IEEE 1588 well into the sub-nanosecond range (see https://en. wikipedia.org/wiki/The_White_Rabbit_Project).
- Adding provisions for addressing time transfer security issues. These provisions will be a combination or use of existing security mechanisms, architectural and deployment practices, and specifications specific to IEEE 1588.
- Addressing the management of IEEE 1588 systems. The existing TLV-based system of IEEE 1588 will be retained. In the longer term P1588 intends to collaborate with IEEE 802.1, IETF, ITU-T, and other organizations to specify a YANG model of the common elements of IEEE 1588 with provision for augmenting the model to handle extensions.

P1588 encourages participation by those interested in network time transfer technology. Please see our public website for details on participation (https://ieee-sa.centraldesktop.com/1588public/).

IEEE 802.11ah: Sub 1 GHz License Exempt Operation

Alfred Asterjadhi (Qualcomm), IEEE 802.11ah Vice-Chair, Yongho Seok (Newracom), IEEE 802.11ah Chair

The IEEE 802.11ah Task Group (TG) is developing a standard specification for targeting the Internet of Things (IoT) and extended range (ER) applications. The TG started the standardization activity in November 2010, and is currently in the last phase of the IEEE ballot procedure. The publication of the IEEE 802.11ah amendment is expected in July 2016.

The IoT is the next major growth area for the wireless industry, with applications across home and industrial automation, asset tracking, healthcare, energy management, and wearable devices.

Unfortunately, the market is fragmented, with multiple non-interoperable technologies, some with coverage issues, some with non-user-friendly network configuration and deployment issues, and some with scalability issues. 802.11ah addresses these deficiencies with an optimized design and operating in the sub 1 GHz spectrum that is available worldwide for the IoT use case.

The 802.11ah amendment defines a narrow band Orthogonal Frequency Division Multiplexing (OFDM) physical layer (e.g. 1/2/4/8/16MHz) operating in the license-exempt bands below 1 GHz that enables an “extended range WLAN” with significantly lower propagation loss through free space and walls/obstructions, augmenting the heavily congested 2.4 GHz band and the shorter-range 5 GHz bands used today.

802.11ah has enabled multiple low rate modes (starting from 150 Kbps) and higher rate modes (up to 78 Mbps per spatial stream, and up to 346 Mbps for 4 spatial streams). Low rate modes, suitable for IoT applications provide whole home coverage for battery operated, small-form factor devices such as temperature and moisture sensors. Higher rate modes, suitable for ER applications, support plug-in devices with a power amplifier, such as video security cameras. Users can now deploy 11ah sensors in attics, backyards, basements, and garages, and have them directly connect to a single 11ah access point (AP). The 11ah MAC layer is optimized for long battery life, with features such as long sleep cycles, scheduled access, and scalability to up to 8191 nodes per AP for industrial applications, shorter packets and faster AP response times to sensor requests.

For ER applications, 802.11ah provides whole home coverage for voice and video-call applications. Users do not have to deal with dropped calls due to handoffs with multiple APs in a home or to a wireless WAN, and also benefit from improved voice quality. 802.11ah scheduled access and low rate modes also enable significantly lower power consumption. Finally, the coexistence between these varieties of devices targeting IoT and ER applications has also been considered as an integral part of the IEEE 802.11ah design.

ITU Focus Group IMT-2020 on 5G Non-Radio Aspects

Peter Ashwood-Smith, Huawei Technologies, Chairman of ITU Focus Group IMT-2020

Unquestionably, wireless networking has played, and will continue to play, an important role in society. Within the ITU, ITU-T Study Group 13, the study group responsible for future networks including cloud computing, mobile and NGN, has initiated work on the standardization of the network aspects of next generation wireless networks, commonly referred to in the industry as 5G, or within the ITU as International Mobile Telecommunications (IMT)-2020, reflecting the anticipated deployment of such systems. IMT-2020 networks are seen as critical for supporting many new applications and devices, including the Internet of Things (IoT). An important aspect of these networks is an expected drastic increase in the number of devices attached. In comparison with current 4G networks, IMT-2020 networks will provide users with hsafetysafetyigher bit rates, high reliability, and low latency to support new services in areas such as heath care, safety, or automation. Technologies such as SDN and NFV, and concepts such as network slicing and softwarization, will be important to realize these goals.

In order to begin developing standards for new mobile networks, Study Group 13 at its May, 2015 plenary meeting established a Focus Group open to non-ITU members to begin an in-depth study related to non-radio aspects of IMT-2020 networks. The primary goal of the Focus Group would be to gather experts from both industry and academia for the purposes of understanding and outlining what gaps may exist in current network standards in order to support IMT-2020 networks. The output of this work would allow ITU-T to define its work plan in conjunction with anticipated work in other SDOs and, where necessary, begin work on new standards.

FG IMT-2020 began its work in June, 2015, and was to run until October 2015, at which time the output would be submitted to the December meeting of Study Group 13. Due to this aggressive target, and considering that many architectural concepts were still being formulated, the Focus Group leadership established a working structure that could allow a significant portion of work to proceed in parallel, making extensive use of virtual meetings. Four face-to-face meetings were held to allow the individual areas to link together.

The five main areas identified for study were:

- High level network architecture.
- Network softwarization.
- End-to-end quality of service.
- Front haul/back haul.
- Emerging network technologies.

The final output was prepared and agreed to at the concluding meeting in October, 2015, which identified 85 "standardization gaps" related to non-radio aspects of IMT-2020 networks. Along with identifying these gaps, the group also discussed and made recommendations on possible future activities that could be undertaken based on the large body of potential new work identified in the first stage. While it was stressed that the objective of standardization activity would be to work collaboratively with other standards bodies in non-overlapping areas, the Focus Group did identify a number of areas where further work on new mobile networks could be beneficial to the industry. These areas included:

- Prototyping or demonstration activity, including the open source community.
- Network softarization and information centric networking (ICN).
- Refinements to the IMT-2020 architecture.
- Fixed/mobile convergence.
- Network slicing for front haul/back haul.
- New traffic models (and associated QoS, OAM).

At the time of this writing, Study Group 13 had not met to discuss the output of the focus group and decide on possible future directions. However, based on a recent ITU CTO meeting of key industry leaders, the overall view by CTOs suggested that the results of the Focus Group and possible future directions are highly relevant.

Update on the OPNFV Project and it's Role in the NVF Ecosystem

Christopher Price, Ericsson, Stockholm, Sweden

Open platform for NFV (OPNFV) is an open source software project focused on establishing an integrated platform to accelerate the introduction of new NFV products and services. The platform is composed solely of open source components and, through its pluggable architecture, offers choices at the component level to facilitate feature growth, variety of applications, and innovative use cases.

The OPNFV community, formed in the autumn of 2014, has been focused on the integration and development of the virtual infrastructure (VNFI) and management (VIM) components of the NFV reference architecture as defined by the ETSI NFV ISG. The infrastructure layer provides a common and consistent platform for the deployment of software running in virtual environments. Common examples of these components include, for instance, OpenStack, OpenDaylight, OVS, and Linux.

Arno, the inaugural release of OPNFV named after the river in Tuscany, was released in June 2015, approximately eight months after formation (https://www.opnfv.org/arno). The first release, intended as a "developmental" release, provided the foundation for the integration and deployment pipeline for the project, the physical infrastructure definitions, and the testing and validation frameworks used by the community.

The Arno release was intentionally targeted to deploy to a physical infrastructure providing redundancy and availability in all aspects of the platform. Approximately three months after the initial release of Arno, the community provided an Arno stable release, delivering improved stability and providing support for nested deployments of the platform aimed at empowering the development community who lack access to the physical infrastructure.

While the foundational elements of the project are being developed and improved, the broader community is focused on solving problems and developing capabilities for the platform. OPNFV now boasts over 40 projects in total addressing functional areas from IPv6 support through platform availability and FCAPs, to advanced service chaining and platform policy functions.

Brahmaputra, named after the river in Asia, is the second planned major release of the OPNFV project. The Brahmaputra release will focus on further stability and increased platform functionality and flexibility. Highlights of the Brahmaputra platform include: the option of selecting from four networking controllers, increased performance and fault management, data center VPN support, and significantly improved deployment and testing solutions. The Brahmaputra release is planned to be available in February 2016.

As OPNFV focuses on deployment of data center technology to physical infrastructure, the management and operations of hardware is an area of significant focus for the community. At the forefront of this area is the Pharos project (https://www.opnfv.org/developers/pharos), which has the responsibility for defining standard deployment architectures, physical connectivity, configuration, and orchestrating the global federation of labs.

There are currently more than a dozen globally distributed labs connected to the OPNFV integration and deployment pipeline. With this number increasing regularly, the management and federation of these infrastructures is a critical activity for the project. The variety of labs is a critical component for the project in developing the OPNFV platform to be fully interoperable at the physical layer, the VI-Ha interface as defined by the ETSI NFV ISG.

The OPNFV project was forged from the success of the ETSI NFV ISG in defining the reference architecture for NFV, and continues to work closely with the ISG as the standards further emerge. The OPNFV community has been proactive in working with other SDOs to ensure interoperability and industry alignment. Examples of such collaboration include the development of service chaining technologies adopting IETF SFC specification, orchestration and lifecycle management of layer 2 services in collaboration with the MEF. It is intended that further development of the platform will result in broader collaboration across the industry.

OPNFV operates under some key tenets: "upstream first" indicates a strong tendency to not fork upstream code, but rather work with source communities, and "diversity of community and interests" keeps the platform responsive to a range of industry needs. With these basic principals and a clear vision, OPNFV is fostering a varied community of developers who bring different requirements, ideas, and knowledge to the table, resulting in faster time to market and stronger code.

IEEE 802.1 Time-Sensitive Networking

Michael Johas Teener, IEEE 802.1 Time-Sensitive Networking Task Group

The charter of the Time-Sensitive Networking (TSN) TG is to provide the specifications that will allow time-synchronized low latency streaming services through 802 networks. The TSN TG was originally created with a different name in 2005 (the "Audio Video Bridging TG", or "AVB") since the intentions were to provide the tight synchronization and low deterministic delays needed for use in studios and live audio and video performances. These capabilites were noticed by people in other industries, so the name was changed to "Time-Sensitive Networking" to make it more descriptive.

The group has been working toward three goals over the years:

1) A precise time synchronization service that works across any standard 802 LAN, in particular IEEE 802.3 Ethernet (including EPON) and IEEE 802.11. Since there was already a successful IEEE time synchronization specification (IEEE 1588 Precision Time Protocol), the TSN TG created an 802-specific profile of 1588 with an extension architecture to support non-Ether- net LANs and performance and management enhancements. The resulting specification, IEEE 802.1AS "Generalized Precision Time Protocol", was approved in 2011, and is now in the process of being revised to further improve performance and enable a future unification of 802.1AS and 1588 (which is also being revised).

2) A unified method to manage the resources needed to guarantee the necessary resources to get a stream through a heterogenious network with low and deterministic delays. The original version of this work was published as IEEE 802.1Qat-2010 "Stream Reservation Protocol" and was based on the earlier IEEE 802.1 "Multiple Attribute Registration Protocol". This is in the process of being revised as project IEEE 802.1Qcc, which will, among other things, specify a data model and managed objects that will enable the kind of centralized management needed for the most difficult of time-sensitive configurations.

3) A set of QoS procedures that enhance the standard 801.1Q bridge frame forwarding rules to meet the requirements of the various TSN interest groups, including the original A/V market, but also useful for industrial and automotive sensing and control networks. The original work was the specification of a simple credit-based shaper (IEEE 802.1Qav-2010) which works in many common A/V environments, but was not adequate for the more critical control systems. To support control systems (and even larger and more complex A/V systems), two new specifications for scheduled queues (IEEE 802.1Qbv) and frame preemption (IEEE 802.1Qbu) have been written, and as of this writing, have passed sponsor ballot and are awaiting processing by REVCOM. Frame preemption requires complementary support at the MAC level, and the Ethenet complement, IEEE 802.3br, is about to enter sponsor ballot. There is a pair of additional QoS procedures that have only recently started: IEEE 802.1Qch Cyclic Queuing and Forwarding, which is effectively a compatible replacement for 802.1Qav with much simpler rules for delay calculations; and something called the "Urgency Based Shaper", which provides the lowest delays of all, at the cost of much more complex queuing mechanisms.

Since automotive and industrial control networks have very high reliablity requirements, a number of semi-proprietary versions of Ethernet were developed for those markets. The TSN TG has tried to extend the 802.1 model to meet those requirements without breaking standard Ethernet and 802.1Q switches. Two of those efforts include:

1) IEEE 802.1CB "Seamless Redundancy", which specifies the ways that frames in a stream can be duplicated and sent on different paths through a network, and then the duplicates removed at a point whre the different paths merge

2. IEEE 802.1Qci "Per-Stream Filtering and Policing", which provides the complementary protection function for the path reservation procedures. It defines the procedures and parameters necessary for a bridge to protect resources (bandwidth, queues, etc) from being overwhelmed by non-compliant streams.

Finally, the TG produces "system specifications" or "profiles" that specifiy a set of usage-specific profiles that will help ensure interoperability between networked devices using the TSN specifications. The first example of this work was published as IEEE 802.1BA-2011 "Audio Video Bridging (AVB) Systems". An additional specification, IEEE 802.1CM "Time-Sensitive Networks for Fronthaul", has just been started to focus on the requirements for networks carrying digitized radio data from antennas to remote processing nodes.

New Revenue-Generating Services: The Broadband 20/20 Vision

Robin Mersh, CEO at Broadband Forum

Software Defined Networking (SDN) and Network Functions Virtualization (NFV) have already made a remarkable impact on mobile networks, with operators implementing the technologies to reduce OpEx and increase network agility. Despite the growing influence of these initiatives, their use has not been deployed in the broadband network nor has it been clarifiedhow these new technologies will enable new revenue-generating services — until now.

The Broadband 20/20 Vision: Broadband 20/20 leverages the innovative use of NFV, SDN, ultra-fast access, Internet of Things (IoT) and, when formally defined, 5G, in the home, small business, and multiuser infrastructure of the broadband network. As a result, providers will be able to enable new opportunities for profitable revenue growth. This includes the delivery of ultra-fast broadband services, as well as distributed compute and storage to anywhere and any device in the home and business context. New services and applications will also be enabled, transforming the way people communicate, purchase, and consume all manner of informational and entertainment content.

Making History: While all service providers are, of course, very focused on profitable revenue-generating services, this is the first time any forum has come together to take a holistic approach to these new technologies and have them deliver real value. NFV and SDN has made the industry think about the speed of service provisioning in this era of intense competition, and it has made us look across the entire ecosystem to converge and develop technologies that will create revenue and services quickly and efficiently.

We have identified and are enabling five areas of innovation: ultra-fast infrastructure services; intelligent home/small business services; hybrid wireline/wireless network services; performance-assured IP broadband services; and personalized network services.

Major Restructuring: In order to realize the promise of the vision, the Forum has undergone a major restructuring into eight work areas, and introduced a new culture of rapid technical development so as to incrementally deliver technical specifications, architecture, management, software data models, APIs, interoperability tests, etc., around which the industry can evolve.

For the foreseeable future, this will see the broadband infrastructure act as a hybrid ecosystem, handling current static and new programmable, virtualized networking and computing as the market naturally evolves. We have also appointed Software Architect William Lupton and have already started working to deliver on many key work areas in the Broadband 20/20 vision.

Ever-Changing Industry: Of course, much remains to be achieved, particularly in the area of data models and APIs. The pace of change in telecommunications never slackens — it always increases — and this new focus for the Forum will support the industry, not only in matching that pace, but at the same time also enabling the creation of important new revenue streams.

Activities of the IEEE 802.24 Vertical Applications Technical Advisory Group

Tim Godfrey, 802.24 TAG Chair, 802.24.1 Chair

The IEEE 802.24 Vertical Applications Technical Advisory Group (TAG) focuses on application categories that use IEEE 802 technology and are of interest to multiple IEEE 802 WGs. The TAG was formed in 2012 with a single scope of Smart Grid. In 2014 a process was defined to enable new application areas, and Internet of Things (IoT) was approved as an additional scope area. The TAG acts as a liaison and point of contact with regulatory agencies, industry organizations, other SDOs, government agencies, IEEE societies, etc., for questions regarding the use of 802 standards in those applications.

The 802.24.1 Smart Grid Task Group addresses the use of IEEE 802 standards in supporting the data link communications needs of Smart Grid systems. In addition to the TAG activities listed above, the Smart Grid TG is active in the following areas:

- Developing white papers, presentations, and other documents that do not require a PAR that describes the application of IEEE 802 standards to Smart Grid applications.
- Acting as a point of contact for industry organizations and alliances for questions regarding the use of IEEE 802 standards in Smart Grid applications.
- Acting as a resource and knowledge base on IEEE 802 standards for certification efforts by industry bodies that require more than one IEEE 802 WG's input.
- Identifying technology gaps in Smart Grid communications and recommending standardization tasks to the appropriate IEEE 802 WG.

The 802.24.1 Smart Grid Task Group (TG) has developed an overall Smart Grid white paper and companion presentation, and is developing a more focused white paper highlighting the smart grid applications of the multiple IEEE 802 wireless standards that operate in bands below 1 GHz.

The 802.24.2 TG was approved in March 2015. It provides a forum to discuss substantive basis for applying 802 standards to the evolving IoT in specific vertical application areas and to gather together like minded experts interested in exploring the topics. The scope of the IoT task group is to:

- Develop white papers, presentations, and other documents that do not require a PAR that describes the application of IEEE 802 standards to identified IoT vertical applications.
- Identify technology gaps in selected IoT vertical applications, and recommend standardization tasks to the appropriate IEEE 802 Working Group.
- Provide a forum for liaison activities with industry organizations, other SDOs, government agencies, IEEE societies, etc., in identified IoT vertical applications.
- Maintain consistency with the IoT architecture framework in development in IEEE P2413.

White paper development in IoT vertical applications areas are intended to provide customers with insight into current industry technical developments related to IoT. The targeted white paper audience includes: end users, service providers, OEMs, chip suppliers, standards developers (e.g. IEEE P802, TIA, ISO, IEC) and industry alliances

Liaison activities are ongoing via a liaison officer with internal SDOs (e.g. IEEE P802 working groups, IEEE P2413), external SDOs (e.g. TIA, ISO, IEC), industry alliances, government agencies, IEEE societies, etc., in identified IoT vertical applications.

Identified IoT vertical applications have broad applicability across numerous users and multiple vendors. These include:

- Industrial Automation.
- Data Center Management.
- Security/Video Networks – Wireless/Wired.
- Building Automation.
- Automotive.

White paper development initiated in Data Center Management, Security, and an overview of (IoT) provides an overview of Internet of Things (IoT) activities for consideration in IEEE 802.

IoT/M2M from Research to Standards: The Next Steps (Part II)

Omar Elloumi *JaeSeung Song* *Yacine Ghamri-Doudane* *Victor C.M. Leung*

As the pace of IoT deployments accelerate, IoT standards are undergoing major evolutions, sometimes revolutions. For instance, cellular network standards are now adding techniques to improve network performance to address traffic patterns generated by an increasing number of IoT devices. Ongoing discussions around 5G requirements may become game changing for M2M communications because the standard will be designed, from the ground-up, for massive-scale IoT deployments. This is a radical shift compared to the "quick-fixes" 3GPP and 3GPP2 have been adding to 2G/3G and 4G standards so far. Another example of this radical shift is related to IoT service platforms (such as the platform standardized by oneM2M) and IoT applications. Semantic interoperability is now emerging as a major trend that allows data exchange between applications, an increased level of interoperability, analytics and reasoning. With ontologies engineering, researchers will soon overcome the limitations of static data models and bridge the gap between the currently deployed vertical silos. Other areas that will see intense standardization activity are IoT security and low power wide area connectivity.

The articles selected for this feature topic can broadly be grouped into four categories: networking including 5G, semantic interoperability, security, and low power wide area connectivity. This Feature Topic, which attracted no less than 40 submissions, out of which we have selected 14 articles, is structured in two parts. Part I, which consists of nine articles and was published in September 2015, addresses network and connectivity topics. Part II of the Feature Topic consists of five articles, and covers security and semantic interoperability topics.

The second part of this Feature Topic starts with an article by Chengzhe *et al.*, "Towards Secure Large-Scale Machine-to-Machine Communications in 3GPP Networks: Challenges and Solutions," which investigates group-based security for large-scale M2M communications in 3GPP network. The article provides an overview of the 3GPP security architecture for Machine Type Communications and identifies several major security challenges, in particular performance issues pertaining to authentication of 3GPP MTC devices. The authors introduce several 3GPP candidate mechanisms of group message protection followed by research directions.

The second article by Ramão *et al.*, "The Importance of a Standard Security Architecture for SOA-Based IoT Middleware," highlights the importance of defining a standard security architecture for Service Oriented Architecture (SOA) based IoT middleware. The authors of this article provide a thorough discussion of the security services and protocols that can be applied to a SOA-based IoT middleware, and propose the definition of standard security architecture together with the importance of lightweight security service from the standard security architecture.

The third article by Kyle *et al.*, "SCALE: Safe Community Awareness and Alerting Leveraging the Internet of Things," describes a safe home architecture and prototype using novel networking technologies, commodity sensor devices, cloud services and middleware abstractions. In particular, the authors of this article introduce a publish/subscribe system based on MQTT for emergency management and safety applications. The article then finishes with lessons learned that will help drive future research and standards for IoT.

In the fourth article by Alaya *et al.*, "Toward Semantic Interoperability in oneM2M Architecture," the authors propose a semantic extension of the oneM2M standard to support semantic interoperability based on an expressive IoT ontology. In the article this IoT ontology merges together a set of popular ontologies and is enriched with new relevant concepts and relationships.

Finally in the fifth article, "Toward Enhanced Data Exchange Capabilities for the oneM2M Service Platform" by Glaab *et al.*, the authors introduce a semantic interoperability framework for the automotive industry based on oneM2M service layer standards. After analyzing the data exchange capabilities of the oneM2M standard, the article proposes several enhancements to oneM2M standards such as a local aggregation of different subscriptions to the same resource.

We hope the readers will enjoy this issue and find the articles useful. We would like to thank all the authors for submitting their proposals and the reviewers for their tremendous help in reviewing the articles. Special thanks to Chonggang Wang for his help in structuring this Feature Topic. We also want to express our thanks for the great leadership and help of Sean Moore (previous Editor-in-Chief), Osman S. Gebizlioglu (current Editor-in-Chief), and Glenn Parsons (Editor-in-Chief of Standards Supplement) who have guided us in this endeav-

or, and Joseph Milizzo from the Communications Society staff for his support.

Biographies

Omar Elloumi [M] (omar.elloumi@alcatel-lucent.com) is head of M2M and smart grid standards within Alcatel-Lucent CTO. He is the chair of the oneM2M Technical Plenary. He joined Alcatel-Lucent in 1999 and has held several positions in areas including research, strategy, system architecture, and more recently standards. He holds a Ph.D. degree in computer science, and he has served on several consortia Board of Directors and chaired several standards committees.

JaeSeung Song [M] (jssong@sejong.ac.kr) is an assistant professor in the Computer and Information Security Department at Sejong University. He holds the position of oneM2M Test Working Group Chair. Prior to his current position, he worked for NEC Europe Ltd. and LG Electronics in various positions. He received a Ph.D. at Imperial College London in the Department of Computing, United Kingdom. He holds B.S. and M.S. degrees in computer science from Sogang University.

Yacine Ghamri-Doudane [M] (yacine.ghamri@univ-lr.fr) is currently a full professor at the University of La Rochelle (ULR), France. Previously, he received an engineering degree in computer science from the National Institute of Computer Science (INI), Algiers, Algeria, in 1998, an M.S. degree from the National Institute of Applied Sciences (INSA), Lyon, in 1999, and a Ph.D. degree in computer science from the University Pierre & Marie Curie in 2003. He has been an officer for two IEEE ComSoc committes.

Victor C. M. Leung [F] (vleung@ece.ubc.ca) is a professor of electrical and computer engineering and holder of the TELUS Mobility Research Chair at the University of British Columbia. He has co-authored more than 800 technical papers in the area of wireless networks and mobile systems. He is a Fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada.

Toward Secure Large-Scale Machine-to-Machine Communications in 3GPP Networks: Challenges and Solutions

Supporting a massive number of machine-to-machine (M2M) communications devices has been considered an essential requirement for mobile operators. Meanwhile, cyber security is of paramount importance in M2M as all applications involving M2M cannot be widely accepted without security guarantee.

Chengzhe Lai, Rongxing Lu, Dong Zheng, Hui Li, and Xuemin (Sherman) Shen

Chengzhe Lai and Dong Zheng are with Xi'an University of Posts and Telecommunications, Xidian University, and Chinese Academy of Sciences.

Rongxing Lu is with Nanyang Technological University.

Hui Li is with Xidian University.

Xuemin (Sherman) Shen is with University of Waterloo.

Abstract

With trillions of machines connecting to mobile communication networks to provide a wide variety of applications, supporting a massive number of machine-to-machine (M2M) communications devices has been considered an essential requirement for mobile operators. Meanwhile, cyber security is of paramount importance in M2M as all applications involving M2M cannot be widely accepted without security guarantees. In this article we focus on the standardization activities of 3GPP, especially group-based security for large-scale M2M communications in 3GPP networks. We first introduce the main components of the machine-type communication (MTC) security architecture. Then we discuss several major challenges for group-oriented secure M2M communications in 3GPP systems, i.e. authentication signalling congestion and overload, and group message protection. Specifically, we identify the performance issues of authentication signalling congestion and overload in no/low mobility scenarios, and propose three group access authentication and key agreement protocols. Moreover, several 3GPP candidate solutions for group message protection are introduced. Finally, we present key issues and research directions related to group-based secure M2M communications, including security, privacy, and efficiency in mobility scenarios of MTC, and flexible and efficient group key management.

Introduction

Machine-to-machine (M2M) communications is an emerging technology empowering full mechanical automation (e.g. in the smart grid, smart transportation, smart city, etc.), and its rapid development is changing our living styles vigorously [1]. M2M technology is drawing overwhelming attention in the standardization and industry areas. Many standards forums and organizations, including the Institute of Electrical and Electronics Engineers (IEEE), the European Telecommunications Standards Institute (ETSI), the Third Generation Partnership Project (3GPP), the China Communications Standards Association (CCSA), oneM2M, etc., have engaged in M2M standard development.

To take full advantage of the opportunities created by a global M2M market over cellular networks, 3GPP[1] has initiated their working groups to facilitate such applications through various releases of their standards [2]. So far, much research effort has focused on the MTC, such as subscription control and network congestion/overload control [3], potential issues on the air interface, including physical layer transmissions, the random access procedure, and radio resource allocation supporting the most critical QoS provisioning [4], mobility management [5], green, reliability, and security of M2M communications [6], etc. As a cutting edge technology for next generation communications, M2M communications is undergoing rapid development and inspiring numerous applications. However, all applications involving M2M cannot be widely accepted without security guarantees. In addition, to support large-scale M2M communications, the 3GPP mobile operator must accommodate its network to support a large number of MTC devices. Therefore, achieving secure large-scale machine-to-machine networking will be a challenge issue in the near future.

In this article we cover some of the standardization activities of 3GPP, focusing especially on the problem of group-based security for large-scale M2M communications in 3GPP networks. First, to address the problems of authentication signaling congestion and overload, we define three types of performance issues in no/low mobility scenarios. Then we propose three group access authentication and key agreement protocols for M2M in 3GPP networks to address them. Second, to solve the key issues in the group based feature (i.e. group based messaging, group based charging optimizations, group based policy control, and group based addressing and identifiers, etc.), several candidate solutions of group message protection are given from the 3GPP point of view.

The remainder of this article is organized as follows. We present the main components of the MTC security architecture. We then discuss several major challenges for group-oriented secure M2M communications in 3GPP systems, i.e. authentication signalling congestion and overload, and group message protection. Furthermore, we introduce new solutions to congestion and overload control for authentication signalling, and provide a summary of the solutions, agreed within 3GPP SA2, for group message protection. Finally, we present potential research directions and conclude the article.

Security Architecture

Figure 1 [7] shows the security architecture for MTC connecting to the 3GPP evolved universal terrestrial radio access network (E-UTRAN) via the LTE-Uu interface. The security architec-

[1] In 3GPP standards, M2M communications is also named machine-type communication (MTC).

0163-6804/15/$25.00 © 2015 IEEE

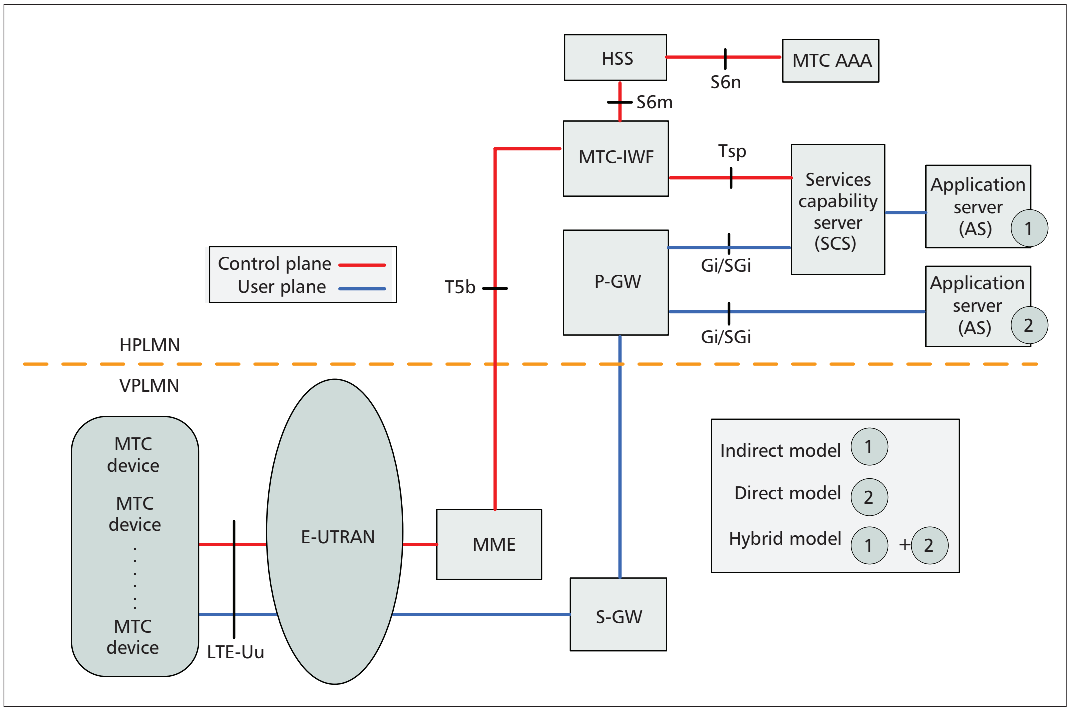


Figure 1. 3GPP security architecture for MTC.

To support large-scale M2M communications, the 3GPP mobile operator has to accommodate its network to support a large number of MTC devices. Therefore, achieving secure large-scale machine-to-machine networking is a challenge issue in the near future.

ture is considered in the roaming scenario, which includes the roaming network domain, i.e. the visited public land mobile network (VPLMN), and the home network domain, i.e. the home public land mobile network (HPLMN).

Table 1 summarizes the functions and descriptions of security related components and reference points of the MTC in this article. The main security related components and reference points of the MTC are introduced as follows.

Network Elements

Home Subscriber Server (HSS): Besides original functions (e.g. authentication and authorization), HSS supporting device triggering mainly supports the following functionalities:

- Stores and provides to MTC-IWF (and optionally to MTC-AAA) the mapping/lookup of the E.164 MSISDN (i.e. the mobile subscriber international ISDN/PSTN number) or external identifier(s) to IMSI (i.e. international mobile subscriber identity) and subscription information used by MTC-IWF for device triggering.
- Mapping of the E.164 MSISDN or external identifiers to IMSI.
- HSS stored "routing information" including serving node information if available for the MTC device (e.g. serving the MME identifier).
- Determines if an SCS is allowed to send a device trigger to a particular MTC device.
- Provides to MTC-AAA the mapping between IMSI and external identifier(s).

MTC Accounting, Authorization, and Authentication (MTC-AAA): To support translation of the IMSI to external identifier(s) at the network egress, an AAA function (MTC-AAA) is used in the HPLMN. The MTC-AAA may be deployed to return the external identifier(s) based on IMSI. Alternatively the MTC-AAA may be deployed as a RADIUS/diameter proxy between the packet data network gateway (P-GW) and the AAA server in the external packet data network (PDN).

MTC Interworking Function (MTC-IWF): The MTC-IWF is the functional entity that hides the internal network topology and relays/translates signaling protocols used over Tsp to invoke specific functionality in the public land mobile network (PLMN) (e.g. control plane device triggering). An MTC-IWF could be a stand-alone network element or a functional entity of another network element and always reside in the HPLMN.

Services Capability Server (SCS): The SCS connects to the 3GPP network via the MTC-IWF in the HPLMN to communicate with the MTC device. The SCS offers capabilities for use by one or multiple MTC applications. An MTC device can host one or multiple MTC applications. The corresponding MTC applications in the external network are hosted on one or multiple application servers (ASs).

Reference Points

The SCS provides an application programming interface (API) to allow different ASs to use the capabilities of the SCS.[2] Tsp is a 3GPP standardized interface to facilitate value-added services motivated by MTC (e.g. control plane device triggering) and provided by an SCS. The T5b interface is intended to provide optimized paths for device trigger delivery and possibly other services (e.g. small data service) to the MTC device. T5b was not standardized in 3GPP Rel-11. The S6m interface is used by the MTC-IWF to interrogate the HSS for mapping an MSISDN or external identifier to the IMSI, retrieving serving node information, and authorizing a device trigger to a particular MTC device. The S6n is an inter-

[2] The interface between SCS and AS is not standardized by 3GPP, but other standards development organizations (SDOs) such as the European Telecommunications Standards Institute (ETSI) Technical Committee on Machine-to-Machine Communications (TC M2M) are expected to standardize APIs.

Network elements	Function
HSS	Main database containing subscription-related information, which is used for authentication, authorization, and supporting device triggering.
MME	The mobility management entity for all mobility related functions and performing the authentication on behalf of the 3GPP core network.
MTC-IWF	The functional entity that hides the internal network topology and relays/ translates security related signaling protocols, e.g. generates group key, encrypts, and signs the group message.
SCS	The entity connects to the 3GPP network via the MTC-IWF in the HPLMN to communicate with the MTC device, e.g. makes group message request.
MTC-AAA	The entity that supports translation of the IMSI to external identifier(s) at the network egress, or plays a "RADIUS/diameter proxy" role between the P-GW and the AAA server.
Reference points	**Description**
Tsp	Reference point used by an SCS to communicate with the MTC-IWF related control plane signaling.
T5b	Reference point used between the MTC-IWF and the serving MME.
S6m	Reference point used by the MTC-IWF to interrogate the HSS.
S6n	Reference point used by the MTC-AAA to interrogate the HSS.

Table 1. Summarizing table of security related network elements and reference points.

face between MTC-AAA and HSS to interrogate HSS for mapping IMSI to external identifier(s) and vice versa at the network egress.

Group-Oriented Secure M2M Communications

There exist several major challenges for group-oriented secure M2M communications in 3GPP systems, including: How to control authentication signalling congestion and overload when a large number of MTC devices want to securely access the 3GPP core network? How to securely and effectively protect group message distribution for the one-to-many or many-to-many communication paradigms?

Challenge 1: Congestion and Overload Control for Authentication Signalling

An MTC group is formed when a group of MTC devices are in the same area and/or have the same MTC features attributed and/or belong to the same MTC user. The MTC group should be identified uniquely across 3GPP networks. When a group of MTC devices want to access the network, they may send their access authentication requests toward the core network successively over a short period of time, or even at the same time, leading to congestion in the different nodes of the network, across the communication path. According to 3GPP TS 22.368 [8], the congestion could happen at different locations, as shown in Fig. 2.

Radio Network Congestion: Radio network congestion because of mass concurrent access authentication requests takes place in some MTC applications. One of the typical applications is tbridge monitoring with a mass of sensors. When a train passes through the bridge, all the sensors may access the network and transmit monitoring data almost simultaneously. The same thing happens in hydrology monitoring during times of heavy rain and in building monitoring when intruders break in. The network should be optimized to enable a mass of MTC devices in a particular area to access the network and transmit data almost simultaneously.

Core Network Congestion: Authentication signalling congestion in the core network is caused by a high number of MTC devices trying almost simultaneously:

- To attach to the network.
- To activate/modify/deactivate a connection.

In a 3GPP system supporting MTC applications, such an overload of the network can be caused by, for example, many mobile payment terminals that become active on a national holiday or by high numbers of metering devices becoming active almost simultaneously after a period of power outage. Also, some MTC applications generate recurring data transmissions at precisely synchronous time intervals (e.g. precisely every hour or half hour). Preferably, the 3GPP system provides the ability to the network operator and MTC user to spread the resulting peaks in the signalling traffic.

To support M2M communications, the 3GPP mobile operator must accommodate their network to support a large number of MTC devices, which can overload network resources and introduce congestion in the network at both the data and control planes. In fact, congestion may occur due to simultaneous authentication signalling messages from MTC devices. Unfortunately, the recent authentication and key agreement (AKA) protocols dedicated to the 3GPP evolved packet system (EPS), known as EPS-AKA [9], or for non-3GPP access networks (e.g. WLAN or WiMAX), known as EAP-AKA [10], cannot provide a group authentication mechanism. If a large number of MTC devices in a group need to access the network almost simultaneously, the traditional authentication protocols (e.g. EPS-AKA or EAP-AKA) will suffer from high signalling overhead, leading to authentication signaling congestion and decreasing the quality of service (QoS) of the network, because every device must perform a full AKA authentication procedure with the HSS, respectively. Because the traditional AKA protocols are not suitable for large-scale M2M communications, we consider designing new group-based access authentication and key agreement protocols.

Our Proposed Solution: Group-Based Access Authentication and Key Agreement

To facilitate system optimization, 3GPP defines a low mobility feature in M2M communications, which is suitable for MTC devices that do not move, move infrequently, or move only within a certain area. This feature enables the network operator to be able to simplify and reduce the frequency of mobility management procedures.

In such no/low mobility scenarios, the following three types of performance issues (PIs) are shown:

PI1: In some applications, a group of MTC devices may want to access the network and send their access authentication requests toward the core network successively over a short period of time. In such a case, if every device still performs a full authentication and key agreement (AKA) procedure with the HSS, the authentication signaling in the network increases. Meanwhile, the overload of HSS will increase because of frequently acquiring authentication vectors (AVs). Moreover, when these devices roam in a visiting domain, which is far from their home domain, the communication may suffer from high network access latency until the completion of authentication procedures by all MTC devices in the same group.

PI2: In some applications, the capabilities of each MTC device, such as computation, battery, and storage, are enough to support a public key cryptosystem. When a group of MTC devices want to access the network and send their access authentication requests toward the core network simultaneously, if every device still performs a full authentication and key agreement (AKA) procedure with the HSS, besides PI1, the authentication signalling congestion occurs at the HSS, MME, and evolved node B (eNB).

PI3: In some applications, a group of MTC devices may want to access the network and send their access authentication requests toward the core network simultaneously. Besides PI2, the capabilities of each MTC device, such as computation, battery and storage, are not enough to support a public key system and thus the symmetric key cryptosystem needs to be applied.

Accordingly, we present three group access authentication and key agreement protocols: GAAKA-1, GAAKA-2, and GAAKA-3.

GAAKA-1: First, the MTC devices form groups based on certain principles (e.g. they belong to the same application, are located within the same region, etc.), then the supplier provides a group identity (ID_{Gi}) and a group key (GK_i) to each group for authentication [11]. When a group of MTC devices want to access the network successively over a short period of time, the first device performs a full AKA procedure and obtains a group temporary key (GTK) for all of the group members. Then the remaining devices in the group only need to perform a simplified AKA procedure with the MME locally without interacting with the HSS. Therefore, the authentication signaling between the MME and the HSS can decrease. Meanwhile, the overload of HSS will decrease as well. Especially when these devices roam in a visiting domain, the performance can be optimized significantly.

GAAKA-2: Similarly, the MTC devices form groups based on certain principles (e.g. they belong to the same application, are located within the same region, etc.), and then the identities of MTC groups (ID_{Gi}) are assigned to each group. Meanwhile, a group leader of MTC devices in the group ($MTCD_{leader}$) will be selected in advance. When each MTC device registers with the EPC, it contacts the key generate center (KGC), provides an identifier, and then receives its private key. Only the authenticated MTC devices can obtain the private keys from the KGC. The KGC can be integrated with the HSS, which has pre-established secure channels with the MME by using the NDS/IP security mechanism. By adopting the certificateless aggregate signature techniques [12], the $MTCD_{leader}$ can collect all signatures of members in the same group and aggregate them to a new signature SIG_{agg}. Then the $MTCD_{leader}$ sends SIG_{agg} to the network, and all members in the group can be authenticated at the same time. Moreover, the independent session key can be negotiated between the core network and each MTC device. Therefore, GAAKA-2 can significantly relieve authentication signalling congestion occurring at the HSS, MME, and eNB.

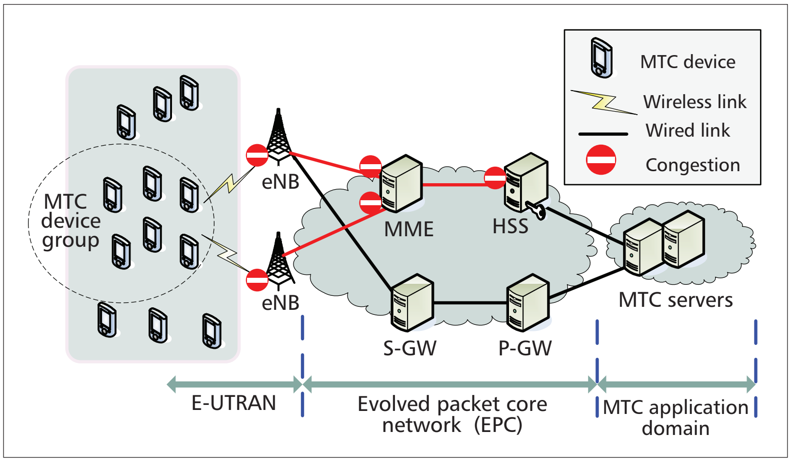


Figure 2. Authentication signalling congestion

GAAKA-3: Constrained by the computation, battery, and storage capabilities of the MTC device, GAAKA-2 may not be suitable for resource-constrained devices due to applying the public key system (e.g. certificateless aggregate signature techniques). Therefore, GAAKA-3 can be proposed by adopting the aggregate message authentication code (AMAC) techniques [13]. Similar to GAAKA-1, the supplier provides a group identity (ID_{Gi}) and a group key (GK_i) to each group for authentication. Each MTC device has a pre-shared secret key (K_{Gi-j}) with HSS when it is first registered in HSS. Meanwhile, a group leader of MTC devices in the group ($MTCD_{leader}$) will be selected in advance. Then the $MTCD_{leader}$ can collect all message authentication codes ($MAC_{indiv}s$) of members in the same group and aggregate them to a new message authentication parameter MAC_{agg}. Then, the $MTCD_{leader}$ sends MAC_{agg} to the network and all members in the group can be authenticated at the same time. Moreover, the independent session key can be negotiated between the core network and each MTC device. Therefore, GAAKA-3 cannot only relieve authentication signalling congestion occurring at the HSS, MME, and eNB, but also is suitable for resource-constrained devices. However, different from GAAKA-2, GAAKA-3 requires two additional authentication signalling exchanges between the MME and HSS.

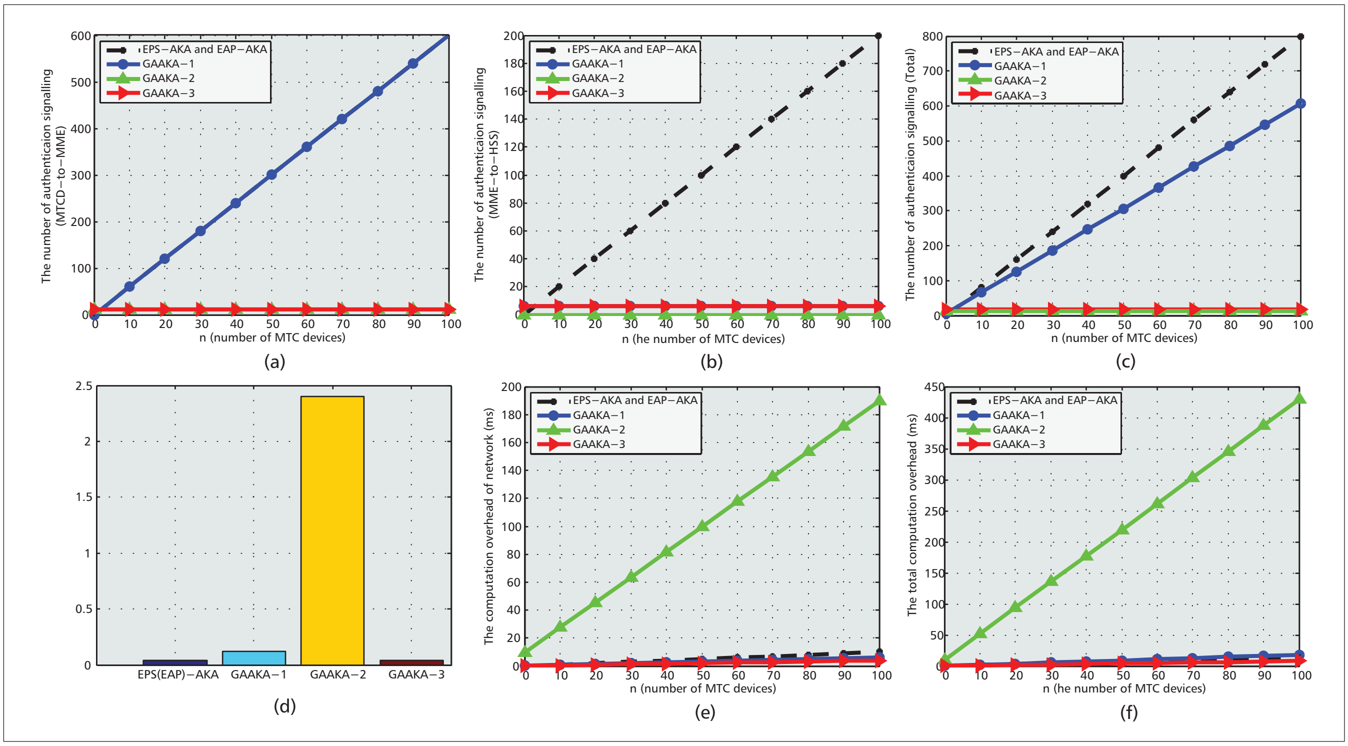


Figure 3. Performance comparison, (a–c): comparison of the authentication signalling; (d–f): comparison of the computation overhead (ms): a) authentication signalling between MTC device and MME (MTCD-to-MME); b) authentication signalling between MME and HSS (MME-to-HSS); c) total authentication signalling; d) computation overhead of each MTC device; e) computation overhead of network; and f) total computation overhead.

Analysis and Evaluation

We assume that there are *n* MTC devices forming *m* groups, obviously, $n > m$. We fix *m*, and plot Fig. 3. According to Fig. 3, we can see that the signalling of GAAKA-2 and GAAKA-3 do not change with *n*, and only depend on *m*; therefore, the authentication signalling incurred by GAAKA-1, 2, 3 are much less than that of EPS-AKA/EAP-AKA. Due to the use of symmetric cryptography (the hash operation T_{hash} takes 0.02 milliseconds (*ms*)), computation overheads of EPS-AKA/EAP-AKA, GAAKA-1 and GAAKA-3 are fairly small. Thus, we mainly consider the cost of the following operations, including a point multiplication T_{mul}, a pairing operation T_{pair}, and a map to point hash operation T_{mtp}. Generally, T_{mtp} takes the same time as T_{mul} (= 0.6*ms*) and T_{pair} = 4.5*ms*. The cost of XOR can be negligible. Therefore, we can see that GAAKA-1 and GAAKA-3 are more efficient than GAAKA-2 in computation, and are close to EPS-AKA/EAP-AKA.

Finally, a comprehensive comparison of design goals among several authentication and key agreement protocols is given in Table 2. We can find that except for GAAKA-2, other protocols are all designed based on the symmetric cryptosystem. Therefore, the computation overhead of GAAKA-2 is larger than that of other protocols. In addition, compared to the existing standard protocols, our proposed protocols have enhanced security properties, including privacy preservation, resistance to redirection attack, and resistance to MITM attack. Most importantly, different from the existing standard protocols, our proposed protocols support group access authentication, and can efficiently control authentication signalling congestion and overload.

Challenge 2: Group Message Protection for Secure M2M Communications

Recently, 3GPP SA2 has been working on the group based feature that includes the following key issues [14]: group based messaging, group based charging optimizations, group based policy control, group based addressing and identifiers, etc. To provide secure M2M group communications, 3GPP SA2 is currently considering the mechanism to distribute a group message from an SCS to those members of an MTC group located in a particular geographic area. According to the current architecture and solutions, MTC-IWF receives a group message from the SCS and forwards it to the target group of MTC devices. As group based messaging can significantly reduce the overhead of network resources, the corresponding session key establishment mechanism should be required to protect the group messages, which can be divided into two cases:

- For the MTC devices in one group, each device may need to communicate with the core network individually so an independent session key for each device may be needed.
- For the MTC devices in one group, the core network may need to distribute the same message (e.g. a trigger request) to those members of one MTC group as a same group session key is needed.

The first case has been discussed above, and we focus on the group message protection issue in this section.

If the broadcast message for a particular MTC group is not protected, then private information related to the particular group is revealed. Therefore, a mechanism should be provided to protect the confidentiality of the group message broadcasted for a particular group. However, confidentiality protection is subject to regional regulatory requirements. Group based messaging would be more prone to tampering and fake triggering attacks, if there is no integrity and replay protection provided by the core network or by the SCS. With a group message, multiple MTC devices can be triggered. Therefore, an unauthorized group message may cause a much more severe problem compared to what a trigger to a single MTC device can cause. Therefore, 3GPP has defined the following security requirements for group based messaging:

- The MTC-IWF should verify if the SCS is authorized to send a group message to a given MTC group.
- The core network should be able to distinguish a group message from other messages.
- The group messages that are distributed to the group of MTC devices should be integrity protected, replay protected, and confidentiality protected.
- Local group ID should not to be exposed to an entity that is located outside the 3GPP network, including the SCS, which is outside the 3GPP network as well.

Candidate Solutions from 3GPP

According to the corresponding security requirements, the 3GPP has proposed the following candidate solutions for secure group based messaging.

Application layer based protection: Security protection applied at the MTC application layer is a straightforward solution. However, the network should trust the SCS and assure/ensure that the SCS protects the group message and MTC application if the MTC device verifies it. In this case, if the security is not applied in the application layer, then there can be attacks on the network. The SCS should apply encryption, signature, and replay protection to the group message. The MTC application on the MTC device should verify the source of the group message and ensure the integrity of the received group message. The MTC device should discard the group message if it is not signed and replay protected by the SCS.

Network based protection for cell broadcast: In network based protection, the MTC-IWF generates the keys for group message protection and protects the group message. Figure 4 [14] shows the message sequence and describes the mechanism.

1. The MTC-IWF creates the group and generates the group encryption key for encrypting the group message. The MTC-IWF uses the public key infrastructure (PKI) for signing the group message, and symmetric key (Gkey) is used for encryption/decryption of the group messages.
2. The MTC-IWF updates the HSS with the public key and the encryption key for a particular group with the group ID. The HSS maintains/maps the group based feature subscription details along with the MTC device subscription data.
3. During individual authentication, the MME fetches subscription data from the HSS. If the MTC device is subscribed for group based features, then the subscription data contains the group based feature information (GID, encryption key, public key, and the key index).
4. After successful authentication, the MME passes the group keys to the MTC device. The MME protects the keys using the non-access stratum (NAS) security context.
5. When the SCS wants to send the group message, it provides the group message over the Tsp interface.
6. The MTC-IWF protects the group message based on the group ID received from the SCS or from the HSS.
7. The MTC-IWF sends the protected group messages to the selected cell broadcast center (CBC). The protected group message includes the key ID and the algorithm ID used for protection.

Multimedia broadcast multicast service (MBMS) based method: MBMS security can provide a shared key for transferring data. Thus it can be used to protect the group message transferred from one MTC application server/MTC SCS to multiple MTC devices in the group when the MTC devices use shared secret keys for transferring. Otherwise, when all MTC devices in one group need to be authenticated together, or the MTC device wants to communicate with the MTC application server/MTC SCS/network individually, or MTC devices want to send uplink data, the current MBMS security solution cannot be applied.

	EPS-AKA [9]	EAP-AKA [10]	GAAKA-1	GAAKA-2	GAAKA-3
TOC	Symmetric	Symmetric	Symmetric	Asymmetric	Symmetric
FTS	Yes	Yes	Yes	No	Yes
PPR	Weak	Weak	Normal	Strong	Normal
RRA	No	No	Yes	Yes	Yes
RMA	No	No	Yes	Yes	Yes
CAH	No	No	Yes	Yes	Yes
CAM	No	No	No	Yes	Yes
CAE	No	No	No	Yes	Yes
CON	Large	Large	Medium	Small	Small
COM	Small	Small	Small	Large	Small
SGA	No	No	Yes	Yes	Yes

TOC: type of cryptosystem; FTS: follow the standard; PPR: privacy preservation; RRA: resistance to redirection attack; RMA: resistance to MITM attack; CAH: congestion avoidance at HSS; CAM: congestion avoidance at MME; CAE: congestion avoidance at eNB; CON: communication overhead of the core network; COM: computation overhead of MTC device; SGA: support group authentication.

Table 2. Comparison of design goals among the authentication and key agreement protocols.

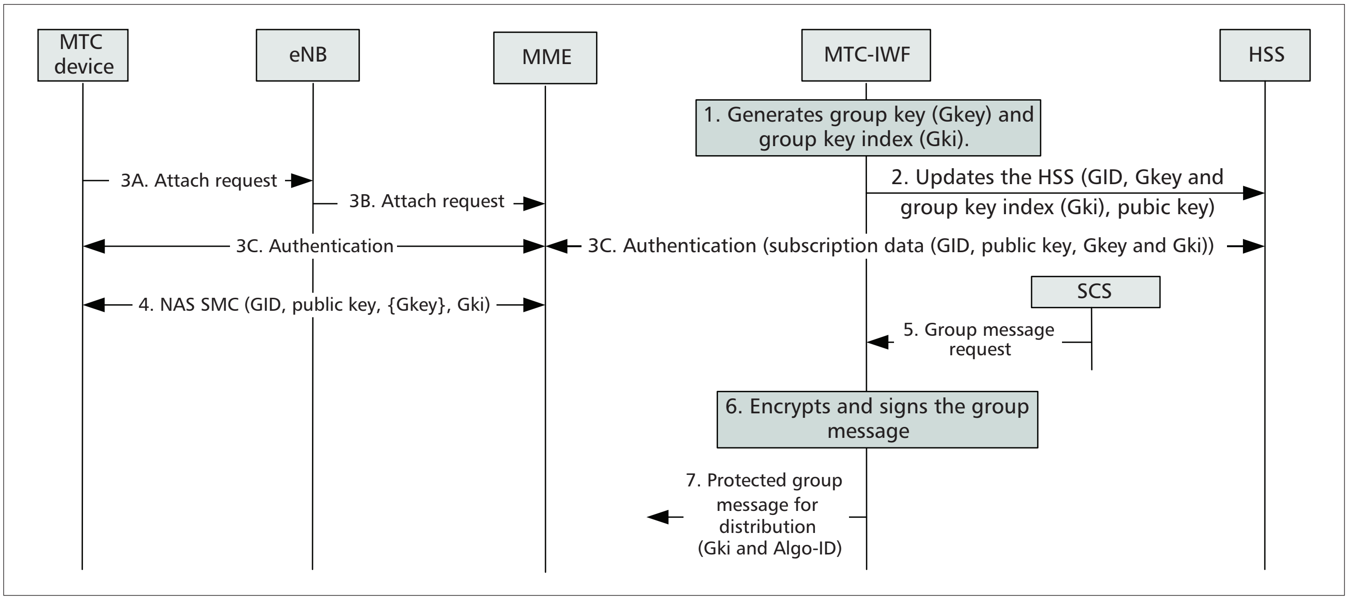

MTC device
eNB
MME
MTC-IWF
HSS
1. Generates group key (Gkey) and group key index (Gki).
2. Updates the HSS (GID, Gkey and group key index (Gki), pubic key)
3A. Attach request
3B. Attach request
3C. Authentication
3C. Authentication (subscription data (GID, public key, Gkey and Gki))
4. NAS SMC (GID, public key, {Gkey}, Gki)
SCS
5. Group message request
6. Encrypts and signs the group message
7. Protected group message for distribution (Gki and Algo-ID)

Figure 4. Network based protection for cell broadcast.

Research Challenges

In this section we present key issues and research directions related to group-based secure M2M communications.

Security, Privacy and Efficiency in Mobility Scenarios of MTC

To facilitate system optimization, the 3GPP defines a low mobility feature in M2M communications, which is suitable for MTC devices that do not move, move infrequently, or move only within a certain area. This feature enables the network operator to simplify and reduce the frequency of mobility management procedures. However, a tremendous number of Internet of Things (IoT) applications in M2M communications, such as fleet management or logistics management, have group-based behavior and high/frequent mobility. Therefore, new requirements for secure mobility management should be put forward. First, to reduce the computation and communication overhead during the move, MTC devices can form temporary groups based on the similarity of their mobility patterns at the location database. However, those MTC devices may not have a pre-established trust relationship and need to establish a temporary group without revealing group member information (i.e. privacy). This is difficult but desirable. Traditional schemes are based on a hierarchical tree, and any network entity that wants to set up a group needs to know the keys of the other group members. Therefore, some emerging cryptographic techniques, e.g. attribute-based cryptography, private set intersection, etc., can be considered to design a privacy-preserving group establishment scheme. In addition, when the MTC groups want to access the network, the new group-based access authentication and key agreement protocols should be studied due to introducing the high/frequent mobility scenario. To this end, fast group-based handover authentication protocols must be proposed.

Flexible and Efficient Group Key Management

3GPP SA2 has pointed out that group key management for application layer based protection is within the scope of 3GPP. Consequently, further research effort should be focused on addressing group key distribution. Generally, group key management schemes fall into two categories:

- Designed for large-scale (e.g. MTC message multicast) groups [15], with a one-to-many communication paradigm. Most of these schemes are centralized key distribution schemes that rely on a single fixed key server to generate and distribute keys to the group.
- Designed to support medium size tightly coupled dynamic peer groups, with a many-to-many communication paradigm.

It is worth noting that the two cases coexist in M2M communications. A group of MTC devices can access the core network to receive the same group message from the SCS, and can also communicate with each other to exchange massages. Therefore, designing a flexible and efficient group key management scheme for hybrid machine-to-machine networking is a desirable and challenging issue. However, traditional centralized key management is not well-suited for dynamic group communication systems, i.e. network partitions or faults may occur randomly. On one hand, the issues with centralized trust and single point of failure/attack should be avoided, and the requirements for strong security properties such as forward and backward secrecy, key independence, etc., should be fulfilled. On the other hand, to improve efficiency, new schemes should significantly reduce memory and computation overhead for each group member (i.e. suitable for a resource-constrained MTC device), efficiently deal with massive membership change by minimizing re-keying messages, and be efficient and very scalable for large-size MTC groups.

Conclusion

In this article we have investigated group-based security for large-scale M2M communications in 3GPP networks.

We first introduced the network elements and reference points of the MTC security architecture. We then identified three types of performance issues for authentication signalling congestion and overload in no/low mobility scenarios, and proposed three group access authentication and key agreement protocols to address them. Moreover, we provided several 3GPP candidate solutions for group message protection. Finally, we proposed future research directions with respect to group-based secure M2M communications. The research work should be useful for both mobile operators and MTC users.

> When the MTC groups want to access the network, the new group-based access authentication and key agreement protocols should be studied due to introducing the high/frequent mobility scenario. To this end, fast group-based handover authentication protocols must be proposed.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. This work is financially supported by the National Natural Science Foundation of China Research Grant (61272037, 61472472, 61402366, 61502386), the International Science and Technology Cooperation and Exchange Plan in Shaanxi Province of China (2015KW-010), and the Natural Science Basic Research Plan in Shaanxi Province of China (2015JQ6236).

References

[1] M. Chen, *et al.*, "A Survey of Recent Developments in Home M2M Networks," *IEEE Commun. Surveys Tut.*, vol. 16, no. 1, pp. 98–114, 2013.

[2] F. Ghavimi and H. Chen, "M2M Communications in 3GPP LTE/LTE-A Networks: Architectures, Service Requirements, Challenges and Applications," *IEEE Commun. Surveys Tut.*, 2014.

[3] T. Taleb and A. Kunz, "Machine Type Communications in 3GPP Networks: Potential, Challenges, and Solutions," *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 178–184, 2012.

[4] S. Lien, K. Chen, and Y. Lin, "Toward Ubiquitous Massive Accesses in 3GPP Machine-to-Machine Communications," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 66–74, 2011.

[5] H. Fu, *et al.*, "Group Mobility Management for Large-Scale Machine-to-Machine Mobile Networking," *IEEE Trans. Veh. Technol.*, 2014.

[6] R. Lu, *et al.*, "GRS: The Green, Reliability, and Security of Emerging Machine to Machine Communications," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 28–35, 2011.

[7] 3GPP TS 23.682 V13.0.0, Architecture Enhancements to Facilitate Communications with Packet Data Networks and Applications, Dec. 2014.

[8] 3GPP TS 22.368 V13.0.0, Service Requirements for Machine-Type Communications (MTC); Stage 1, Jun. 2014.

[9] 3GPP TS 33.401 V12.5.0, 3GPP System Architecture Evolution (SAE); Security Architecture, Sep. 2012.

[10] 3GPP TS 33.402 V12.5.0, 3GPP System Architecture Evolution (SAE); Security Aspects of Non-3GPP Accesses, Dec. 2014.

[11] C. Lai, *et al.*, "SE-AKA: A Secure and Efficient Group Authentication and Key Agreement Protocol for LTE Networks," *Computer Networks*, vol. 57, no. 17, pp. 3492–3510, 2013.

[12] C. Lai, *et al.*, "SEGR: A Secure and Efficient Group Roaming Scheme for Machine to Machine Communications between 3GPP and WiMAX Networks," in *Proc. IEEE ICC 2014*, pp. 1011–1016.

[13] ——, "LGTH: A Lightweight Group Authentication Protocol for Machine-Type Communication in LTE Networks," in *Proc. IEEE GLOBECOM 2013*, pp. 832–837.

[14] 3GPP TR 33.868 V12.1.0, Study on Security Aspects of Machine-Type Communications (MTC) and Other Mobile Data Applications Communications Enhancements, Jun. 2014.

[15] H. Zhang, *et al.*, "Optimal DoS Attack Scheduling in Wireless Networked Control System," *IEEE Trans. Control Syst. Technol.*, to appear.

Biographies

Chengzhe Lai [M'15] (lcz.xidian@gmail.com) received a B.S degree in information security from Xi'an University of Posts and Telecommunications in 2008, and a Ph.D. degree from Xidian University in 2014. He was a visiting Ph.D. student with the Broadband Communications Research (BBCR) Group, University of Waterloo from 2012 to 2014. At present he is with the School of Telecommunication and Information Engineering, Xian University of Posts and Telecommunications, and with the National Engineering Laboratory for Wireless Security, Xian, China. He is also a visiting researcher at the State Key Laboratory of Integrated Services Networks and State Key Laboratory of Information Security. His research interests include wireless network security, privacy preservation, and M2M communications security.

Rongxing Lu [S'09, M'11, SM'15] (rxlu@ntu.edu.sg) received a Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China in 2006, and a Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2012. He is currently an assistant professor with the Division of Communication Engineering, School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore. His research interests include wireless network security, applied cryptography, and trusted computing.

Hui Li [M'10] (lihui@mail.xidian.edu.cn) received his B.Sc. degree from Fudan University in 1990, and M.A.Sc. and Ph.D. degrees from Xidian University in 1993 and 1998, respectively. He is a professor with the School of Telecommunications Engineering, Xidian University, Xian, China. In 2009 he was with the Department of ECE, University of Waterloo as a visiting scholar. His research interests are in the areas of cryptography, security of cloud computing, wireless network security, information theory, and network coding. He is the co-author of two books. He served as TPC co-chair of ISPEC 2009 and IAS 2009, and general co-chair of e-forensic 2010, ProvSec 2011, and ISC 2011.

Dong Zheng (zhengdong@xupt.edu.cn) received an M.S. degree in mathematics from Shaanxi Normal University, Xian, China, in 1988, and a Ph.D. degree in communication engineering from Xidian University, Xian, in 1999. He was a postdoctoral fellow in the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, from 1999 to 2001, and a research fellow at Hong Kong University, Hong Kong, in 2002. He was a professor in the School of Information Security Engineering, Shanghai Jiao Tong University. He is also with the State Key Laboratory of Integrated Service Networks, Xidian University. He is currently a professor in the School of Telecommunication and Information Engineering, Xian University of Posts and Telecommunications, and is also connected with the National Engineering Laboratory for Wireless Security, Xian, China. His research interests include provable security and new cryptographic technology.

Xuemin (Sherman) Shen [M'97, SM'02, F'09] (xshen@bbcr.uwaterloo.ca) received his B.Sc. degree from Dalian Maritime University, China, in 1982, and his M.Sc. and Ph.D. degrees from Rutgers University, New Jersey, in 1987 and 1990, respectively, all in electrical engineering. He is a professor and University Research Chair in the Department of Electrical and Computer Engineering, University of Waterloo. His research focuses on resource management in interconnected wireless/wired networks, UWB wireless communications networks, wireless network security, wireless body area networks, and vehicular ad hoc and sensor networks. He has co-authored three books and has published more than 400 papers and book chapters in wireless communications and networks, control, and filtering. He is a former Editor-in-Chief of *IEEE Network* and served as a Technical Program Committee Co-Chair for IEEE INFOCOM 2014. He is the Chair of the IEEE ComSoc Technical Committee on Wireless Communications, and P2P Communications and Networking, and a voting member of GITC. He was a founding area editor of *IEEE Transactions on Wireless Communications*, and a guest editor for *IEEE JSAC*, *IEEE Wireless Communications*, and *IEEE Communications Magazine*. He also served as the Technical Program Committee Chair for GLOBECOM'07, Tutorial Chair for ICC'08, and Symposia Chair for ICC'10. He received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004, 2007, and 2010 from the University of Waterloo, and the Premier's Research Excellence Award in 2003 from the Province of Ontario, Canada. He is a registered professional engineer of Ontario, Canada, an IEEE Fellow, a Fellow of the Engineering Institute of Canada, and a Fellow of the Canadian Academy of Engineering. He has been a ComSoc Distinguished Lecturer.

The Importance of a Standard Security Architecture for SOA-Based IoT Middleware

The authors discuss the importance of defining a standard security architecture for SOA-based IoT middleware, analyze concepts and existing work, and make considerations about a set of security services that can be used to define a security architecture to mitigate the security threats in SOA-based IoT middleware systems.

Ramão Tiago Tiburski, Leonardo Albernaz Amaral, Everton de Matos, and Fabiano Hessel

COMMUNICATIONS STANDARDS

Abstract

The proliferation of the Internet of Things (IoT) in several application domains requires a well-defined infrastructure of systems that provides services for device abstraction and data management, and also supports the development of applications. Middleware for IoT has been recognized as the system that can provide this necessary infrastructure of services and has become increasingly important for IoT in recent years. The architecture of an IoT middleware is usually based on an SOA (service-oriented architecture) standard and has security requirement as one of its main challenges. The large amount of data that flows in this kind of system demands a security architecture that ensures the protection of the entire system. However, none of the existing SOA-based IoT middleware systems have defined a security standard that can be used as a reference architecture. In this sense, this article discusses the importance of defining a standard security architecture for SOA-based IoT middleware, analyzes concepts and existing work, and makes considerations about a set of security services that can be used to define a security architecture to mitigate the security threats in SOA-based IoT middleware systems.

Introduction

The Internet of Things (IoT) is a computing paradigm that basically aims to interconnect our everyday life objects (computing devices or things) using the Internet as the communication medium. Beyond Internet-based communication, IoT also aims to provide information processing capabilities to enable things to sense, integrate, and present data, reacting to all aspects of the physical world [1].

The IoT ecosystem is based on a layered architecture style and uses this view to abstract and automate the integration of objects, and to provide smart service solutions to applications [2]. In IoT, high-level system layers, such as the application layer, are composed of IoT applications and a middleware system, which is an entity that simplifies the development of applications by supporting services to cope with the interoperability requirement among heterogeneous devices.

IoT middleware systems have evolved from hiding network details from applications into more sophisticated systems to handle many important requirements, providing support for heterogeneity and interoperability of devices, data management, and security, just to name a few.

Although there are IoT middleware systems designed to be applied in specific areas of applications, most existing IoT middleware was designed to cover different domains, such as industry, environment, and society, and lack a generic and well-defined system architecture capable of allowing the interoperation of these different areas. The SOA (service-oriented architecture) standard has been used as a viable choice to cope with this gap, as it helps ensure high levels of system interoperability, providing system services that are based on devices and used by applications.

The deployment of a SOA-based IoT middleware in single or multiple domains requires the support of a security architecture, since the major problem faced by IoT middleware is the massive volume of data that flows into the middleware components, which creates security vulnerability points in the architecture of the middleware.

Even though there are SOA-based IoT middleware systems that have some kind of security in their architectures, the security approaches provided by these systems are specific for some application requirements, and in most cases do not present implementation details. In fact, none of the analyzed works provide a security architecture that can be used as a reference for the implementation of a security standard in SOA-based IoT middleware.

In this article we discuss the importance of having standard security architecture support for SOA-based IoT middleware. We also present security services that could be used to ensure the protection of the entire middleware regardless of the specific requirements of any application domain.

In the following sections we present the concepts regarding both SOA-based IoT middleware and security, as well as an overview of the existing work related to security in IoT middleware. We conclude the article with a definition of a standard security architecture and discuss the importance of having this standard used by SOA-based IoT middleware systems.

SOA-Based IOT Middleware

Current research trends regarding the pervasive and ubiquitous computing field consider IoT as the interconnection between things-embedded computing devices and the existing Internet and web infrastructure. Thus, IoT is expected to offer to users advanced connectivity of devices, systems, and services in a way that goes beyond machine-to-machine (M2M) communications, and that furthers the integration of things not only to the Internet (the network), but also to the web (application layer), allowing the develop-

The authors are with the Pontifical Catholic University of Rio Grande do Sul (PUCRS).

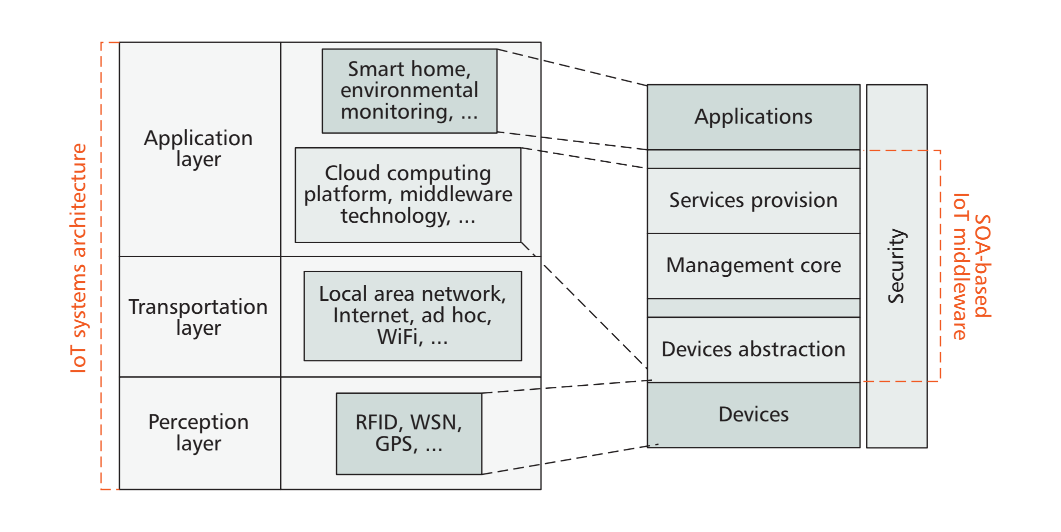


Figure 1. IoT systems architecture and SOA-based IoT middleware.

SOA-based techniques provide to IoT applications a uniform and structured abstraction of services for communication with IoT devices. On the other hand, ROA-based approaches realize the necessary requirements to make the devices (things) addressable, searchable, controllable, and accessible to IoT applications through the Web.

ment of things-oriented and service-based applications built upon a large number of networked physical elements [1].

The notion of service-based IoT systems has been realized according to the principles of SOA and ROA (resource-oriented architecture) architecture styles, which increasingly coexist in the IoT ecosystem, as ROA allows the deployment of lightweight SOA-based communication mechanisms embedded into resource-constrained IoT devices [3]. SOA-based techniques provide to IoT applications a uniform and structured abstraction of services for communication with IoT devices. On the other hand, ROA-based approaches realize the necessary requirements to make the devices (things) addressable, searchable, controllable, and accessible to IoT applications through the web.

IoT systems architecture can be divided into three layers [2], perception, transportation, and application, as shown in Fig. 1. Perception is the layer responsible for the recognition and control of physical devices, and also for the collection of the information provided by these devices. Transportation is the layer that provides ubiquitous network access for the elements of the perception layer. The application layer refers to the domains in which IoT applications can be developed. This layer is responsible for supporting the provision of services, and also for the realization of intelligent computation and logical resources allocation.

IoT middleware is a software layer or a set of sub-layers interposed between technological layers (perception and transportation layers) and application layers. The middleware's ability to hide the details of different technologies is fundamental to exempt the programmer from issues that are not directly pertinent to their focus, which is the development of specific applications enabled by IoT infrastructures [1]. In this way, IoT middleware has received much attention in recent years due to its major role of simplifying the development of applications and the integration of devices.

Many of the system architectures proposed for IoT middleware comply with the SOA approach. The adoption of SOA principles allows the decomposition of complex systems into applications consisting of a system of simpler and well-defined components. In SOA architecture, each system offers its functionality as standard services. Moreover, the SOA architecture supports open and standardized communication through all layers of web services [1].

An SOA structure for IoT middleware is also illustrated in Fig. 1. According to this structure, the applications layer allows end users to request information services and interact with the middleware. The devices layer can be composed of any IoT device that can connect to the middleware to provide services based on its features/resources. The devices abstraction layer can be embedded into both devices and middleware. Each service in the services provision layer is composed of one or more services from the devices. The devices function is abstracted into services by the devices abstraction layer and provided by the middleware through the services provision layer. The applications should use an API from the services provision layer to consume the provided services. All the processing activity is generated in the management core layer, also called the middleware core. The security layer must ensure security in all exchanged and stored data, since the middleware architecture enables some vulnerability points that can be explored by security threats.

Security Threats and Requirements

Security support is an important requirement for the correct behavior of SOA-based IoT middleware. It is mandatory to know what kind of attacks these systems can suffer and then be able to propose or decide which security mechanisms or countermeasures must be implemented in order to protect the entire system.

In Fig. 2 we present a security taxonomy for SOA-based IoT middleware which identifies the regions of attack and the security requirements for this system. According to the taxonomy, the

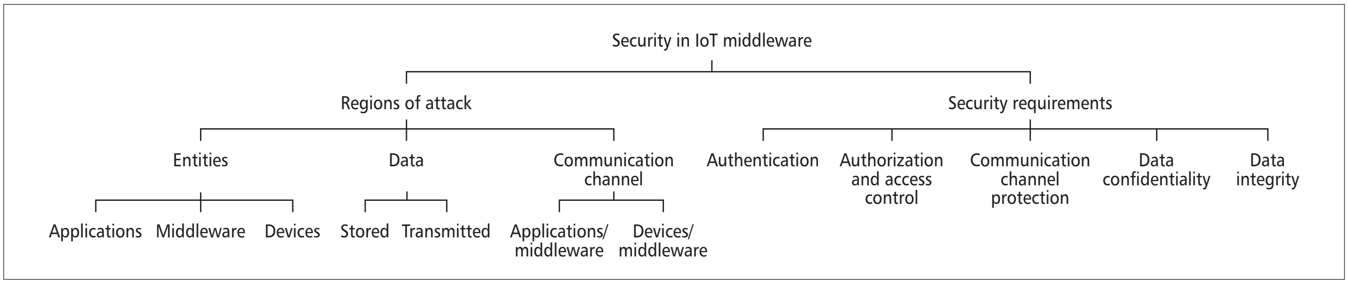


Figure 2. Security taxonomy for SOA-based IoT middleware.

Possible attacks	Authentication	Authorization and access control	Communication channel protection	Data confidentiality	Data integrity
Unauthorized access in entities	✓	✓			
Attacks in transmitted data				✓	✓
Attacks in stored data				✓	✓
Attacks in communication channels			✓		

Table 1. Relation between threats and requirements.

attacks can occur in three specific regions [4]: entities, data, and the communication channel.

Entities attacks are related to unauthorized access and physical attacks in applications, middleware, or devices. Data attacks can happen in two ways: when data are changed (e.g. tampering) or spied (e.g. eavesdropping) during the transmission between entities, and/or when the stored data are illegally modified or spied in the data repository (e.g. cloning or stealing credentials). Communication channel attacks can happen when the communication between system entities is attacked. IoT middleware has two communication channels, one with applications and another with devices. Both channels can be explored by attacks.

In order to protect IoT middleware from these attacks, security countermeasures must be developed and deployed in the middleware architecture. In the following sections we describe a set of security requirements (according to Fig. 2) that can be used to protect the middleware [4].

Authentication: This requirement must be provided for both applications and devices. It includes features such as credentials and trust management, and guaranteeing the correct identity of the application or device. The main function is to prevent unauthorized access.

Authorization and Access Control: This requirement is strongly related to authentication because when a device or an application is authenticated, the system must apply a correct set of rules to these entities, which determines their level of access to the system. Authorization is a mandatory requirement for applications and devices as they have different privileges for accessing specific resources and services of the middleware.

Communication Channel Protection: The role of this requirement is to protect the communication channels between applications/devices and middleware. The goal is to protect the data exchanged by entities against eavesdropping or tampering during transmission through the use of security protocols that must ensure communication channel protection independent of the security mechanisms used by them.

Data Confidentiality: This requirement involves the use of a cryptographic mechanism to preserve the exchanged data in the entire architecture of the middleware. Data confidentiality can also provide protection for data stored in entities (e.g. output data, requests, authentication credentials, etc.).

Data Integrity: This requirement is important to guarantee that an exchanged message has not been changed during the transmission by an unauthorized entity. This feature must validate the data and verify if the data has been violated during communication. This requirement can also be used to protect data stored in entities.

The relationship between potential attacks and security requirements is presented in Table 1. Regarding the attacks dedicated to entities, the unauthorized access can occur in applications, middleware, or devices, and must be prevented through authentication and access control mechanisms, which are strongly related. The middleware is responsible for controlling the security policies of these mechanisms, which should also protect the middleware from illegal access by applications or devices.

Regarding data, the vulnerability can occur when data is exchanged between entities. Data manipulation (or tampering) can be prevented with integrity checking, as this approach makes it possible to verify if data has been modified during transmission, and data leakage can be prevented with confidentiality mechanisms. Another vulnerability is related to data stored in entities, where threats (e.g. physical attacks) can attack these entities in order to access or modify the important data (e.g. authentication credentials and entities features). The stored data

SOA-based IoT middleware	Authentication	Authorization and access control	Communication channel protection	Data confidentiality	Data Integrity
SIRENA	✓		✓		
COSMOS	✓	✓		✓	
SOCRADES	✓	✓			
HYDRA	✓	✓	✓	✓	

Table 2. Security mechanisms implemented in SOA-based IoT middleware.

must be protected through data encoding and integrity checking mechanisms. Regarding the communication channel, violation of communication between both applications/devices and middleware should be prevented with channel protection protocols.

The relation between requirements and threats presented in Table 1 is intended to highlight the possible countermeasures that can be used to mitigate the mentioned security vulnerabilities. Moreover, it is possible to improve system security by adding more security levels for each possible attack (e.g. using authentication and authorization control to prevent attacks on the stored data).

Related Work

This section identifies the related works that have proposed security approaches regarding SOA-based IoT middleware. Further, it also presents some examples of security standards that can be used to mitigate security threats. The intention with this section is not to make an exhaustive comparison of existing work, but to identify security approaches and standards that can be used in the consolidation of the desired standard security architecture.

SIRENA middleware [5] concentrates its security approaches on communication channel protection and authentication for applications and devices. It uses the DPWS (Devices Profile for Web Services) technology in its framework, which defines a minimal set of implementation requirements to enable secure web service messaging on resource-constrained devices. DPWS uses TLS/SSL to establish a connection between applications and devices. Moreover, it uses the x.509.v3 certificate as a cryptographic credential to allow authentication.

COSMOS middleware [6] focuses its security approaches on authentication for applications and devices, access control, and data confidentiality. It has a module, called security manager, that controls access for sensor networks in the middleware. This module provides the protection of the system.

SOCRADES [7] focuses its security approaches on access control and authentication for applications and devices. In this work, devices and back-end services may only be accessed by clients that have certain authorization privileges and provide correct credentials for authentication.

HYDRA [8], one of the most consolidated middlewares, does not address functions related to data integrity. However, it implements all other security requirements, supporting security on different abstraction levels and using the middleware layer to build security by authentication, access control, communication, and data protection. HYDRA uses an access control policy framework, which is an implementation of the XACML (eXtensible Access Control Mark-up Language) processing model to ensure protection against unauthorized access [9].

As we can see in Table 2, most SOA-based IoT middleware addresses authentication and access control in their security approaches. On the other hand, data confidentiality and communication channel protection have low coverage, while data integrity is not cited in any work. A thorough comparison of the security mechanisms used in these systems was hindered since some of these works did not present information about the technologies used in their implementations.

We observed that none of the analyzed works propose solutions that span all the middleware security requirements. However, all these requirements are important in order to accomplish a standard security architecture. The use of well-defined and established standards that can be deployed in SOA architectures is essential to provide the intended security standardization. In this sense, researchers and industry have been proposing the definition of standards that can be used in this architecture [10–14].

Many IoT/M2M standards have been working on ROA architectures. The oneM2M group recently specified a security architecture for ROA-based M2M systems [10]. This specification consists of three security layers responsible for protecting all the architecture. They address the main important security requirements cited previously. Moreover, once ROA is coexisting with SOA in current IoT ecosystems, the well-defined security architecture of the oneM2M can assist in the composition of a standard security architecture for SOA-based IoT middleware.

In this sense, the work in [11] presents a snapshot of the latest progress in oneM2M standardization such as architecture, protocols, security aspects, device management, and abstraction technologies. They adopted a resource-based data model. All services are represented as resources, which are associated with the common service functions to support registration, configuration, management, and security. This kind of effort cooperates with the definition of a standard security architecture.

The CoRE (Constrained RESTful Environments) Working Group within the IETF (Internet Engineering Task Force) has defined the

We observed that none of the analyzed works propose solutions that span all the middleware security requirements. However, all these requirements are important in order to accomplish a standard security architecture. The use of well-defined and established standards that can be deployed in SOA architectures is essential to provide the intended security standardization.

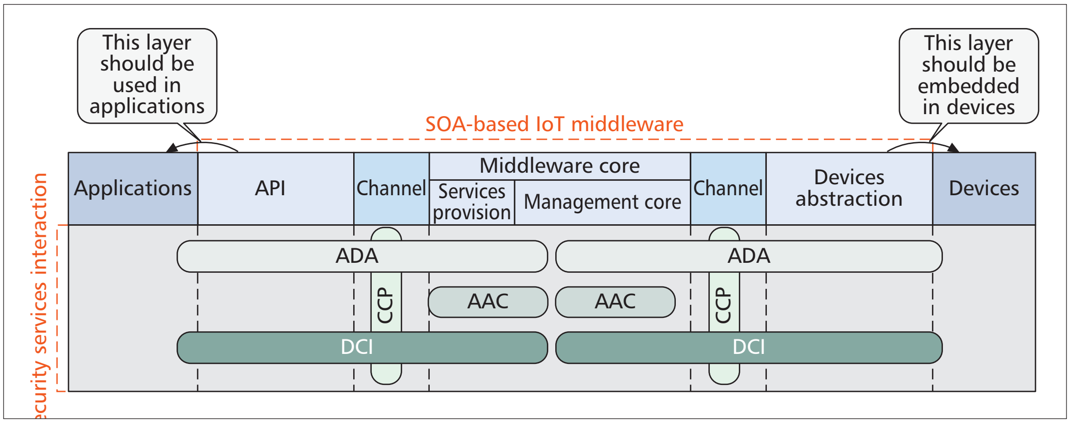


Figure 3. Security on SOA-based IoT middleware.

Custom security approaches offered by the IoT research community mostly offer specific improvements or solutions. However, these approaches rarely help to understand the big picture for securing a SOA-based IoT middleware system.

Constrained Application Protocol (CoAP) as a generic web-based protocol for RESTful-constrained environments, targeting M2M applications that cope with HTTP for integration with the existing web [12]. In CoAP, the universal resource identifier (URI) is used to access the resources on a given host.

CoAP is a relatively simple request and response protocol providing both reliable and unreliable forms of communication. To protect the transmission of sensitive data, secure CoAP mandates the use of Datagram Transport Layer Security (DTLS) as the underlying standard security protocol to guarantee communication channel protection on a point to point basis [13].

Another important technology that has become the choice for implementing SOA architectures is web services. This communication standard provides a framework for systems integration independent of programming language and operating system. However, the security of web services depends not only on the security of the services themselves, but also on the confidentiality and integrity of the XML (eXtensible Markup Language) based SOAP (Simple Object Access Protocol) messages used for communication [14].

XML Signature and XML Encryption are used in web services to provide integrity and confidentiality, respectively, and can be reused to provide SOAP security as well. WS-Security specifies how to effectively apply XML Signature and XML Encryption to the SOAP standard, providing integrity and confidentiality to SOAP messages. In addition, WS-Security provides a mechanism to avoid replay attacks and a way to include security tokens in SOAP messages. Security tokens are typically used to provide authentication and authorization.

Standard Security Architecture Definition

In order to promote a secure interoperability between system entities in an IoT environment controlled by SOA-based middleware, the security services to be used by each entity should preferably be based on established standards. In this way, the correct choice of the "common set of security standards" that can be used in the addressed secure architecture is a step forward toward the definition of the "standard security architecture". Developing the required standard for SOA-based IoT middleware guarantees confidentiality, integrity, communication channel protection, authenticity, and access control for all entities and data exchanged and stored in the middleware.

The set of security requirements that is mandatory for SOA-based IoT middleware can be summarized in four security services: application and device authentication (ADA); authorization and access control (AAC); data confidentiality and integrity (DCI); and communication channel protection (CCP). We believe the implementation of these services in a security architecture for SOA-based IoT middleware would help to protect the system against the threats described earlier.

Figure 3 identifies where each security service could be embedded in the middleware in order to ensure security in the whole system architecture. ADA and DCI services could be provided in all layers of the middleware. The CCP service could be responsible for protecting the communication channels between these layers, and the AAC service could be presented only in the middleware core. The security services for applications and devices have distinct boundaries, so they are separated in the system architecture. Next we present a brief definition for each service in order to explain how they could provide security for a SOA-based IoT middleware.

The ADA service would be responsible for enabling the authentication of applications and devices in the middleware core. The main part of the authentication service would be contained in the middleware core. However, some techniques and methods used for authentication by applications and devices would be contained in the API layer and devices abstraction layer, respectively.

The AAC service would be responsible for allowing access to information by an authenticated application or device. It would be present in the middleware core and also would be directly related to the authentication service (ADA). AAC should prohibit the unauthorized access of any object or information present in the middleware core.

The CCP service would be responsible for protecting the communication channel in order to ensure the protection of data sent by devices or applications against eavesdropping and/or tampering during the transmission. It can be provided by the implementation of security protocols and could take place between the API/devices abstraction and middleware core layers.

The DCI service would involve all layers of the architecture. Part of this service could be based on data encoding and decoding techniques, which would be responsible for encrypting and decrypting data in all layers. This service could also be responsible for protecting the stored data on entities. The other part of this service is related to data integrity techniques and would be responsible for allowing integrity checking of the transmitted data in the network (between application/devices and middleware). Moreover, it could also be used to protect the stored data.

The fact that the API is used by applications and the devices abstraction layer is embedded into the devices should provide protection for the entire system architecture. This feature is mandatory in middleware architecture since the middleware layers are distributed between applications and devices, which enables the use of confidentiality and integrity services to ensure data authenticity during the entire data life cycle.

Discussion

Security for SOA-based IoT middleware encompasses procedures that include: embedding keys in system entities; generation of new keys; establishing access control policies together with authentication to allow access to networks, services and data; usage of security services to protect data against tampering and eavesdropping; and the development and selection of efficient cryptographic methods. Custom security approaches offered by the IoT research community mostly offer specific improvements or solutions. However, these approaches rarely help to understand the big picture for securing an SOA-based IoT middleware system.

A standard security architecture that provides a full stack of security services composed of authentication, authorization, integrity, communication channel protection, and confidentiality should be defined. Many security protocols have already been standardized, and adapting them to be used in different IoT environments would be beneficial for the standard definition. Moreover, a standardized architecture when deployed on an SOA-based IoT middleware could interoperate more easily with existing Internet/web infrastructures and services.

The importance of having a standard security architecture is strongly related to the definition of the set of security standard services that can be applied in the secure architecture to enable real interoperability between system entities. Thus, every entity can be developed according to a common knowledge of the desired security requirements.

The services proposed previously could compose a SOA-based security architecture and would be useful to ensure protection against the attacks mentioned previously. A standard security architecture is intended to ensure protection for an entire system, providing a standard that could be used by other middleware systems as none of the security mechanisms mentioned in this article are able to ensure, by themselves, adequate protection for an entire system with different application requirements.

The use of the security standards protocols described previously could be the basis for the definition of a standard security architecture. XML and web services standards provide a flexible framework to comply with basic security requirements such as confidentiality, integrity, and authentication, as well as more complex requirements such as non-repudiation, authorization, and identities. Furthermore, these standards offer high flexibility with respect to the cryptographic algorithms used, facilitating the adaptation of complex algorithms if required. In this sense, the DTLS protocol can provide protection for the communication channel, while CoAP can ensure a secure interoperation between entities or systems.

Although the definition of a standard security architecture for SOA-based IoT middleware is noteworthy, it is mandatory to be careful with the way in which the involved protocols can be deployed, since some IoT devices may not have sufficient computing resources to perform complex security mechanisms. An important challenge related to IoT middleware, which is also common for M2M platforms, is the necessity of having lightweight security solutions [2, 13]. Providing solutions such as key management, authentication, access control, confidentiality, and integrity is considered a significant challenge, mainly when applied in resources-constrained environments such as the physical infrastructure of devices of the IoT.

In this sense, the SOA-based IoT middleware architecture presented previously can coexist with an ROA-based architecture to ensure the abstraction of devices as resources in order to be used by the middleware management layer. This is possible because we can embed lightweight web services on devices and offer them as resources.

Regarding security, the coexistence of SOA and ROA creates a new set of security requirements that must be followed for resource-constrained environments in order to ensure system protection. These requirements are security mechanisms that are suitable for device constraints and which must ensure device protection. To accomplish this goal, the devices must implement a well defined security architecture that addresses the security requirements and that can interact with the top layer of the IoT ecosystem (middleware core).

In this way, the use of ROA-based security architecture standards in the devices would be one of the keys to ensure protection along with the rest of the SOA-based security architecture. We could use security mechanisms such as authentication, access control, channel protection, confidentiality, and integrity in order to ensure the protection of each device according to its needs. With this approach, we would be ensuring security and abstraction of resources (devices) that connect and interact with the middleware, and also maintaining the service-oriented architecture for applications and for the middleware itself. Therefore, we believe that a well defined security architecture, as presented in [10, 11], along with other security standards [12–14], can coexist with an SOA-based security architecture to provide a standard security archi-

Although the definition of a standard security architecture for SOA-based IoT middleware is noteworthy, it is mandatory to be careful with the way in which the involved protocols can be deployed, since some IoT devices may not have sufficient computing resources to perform complex security mechanisms.

tecture for SOA-based IoT middleware. In this sense, it is important to find an efficient strategy to deal with the massive data generated by IoT systems and to provide secure services to efficiently handle, organize, and protect all this information.

Conclusion

SOA-based IoT middleware has an important role in implementations of IoT. However, to apply this kind of system in different application domains, ensuring interoperability and integration of devices in a secure way, requires the definition of a standard security architecture for the middleware system. This article discussed the importance of a standard security architecture for SOA-based IoT middleware. We identified reasons why a standard security architecture would be helpful. We also discussed the security services and protocols that can be applied in an SOA-based IoT middleware to provide security in the entire architecture of the middleware, as well as to enable its secure use in different domains. Finally, we discussed the importance of lightweight security services for the consolidation of the standard security architecture.

SOA-based IoT middleware has an important role in scenarios of IoT. However, to apply this kind of system in different application domains, ensuring interoperability and integration of devices in a secure way, requires the definition of a standard security architecture for the middleware system.

Acknowledgments

Our thanks to CAPES/CNPq for funding within the scope of the project numbers 058792/2010 and 382169/2014-0.

References

[1] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A Survey," *Computer Networks*, vol. 54, no. 15, 2010, pp. 2787 – 2805.

[2] Q. Jing, A. Vasilakos, J. Wan, J. Lu, and D. Qiu, "Security of the Internet of Things: Perspectives and Challenges," *Wireless Networks*, vol. 20, no. 8, 2014, pp. 2481–2501.

[3] D. Guinard, V. Trifa, and E. Wilde, "A Resource Oriented Architecture for the Web of Things," *IEEE Internet of Things (IOT)*, 2010, pp. 1–8.

[4] oneM2M-TR-0008, "oneM2M Technical Report v1.0.0," Tech. Rep., 2014.

[5] H. Bohn, A. Bobek, and F. Golatowski, "SIRENA — Service Infrastructure for Real-Time Embedded Networked Devices: A Service-Oriented Framework for Different Domains," *Int'l. Conf. Networking, Int'l. Conf. Systems and Int'l. Conf. Mobile Communications and Learning Technologies*, Apr. 2006, pp. 43–43.

[6] M. Kim *et al.*, "COSMOS: A Middleware for Integrated Data Processing over Heterogeneous Sensor Networks," *ETRI J.*, vol. 30, no. 5, 2008, pp. 696–706.

[7] P. Spiess *et al.*, "SOA-Based Integration of the Internet of Things in Enterprise Services," *IEEE Int'l. Conf. Web Services*, July 2009, pp. 968–75.

[8] A. Badii *et al.*, "HYDRA: Networked Embedded System Middleware for Heterogeneous Physical Devices in a Distributed Architecture," *Final External Developers Wksps. Teaching Materials*, 2010.

[9] A. Badii, M. Crouch, and C. Lallah, "A Context-Awareness Framework for Intelligent Networked Embedded Systems," *3rd Int'l. Conf. Advances in Human-Oriented and Personalized Mechanisms, Technologies and Services (CENTRIC), IEEE*, 2010, pp. 105–10.

[10] oneM2M-TS-0003, "oneM2M security solutions," Tech. Rep., 2014.

[11] J. Swetina *et al.*, "Toward a Standardized Common M2M Service Layer Platform: Introduction to oneM2M," *IEEE Wireless Commun.*, vol. 21, no. 3, June 2014, pp. 20–26.

[12] S. L. Keoh, S. Kumar, and H. Tschofenig, "Securing the Internet of Things: A Standardization Perspective," *IEEE Internet of Things J.*, vol. 1, no. 3, June 2014, pp. 265–75.

[13] S. Raza *et al.*, "LITHE: Lightweight Secure CoAP for the Internet of Things," *IEEE Sensors J.*, vol. 13, no. 10, Oct 2013, pp. 3711–20.

[14] N. Nordbotten, "XML and Web Services Security Standards," *IEEE Commun. Surveys Tutorials*, vol. 11, no. 3, 2009, pp. 4–21.

Biographies

Ramão Tiago Tiburski (ramao.tiburski@acad.pucrs.br) received his B.Sc. degree in computer science from the University of Passo Fundo. He is an M.Sc. student of computer science at Pontifical Catholic University of Rio Grande do Sul (PUCRS), studying at the Embedded System Group (GSE). He has experience in computer science and his research interests are IoT, middleware, and security in IoT environments.

Leonardo Albernaz Amaral (leonardo.amaral@acad.pucrs.br) received his B.Sc. degree in systems analysis from UCPEL, and M.Sc. and Ph.D. degrees in computer science from PUCRS. He is a research leader and project manager at GSE/PUCRS. He is currently a researcher in the Graduate Program in Electrical Engineering and Computer Science, both at PUCRS. He has experience in computer science, with an emphasis on middleware systems, RFID, IoT, smart cities, and pervasive systems for healthcare environments.

Everton de Matos (everton.matos.001@acad.pucrs.br) received his B.Sc. degree in computer science from the University of Passo Fundo. He is an M.Sc. student of computer science at Pontifical Catholic University of Rio Grande do Sul (PUCRS), studying at the Embedded System Group (GSE). He has experience in computer science and his research interests are IoT, middleware for IoT, and context aware in IoT environments.

Fabiano Hessel (fabiano.hessel@pucrs.br) is an associate professor of computer science at PUCRS, Brazil. He received his Ph.D. in computer science from UJF, France (2000). He is the head of the Embedded System Group (GSE). He was general chair and program chair, and a participant, in several technical/program committees of prestigious conferences and journals. He has had several publications in prestigious conferences, journals, and books. His research interests are embedded real-time systems, RTOS, and MPSoC systems applied to IoT/SmartCities.

SCALE: Safe Community Awareness and Alerting Leveraging the Internet of Things

The authors propose the Safe Community Awareness and Alerting Network (SCALE), a cyber-physical system (CPS) leveraging the pervasive Internet of Things (IoT) to extend a smarter, safer home to all residents at a low incremental cost. SCALE uses novel networking technologies, commodity sensor devices, cloud services, and middleware abstractions to sense, analyze, and act on sensed events in a distributed manner.

Kyle Benson, Charles Fracchia, Guoxi Wang, Qiuxi Zhu, Serene Almomen, John Cohn, Luke D'Arcy, Daniel Hoffman, Matthew Makai, Julien Stamatakis, and Nalini Venkatasubramanian

Kyle Benson, Guoxi Wang, Qiuxi Zhu, and Nalini Venkatasubramanian are with University of California.

Charles Fracchia is with BioBright.

Serene Almomen and Julien Stamatakis are with Senseware, Inc.

John Cohn is with IBM.

Luke D'Arcy is with Sigfox.

Daniel Hoffman is with Montgomery County.

Matthew Makai is with Twilio.

COMMUNICATIONS STANDARDS

Abstract

We propose the Safe Community Awareness and Alerting Network (SCALE), a cyber-physical system (CPS) leveraging the pervasive Internet of Things (IoT) to extend a smarter, safer home to all residents at a low incremental cost. SCALE uses novel networking technologies, commodity sensor devices, cloud services, and middleware abstractions to sense, analyze, and act on sensed events in a distributed manner. It monitors environmental factors (i.e. smoke, explosive gas) and automatically alerts residents via phone upon discovery of a possible emergency, enabling them to confirm the event and contact emergency dispatchers with minimal effort. This article describes the inception, design, development, and deployment of a prototype system to achieve these goals. We discuss lessons learned and future directions for general CPS/IoT platforms.

Introduction

With the increasing pervasiveness of computers in our daily lives, the Internet of Things (IoT) concept is transitioning from a future prediction to real-world deployments. With this manifestation comes a myriad of possible applications, from manipulating devices in our homes to large-scale automation of industries and public utilities. A common human-facing aspect of each of these applications is that they aim to improve our quality of life through inexpensive, commonly available technology. While home security systems have existed for decades, they are rather expensive services, and only in recent years have we seen components become cheap and available enough that hobbyists experiment with do-it-yourself systems. So it seems natural that an open system, made possible with these recent advances, should be created to improve the lives of under served populations that previously could not afford such advanced home security and safety monitoring systems. This motivation led our team, assembled in response to the SmartAmerica Challenge,[1] to envision, design, build, and demonstrate the Safe Community Awareness and Alerting Network (SCALE).

SCALE aims to improve the safety of residents through the use of modern connected devices and computer systems, particularly lower-income and elderly residents who often do not have access to advanced technologies such as home security systems, smartphones, and computers with Internet connections. To accomplish this goal, we designed an event-driven distributed system to sense safety-related data from devices in homes or on individuals, analyze it locally or within the cloud to detect possible emergency events, and automatically contact individuals (e.g. homeowners, caretakers, even emergency dispatchers) to notify them and confirm if there is indeed an emergency. We implemented a prototype of this system and deployed it in Montgomery County, Maryland, USA to enable rapid integration of components and testbeds from different partners.

The immediate goals of the SCALE project are:

- Demonstrate our ability to extend a connected safe home to everyone at a low incremental cost.
- Jump-start a live testbed for identifying and researching IoT challenges (e.g. middleware, networking, etc.).
- Identify suitable sensors, data schemas, and algorithms for detecting possible emergency events.
- Implement and test workflows for cloud-based analytics and alerting.
- Demonstrate an open data platform for connecting disparate systems with minimal coordination.

System Architecture

SCALE devices upload sensed events, and the analytics service looks for possible emergencies, it sends residents emergency alerts to confirm or reject, and interested individuals (i.e. emergency dispatchers) visualize events through a dashboard. This section discusses the high-level requirements, logical components, architectural design decisions, and implementation details of the system prototype. It first discusses a *cloud data exchange* and then the components of the system that perform *sensing*, *analysis*, and *actuation*.

Cloud Data Exchange for IoT

To facilitate machine-to-machine (m2m) communication for exchanging IoT data in SCALE (sensed events, analytics, alerts, etc.), we propose the Data in Motion Exchange (DIME) system, shown in Fig. 1. We envisioned DIME as an open communications hub for IoT that simplifies the development and deployment processes.

DIME allows any device or service to publish or subscribe to any other data feed, regardless of the protocols used at the device level. This simple loose coupling enables developers to incorporate new services and devices without the need to modify existing ones. This simplifies system evolution, and it also creates a level playing field for innovation. Any party can introduce new

[1] http://smartamerica.org/

0163-6804/15/$25.00 © 2015 IEEE

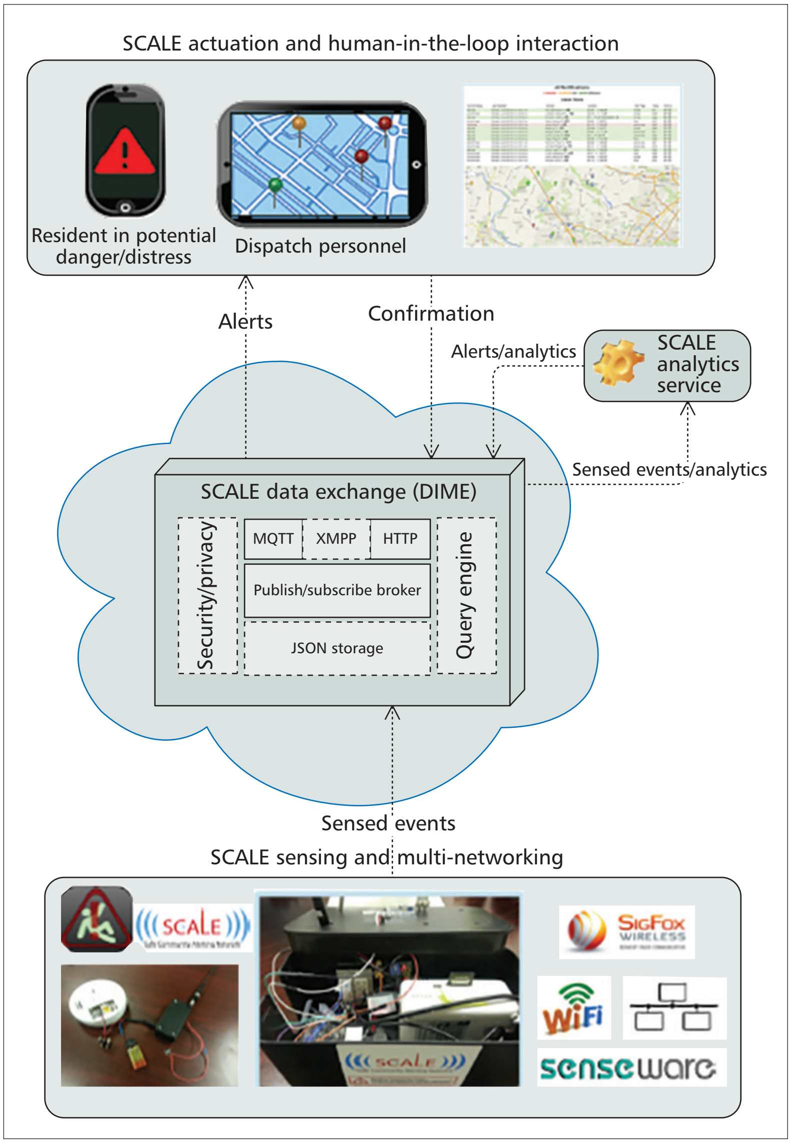


Figure 1. DIME facilitates the exchange of data between main SCALE components. DIME Components shown in solid boxes have been implemented, and those in dashed boxes remain as future work.

capabilities, or improvements to existing ones, to the system with minimal need for coordination among current components. They can perform analysis on sensed data, or even higher-level events, and contribute the results back to the exchange, driving science and innovation faster as more devices connect.

Sensed Event Data: To build an exchange for IoT data, we first defined the type of data that DIME should handle. We decided to treat raw sensed data and higher-level events equivalently. This aligns with our concept of *virtual sensors*, previously proposed in [1]. *Virtual sensors* abstract low-level data by processing sensor data streams, which may be directly or indirectly derived from physical sensor devices, and exposing higher-level semantics through advanced analytics.

Recognizing the rich amount of information contained within a higher level event as well as subtle device differences that affect lower-level events, we wanted a well-adopted flexible schema that could allow, but not require, inclusion of additional information fields beyond what is necessary to convey the sensor reading. These additional fields should not break the schema or require all entities in the system to understand them.

For simplicity and flexibility, we opted to use JSON to format the data for transmission to the broker, as it provides a commonly-used self-describing format supported by mature software modules. We defined what we thought was a reasonable starting point for the schema. It includes information about the platform (hardware, operating system, etc.), sensor (device type, identifier, etc.), data (units, value, timestamp, priority, etc.), a pointer to the specific schema in use to facilitate interoperability, and any other miscellaneous domain-specific information that developers want to include. One should note that we do not believe this schema to be comprehensive; rather, we envision a system where different domains could define their own schema and publish information about how to interpret it so as to encourage interoperability between vendors/systems.

Current Implementation: In its current form, DIME uses MQTT,[2] a fast, lightweight, publish-subscribe-style protocol. It was developed by IBM for lightweight telemetry, donated to open source, and has since gained popularity for use as an m2m protocol for IoT data. The publish-subscribe model allows multiple servers to collect data from DIME and multiple clients to send it without requiring any configuration on our part. The DIME server currently uses the open source Eclipse Paho MQTT broker.[3] While Paho could be run anywhere, we used IBM's MessageSight[4] software appliance, which handles millions of concurrent data streams, running on the IBM SoftLayer Cloud.

In DIME, sensor data is published to a particular topic, which consists mainly of a device identifier and sensed event type. Other services, such as the SCALE server, subscribe to this data by a particular device, sensor type, or just to all data.

For compatibility, DIME also provides a RESTful interface, implemented via HTTP, initially residing on the SCALE server for ease of deployment. This interface translates incoming data into the proper format and publishes them via MQTT. In this manner, we quickly implemented DIME as a simple MQTT server, though we plan to extend it to directly support other protocols (e.g. HTTP and XMPP).

Sensing

This section describes the development and deployment of several SCALE clients that sense, minimally analyze, and report data to DIME for ingestion by the analytics server.

Networking Technologies: To support a heterogeneous mix of devices and improve the client's flexibility in deployment, we integrated multiple networking technologies. In addition to the standard Wi-Fi and Ethernet connections, the clients supported ultra-narrowband (UNB) wireless adapters. UNB allows for long-range, low-power, low-bandwidth uplinks. Sigfox provided a UNB basestation to install in Montgomery County. We

[2] http://mqtt.org.

[3] http://www.eclipse.org/paho/

[4] http://www-03.ibm.com/software/products/en/messagesight

were able to deploy several SCALE devices with Sigfox UNB adapters in Rockville, Maryland, and send data to DIME via the basestation from up to several kilometers away, despite using lower-powered basestation and client adapter antennas.

Sigfox adapters send data in 12-byte packets, so MQTT was not an option. Instead, we coded the data to fit within this packet and created the aforementioned HTTP interface where Sigfox directed this data. We also integrated Senseware's proprietary mesh networking solution into the SCALE system, as described below.

Hardware Platforms: We wanted a flexible client platform to allow deploying heterogeneous sensors, devices, and networking technologies. Some clients may plug into a stable power source and Internet access to support a multitude of sensors and more advanced local data analytics, while others may be battery-powered and just upload raw sensed data via wireless. To support the pervasiveness of these systems and address the latter of these device types by reducing reliance on home Internet access, crucial for our mission to support under served populations, we aimed to integrate platforms and technologies that could provide long-range low-power connections. To address both styles, we chose to use commodity off-the-shelf components wherever possible, which had additional benefits of reducing infrastructure costs; increasing the number of possible integrated devices and sensors; reducing development costs by leveraging extensive community support; and allowing other researchers, hobbyists, and new team members to easily understand our design so that they may copy and extend it.

We first built a general-purpose sensor box named FlexSCALE that supports many different sensors and network adapters. The compute units and sensors are housed within a large cable box to protect wires and maintain a cleaner facade. Environmental sensors (e.g. light and temperature) were fastened on top so they protruded from holes in the lid, gaining external access with minimal wiring exposed. The initial version housed both a Raspberry Pi and a Sheevaplug, each running a form of Debian Linux, as the compute units. We transitioned to just using the Raspberry Pi to simplify platform support and handle a greater variety of peripherals thanks to I/O ports and pins other than USB and Ethernet.

Each FlexScale box has light (luminance), explosive gas, passive infrared (motion detection) sensors, an accelerometer (acting as a seismograph), and thermometer as well as a Wi-Fi dongle and a Sigfox UNB adapter. A powered USB hub supported the two USB sensors and two USB network adapters on the Sheevaplug and older Raspberry Pis.

In contrast with the larger and more extensible FlexSCALE Box, we experimented with dedicated devices to monitor a single sensor and report its readings with almost no analysis. We were particularly interested in retrofitting existing household sensing devices and connecting them with the SCALE service. Therefore, we modified an off-the-shelf 9-volt smoke detector and attached it to an Arduino Micro for the purpose of monitoring the voltage level of the battery. See Fig. 2 for the wiring diagram used for this device. The Arduino constantly sends (every ~4 sec.) the measured voltage level to DIME via a Sigfox adapter. If this level drops significantly, indicative of the alarm going off, the server sends an alert. The theory here is that the alarm consumes more power than just the sensor itself and so the additional function drops the voltage level of the battery significantly. The Arduino and Sigfox devices fit into a small project box, similar in size to a mint tin.

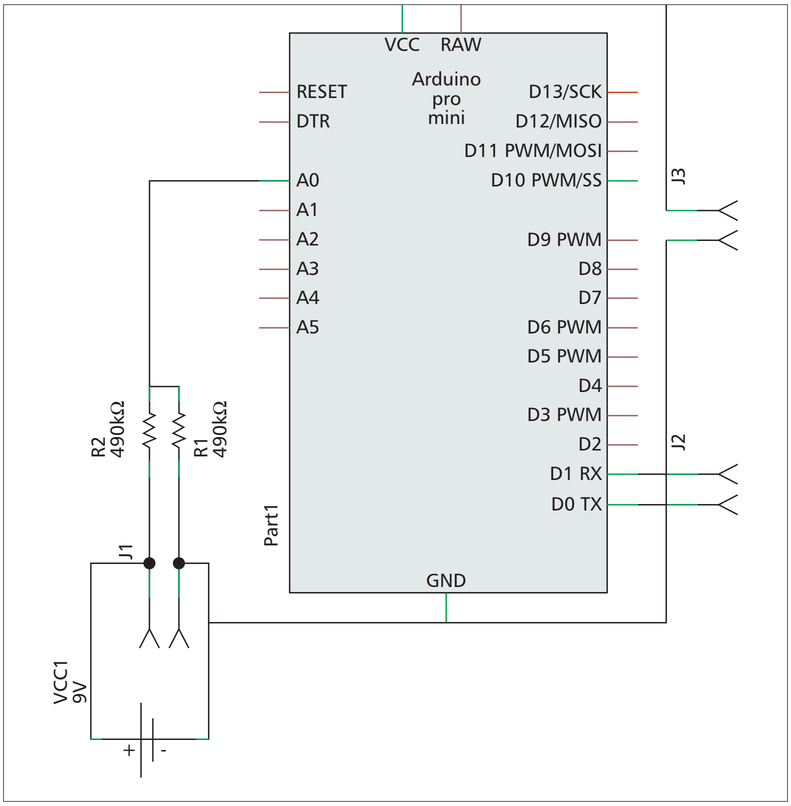


Figure 2. Wiring diagram for the hacked smoke detector device.

To complement the aforementioned dedicated and flexible sensing platforms, we also built an Android application for personal fall detection. It analyzes the device's accelerometer readings using the algorithm presented in [2]. Upon detecting a user falling, the application presents them with an option to cancel the alert, thus preventing false alarms, before a countdown timer expires and the phone publishes the alert via MQTT to call for help.

To test and showcase how existing proprietary systems could integrate with DIME and SCALE with minimal modifications, we partnered with Senseware, a Virginia-based startup. They build modular sensor devices that transmit data via mesh networks to a gateway for upload to a web-based cloud service. The user-friendly devices are easy to deploy and can have a variety of connected sensors (i.e. air quality, humidity), making them an ideal candidate to expand the SCALE testbed with commercial hardware. Senseware integrated their sensors' data by forwarding it to a Senseware-specific HTTP endpoint to facilitate this connection, similar to how we integrated Sigfox devices.

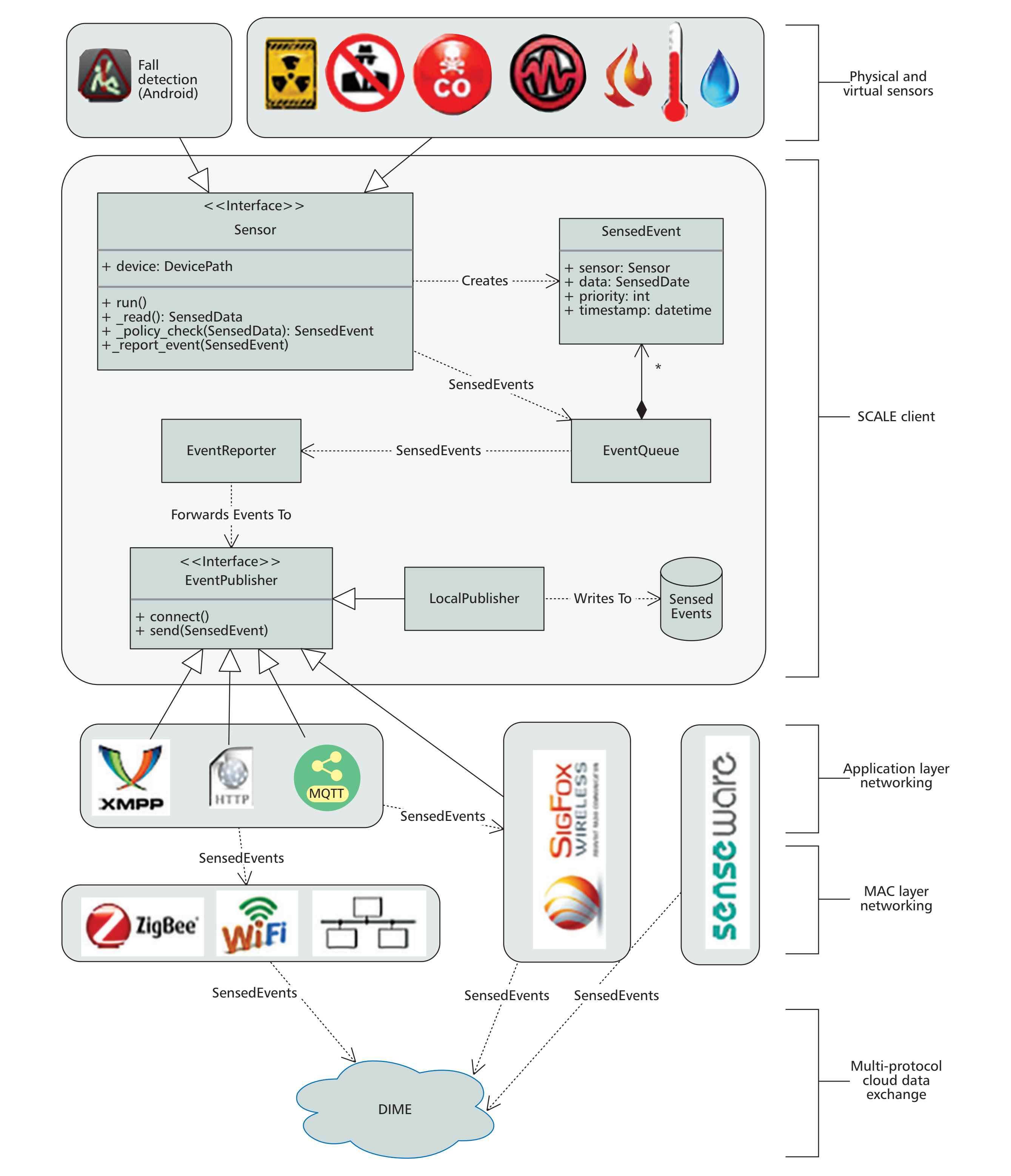


Figure 3. The SCALE Client architecture.

Software Design: We wanted a cross-platform extensible software package that runs on the majority of devices. This package should be modular and support plugging in different component implementations (e.g. new sensors or network protocols) without disrupting other modules. Adding or changing hardware components should not require any software changes, but should rather be handled through a simple configuration file.

Figure 3 shows the prototype FlexSCALE software we built to address the above requirements. This Python package connects with various sensor devices attached to the compute system, records data, and publishes events according to some policy. Data originates at an instantiation of the abstract sensor class, which allows us to rapidly connect new sensor types and define new *virtual sensors*. Sensors create SensedEvents, which encapsulate the sensor data schema described earlier, and place them in a queue for reporting to DIME or further analysis by relevant VirtualSensors.

Each networking protocol that connects the client to DIME is abstracted with a concrete instantiation of the EventPublisher class. Similar to adding new sensors, this allows us to easily add new protocols and API endpoints with minimal additional code. It currently supports MQTT via Wi-Fi or Ethernet, Sigfox ultra-narrowband (UNB), and local storage. EventPublishers also provide a degree of control over quality of service (QoS), currently just in the form of transmitting higher-priority events first. We added this feature early on to address the UNB transmitters' low bandwidth.

We used SaltStack[5] for configuration management: remotely deploying and updating software on the sensor boxes. We chose SaltStack because it is highly scalable, supports redundant master servers, and (most importantly) connects with devices deployed behind network address translators (NATs) as are commonly found in residential homes.

Analytics

The SCALE analytics service monitors sensor data and events streaming from DIME and publishes detected emergency events, which may trigger alerts to individuals when appropriate as described earlier. Refer to Fig. 4 and the description below for how we designed and implemented the analytics server.

We implemented the analytics engine as an asynchronous event-driven Python server that acts on sensed events in accordance with their type using appropriate event-handlers. Thus, adding new sensor and event types only requires additional programming by end application developers, not those responsible for server development.

The server, deployed on the IBM BlueMix platform, receives sensed data through Eclipse Paho's MQTT client[6] and routes it to the appropriate event-detection function. These functions, which we refer to as *virtual sensors*, convert lower-level events to higher-level ones (e.g. alarm buzzing to smoke detected to possible fire), escalating events and publishing them back to DIME. When a possible emergency event is detected, SCALE may alert a resident as described earlier.

We used the above incremental approach as it allows different server components to live on or replicate across separate machines and locations, improving scalability, response times, modularity, reliability, and ease of creating an audit trail. An audit trail exposes intermediate events to external entities, which helps in building trust in particular event sources (i.e. sensing devices, event-detection algorithms) and adding new hooks for separate services to make use of these states. We accomplished this distributed approach using the Celery task queue manager[7] that distributes event-handling across worker processes.

Some historical storage of recent events is necessary to detect changes over time and disambiguate sensor readings indicative of the same event. We used the Django framework's object-relational mapping (ORM) to abstract the PostgreSQL database tables seen in Fig. 4 as Python objects. Periodically, the database removes old events, though in the future we will instead archive them for historical analysis and audit purposes.

Actuation

This subsection describes SCALE's mechanisms for interacting with and alerting human users.

Alerting: Once possible emergency events are detected, concerned individuals must be notified in a timely, reliable, accessible, and interactive manner. Users will receive alert messages after connecting their home monitoring devices to SCALE and registering these devices and contact information with the alerting system. Ideally, this system could eventually integrate with emergency dispatch centers to automatically alert authorities. To mitigate false-positives, it supports a confirmation step in which the user determines whether the emergency is real and emergency personnel should be alerted.

Because SCALE especially aims to make the system as accessible as possible, especially for

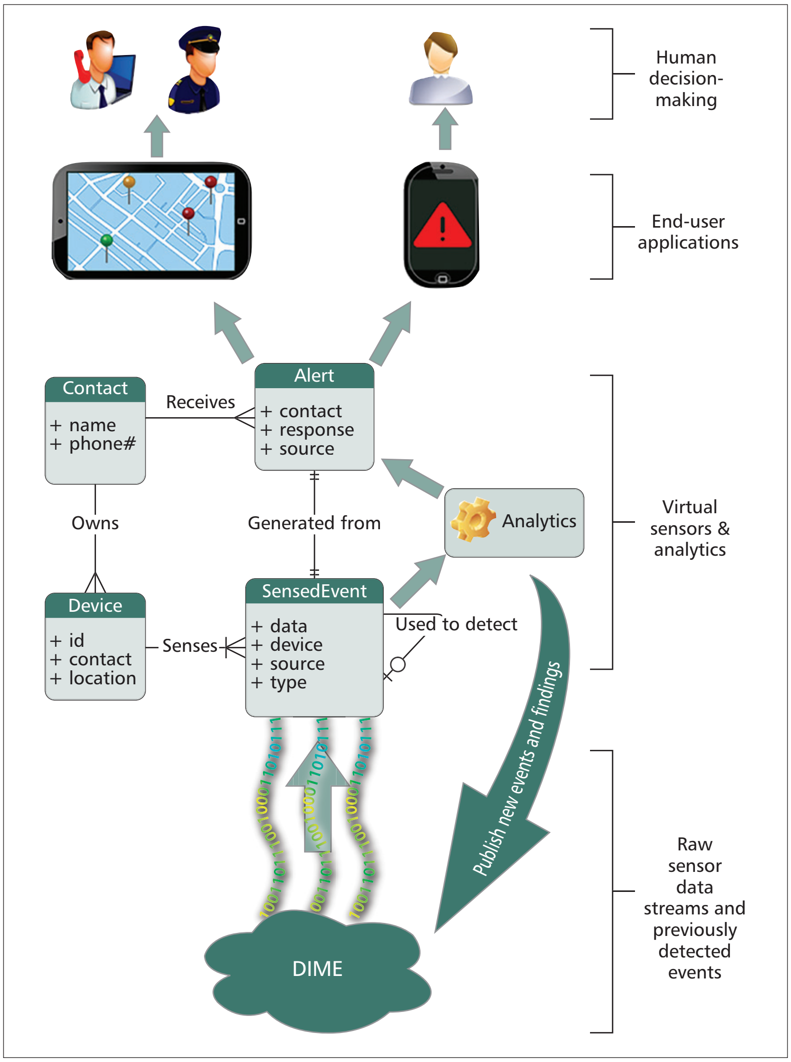
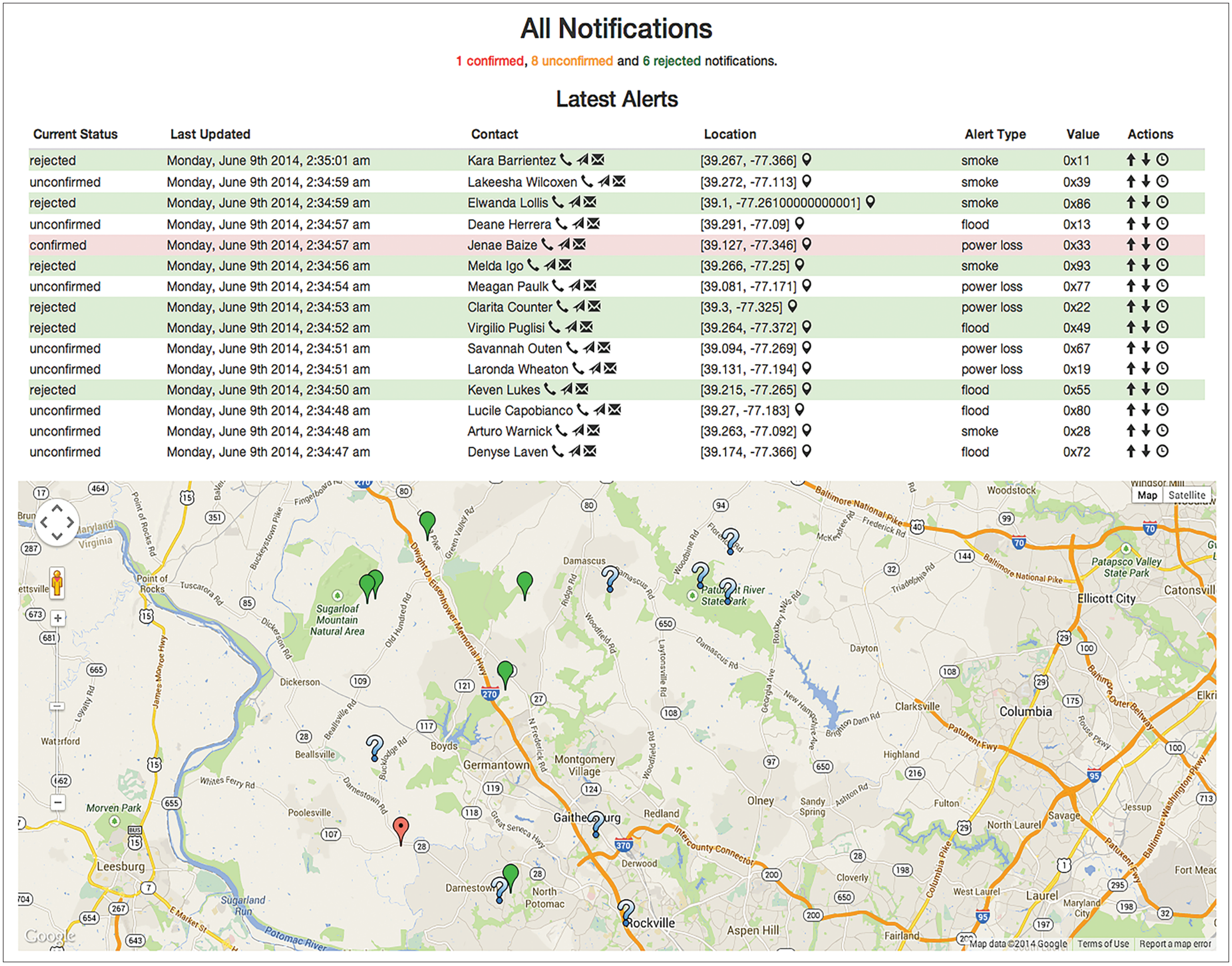


Figure 4. Depiction of data flow and database tables in the analytics and alerting services.

[5] http://www.saltstack.com/

[6] http://www.eclipse.org/paho/

[7] http://www.celeryproject.org



All Notifications

1 confirmed, 8 unconfirmed and 6 rejected notifications.

Latest Alerts

Current Status	Last Updated	Contact	Location	Alert Type	Value	Actions
rejected	Monday, June 9th 2014, 2:35:01 am	Kara Barrientez ☎ ✈ ✉	[39.267, -77.366] 📍	smoke	0x11	↑ ↓ ◷
unconfirmed	Monday, June 9th 2014, 2:34:59 am	Lakeesha Wilcoxen ☎ ✈ ✉	[39.272, -77.113] 📍	smoke	0x39	↑ ↓ ◷
rejected	Monday, June 9th 2014, 2:34:59 am	Elwanda Lollis ☎ ✈ ✉	[39.1, -77.26100000000001] 📍	smoke	0x86	↑ ↓ ◷
unconfirmed	Monday, June 9th 2014, 2:34:57 am	Deane Herrera ☎ ✈ ✉	[39.291, -77.09] 📍	flood	0x13	↑ ↓ ◷
confirmed	Monday, June 9th 2014, 2:34:57 am	Jenae Baize ☎ ✈ ✉	[39.127, -77.346] 📍	power loss	0x33	↑ ↓ ◷
rejected	Monday, June 9th 2014, 2:34:56 am	Melda Igo ☎ ✈ ✉	[39.266, -77.25] 📍	smoke	0x93	↑ ↓ ◷
unconfirmed	Monday, June 9th 2014, 2:34:54 am	Meagan Paulk ☎ ✈ ✉	[39.081, -77.171] 📍	power loss	0x77	↑ ↓ ◷
rejected	Monday, June 9th 2014, 2:34:53 am	Clarita Counter ☎ ✈ ✉	[39.3, -77.325] 📍	power loss	0x22	↑ ↓ ◷
rejected	Monday, June 9th 2014, 2:34:52 am	Virgilio Puglisi ☎ ✈ ✉	[39.264, -77.372] 📍	flood	0x49	↑ ↓ ◷
unconfirmed	Monday, June 9th 2014, 2:34:51 am	Savannah Outen ☎ ✈ ✉	[39.094, -77.269] 📍	power loss	0x67	↑ ↓ ◷
unconfirmed	Monday, June 9th 2014, 2:34:51 am	Laronda Wheaton ☎ ✈ ✉	[39.131, -77.194] 📍	power loss	0x19	↑ ↓ ◷
rejected	Monday, June 9th 2014, 2:34:50 am	Keven Lukes ☎ ✈ ✉	[39.215, -77.265] 📍	flood	0x55	↑ ↓ ◷
unconfirmed	Monday, June 9th 2014, 2:34:48 am	Lucile Capobianco ☎ ✈ ✉	[39.27, -77.183] 📍	flood	0x80	↑ ↓ ◷
unconfirmed	Monday, June 9th 2014, 2:34:48 am	Arturo Warnick ☎ ✈ ✉	[39.263, -77.092] 📍	smoke	0x28	↑ ↓ ◷
unconfirmed	Monday, June 9th 2014, 2:34:47 am	Denyse Laven ☎ ✈ ✉	[39.174, -77.366] 📍	flood	0x72	↑ ↓ ◷

Figure 5. The main view of the SCALE dashboard.

lower-income and/or less technologically-savvy users, it does not require access to a computer or smartphone when receiving and acting on alerts. It supports simple phone calls so that users with land lines, but no cell phones, can still use it. It does support SMS (text messaging) as most people in the U.S. nowadays have cell phones, especially since government programs[8] exist to provide them to low-income residents.

While a smartphone application is a potential future addition, we opted to use an Internet telephone service for alerting. We chose Twilio, which has a rich API for programming interactions with users through the server's web interface, to issue SMS/phone call alert messages and handle correspondence with participants (event confirmation/rejection, registration/unregistration, contact method preferences, etc.).

When the analytics subsystem detects a possible emergency, it sends an Alert message through MQTT to instruct the alerting subsystem to contact the registered user(s) of the device from which the sensor data originated. This contact info is retrieved from the database as shown in Fig. 4, and the database stores an Alert entry representing this communication. When the contact responds, the database updates the state of the Alert to *rejected* if the user responds with "emergency" or presses 1 and *event confirmed* if the user responds with "okay" or presses 2.). If no one responds within some amount of time (currently 30 to 60 seconds) of initiating the alert, a trigger fires that escalates the emergency event. Currently, it is set to confirm the event, but public officials likely would adopt a different policy that perhaps dispatches an individual to investigate further rather than scrambling an entire unit.

Dashboard: To help dispatch personnel visualize alert events and sensed data, we built a dashboard for the SCALE system. We wanted an intuitive, lightweight, web-based solution so we could later port it to mobile devices and borrow functionality for a smartphone application aimed at residents for monitoring their personal SCALE deployment(s). Figure 5 shows the main user interface.

The main view of the SCALE dashboard presents a list of recent alerts, their locations in a Google Maps view, and a summary of the number of high, medium, and low priority events. It includes contact information about the individual alerts and a currently non-functional interface for calling,

[8] http://www.fcc.gov/lifeline

texting, or emailing residents. The user can also confirm or reject events manually. A second view presents raw sensed events as they arrive, which is simply for debugging purposes. The dashboard, also hosted on BlueMix, is built on top of software designed by BioBright. It is written using Node.js at the backend, Javascript and Twitter Bootstrap in the front-end, and a browser MQTT client.

Conclusions and Future Research

Our experience in designing, developing, and deploying the first iteration of SCALE described in this article has proven the feasibility of a distributed IoT approach to improving resident safety at a low incremental cost. This initial exploration requires much further development before this system could be deployed in any real capacity. However, the lessons we learned will help drive future research and development for IoT. We discuss some of these topics below and present our future plans and vision for SCALE.

Resilience Concerns

From a hardware perspective, our biggest lesson learned was that *cheap sensors break*. We purchased many of our devices online for under $10 to $20 (U.S.) and some failed. The explosive gas sensors in particular tended to burn out after a few uses. While some of these issues can be alleviated by using better quality components, this likely drives up the price of the device without completely ensuring reliable operation, and so care must be taken to plan for these issues.

Regarding power, we found that the number of peripherals on the Raspberry Pi required a higher-amperage power adapter. We also used a powered USB hub to support all of the USB sensors and wireless adapters used on the Sheevaplug. We experimented with battery backup and determined that the system dies after about 10 hours. Future work will explore graceful degradation of devices (turning off network adapters, adjusting sampling rates, etc.) to improve this battery life.

From a software perspective, we found the design of the client around simple abstract pipeline components to be very flexible when adding new hardware support.

Currently, we are experimenting with additional networked sensor devices to extend the coverage of a SCALE deployment and improve its resilience. We are adding support for an ad-hoc Wi-Fi mode that supports distributed emergency detection and alerting even during power and network failures by having FlexSCALE boxes exchange data with each other directly. We are also adding inexpensive battery-powered microcontrollers with attached sensors and IEEE 802.15.4-based wireless so additional sensors can be deployed throughout a residence without requiring additional FlexSCALE boxes.

Data Exchange

While we found MQTT suitable for rapidly developing an IoT system, we did find it limited due to its simplistic lightweight approach. Below we outline some considerations for future IoT protocol standardization efforts and security considerations for data management in IoT systems.

Standards Considerations: When designing our analytics system and topic hierarchies, we found MQTT's lack of support for fine-grained queries somewhat limiting. It does not handle ranges at all, and the expressiveness of wildcards cannot match that of regular expressions. For example, to perform a query over a target geography one would need to define a tag for that geography, which limits flexibility for defining new targets. Our current inefficient solution is to subscribe to all events and filter them based on content. Future protocols should consider the desire to issue such queries and filters, as sorting through the results by content on the client-side may be intractable with the larger-scale systems the IoT vision promises.

A major advantage of MQTT is its lightweight nature and simplicity. Future IoT standards should follow this model, while allowing for extensions that provide additional services when, but only when, developers/deployers wish to use them. For example, the size of the DDS[9] standard may intimidate some newcomers, whereas getting started with MQTT takes only a matter of minutes. Protocol designers must keep in mind that many IoT developers will enter the market with little systems experience or come from a Web 2.0 background. As such, providing a simple intuitive starting point, perhaps with RESTful APIs, for them to develop systems will help lower the barrier to entry, resulting in more projects with diverse applications. This approach appears to have worked very well with Node.js, which has enjoyed rapid adoption in part because it gives the developer community the freedom to pick from a variety of options for accomplishing a given task rather than specifying one standard approach. To further lower this barrier, future standards should also allow developers to use familiar tools, languages, etc. whenever possible. For example, they should emphasize interoperability with other protocols, such as how CoAP [3] can interoperate with HTTP.

Security and Privacy: While we did not implement security mechanisms in SCALE beyond requiring SSH keys to remotely access devices, we did discuss security and privacy implications throughout the project and plan to address them in future versions. MQTT supports authentication and identification directly, but not authorization. It can be run using TLS so that the user name and password used for authentication are encrypted during transmission. Identification is handled using a unique identifier or a public digital certificate, with the latter clearly involving management of keys. Some MQTT server implementations provide authorization as an added service. In a scenario where user privacy and integrity of the data and communications is crucial, such a server should be used. This allows the server to determine which client devices have access to which resources, i.e. which topics they are allowed to publish and subscribe to. This would prevent unauthorized individuals from retrieving readings from devices they do not own, as well as prevent publishing of information to a topic representing a different device. However, this does not validate the actual data in question, which could still be faked by an individual with the proper secret keys.

> Because SCALE especially aims to make the system as accessible as possible, especially for lower-income and/or less technologically-savvy users, it does not require access to a computer or smartphone when receiving and acting on alerts. It supports simple phone calls so that users with land lines, but no cell phones, can still use it.

[9] http://portals.omg.org/dds/

One open concern is that of the devices' physical security. As they are located in residents' homes, they could be physically tampered with, moved, or have their code modified by knowledgeable users. This could result in undefined behavior, misleading event reports, or completely spoofed data. This is one of the main reasons for involving human-in-the-loop sensing in order to confirm events before notifying emergency personnel. Whether this step is truly enough to ensure correctness of the data in question is a policy question outside the scope of this article.

One open concern is that of the devices' physical security. As they are located in residents' homes they could be physically tampered with, moved, or have their code modified by knowledgeable users. This could result in undefined behavior, misleading event reports, or completely spoofed data.

Acknowledgments

This work was supported in part by the National Science Foundation award nos. CNS 1143705, CNS 0958520, and CNS 1450768. The authors would also like to thank NIST and the Presidential Innovation Fellows for issuing the SmartAmerica Challenge that initiated our team's creation; Beall's Grant Apartments for hosting our initial test deployment; Kevin Malby and Phu Nguyen from UCI for contributing to SCALE; Mani Chandy from Caltech and Sharad Mehrotra from UCI for their discussions involving sensing and SCALE; World Sensing, Anomaly Systems and the Community Seismic Network at Caltech for providing additional sensors.

References

[1] B. Hore *et al.*, "Design and Implementation of a Middleware for Sentient Spaces," ***2007 IEEE Intelligence and Security Informatics***, May 2007, pp. 137–44.

[2] A. Bourke, J. O'Brien, and G. Lyons, "Evaluation of a Threshold-Based Tri-Axial Accelerometer Fall Detection Algorithm," *Gait & Posture*, vol. 26, no. 2, 2007, pp. 194–99.

[3] Z. Shelby *et al.* RFC 7252 — The Constrained Application Protocol (CoAP).

Biographies

Kyle Benson (kebenson@uci.edu) is a computer science Ph.D. student at UC Irvine and NSF GRFP Honorable Mention. He researches resilient pervasive sensing communications platforms leveraging low-cost Internet-connected devices. During the SmartAmerica Challenge, he led development on SCALE. His current research focus is on the use of geo-aware overlays to improve IoT communications during disaster scenarios, thereby enhancing situational awareness and emergency response efforts.

Charles Fracchia is an IBM Ph.D. Fellow at the MIT Media Lab in Joe Jacobson's Molecular Machines group, and in the Church lab at the Wyss Institute at Harvard Medical School. Charles is a founder of BioBright, a company building a 'smart lab' user interface that can capture and track everything that happens in a biological experiment.

Guoxi Wang is a master's student at the University of California, Irvine, majoring in networked systems. He received his B.S. in information security from Wuhan University in 2012. His research interests include wireless and sensor networks and software-defined networking.

Qiuxi Zhu is a Ph.D. student in the Department of Computer Science at the University of California, Irvine. He received his B.E. degree in automation from Zhejiang University in 2013. His research interests include mobile sensing, mobile networking, and Internet of Things.

Serene Almomen is CEO and co-founder of Senseware, whose cloud-based data platform, wireless modular technology, and patent pending universal sensor interface provides clients with real-time data to optimize facility performance, meet regulatory requirements, and reduce costs. Serene is passionate about using technology to improve the world around us.

John Cohn is an IBM and IEEE Fellow in the IBM Internet of Things Division. He received a BSEE from MIT, and a Ph.D. in CE from Carnegie Mellon. John is eager to share his love of science and technology with anyone who will listen.

Luke D'Arcy is working to build a SIGFOX network covering the entire U.S. Previously he was a founder of Neul, an IoT networking company that was bought last year by Huawei, and of the Weightless SIG, a standards body defining a standard for low power IoT networks. Before that he was one of the first full time marketing staff at the chip company CSR.

Daniel Hoffman is the first chief innovation officer for Montgomery County, Maryland, a position he has held since October 2012. The program he oversees serves as a laboratory for civic improvement and a safe place to test out new processes, technologies, and ideas. Project topics vary from the Internet of Things (IoT) to autism technology to food security and more.

Matthew Makai is a Twilio developer evangelist and software developer with an affinity for Python. He was a speaker at EuroPython, DjangoCon US in 2014, and PyCon in 2015. Matt also writes Full Stack Python, which helps more than 25,000 developers a month learn to build and deploy Python web applications.

Julien Stamatakis is CTO and co-founder of Senseware, whose cloud-based data platform, wireless modular technology, and patent pending universal sensor interface provides clients with real-time data to optimize facility performance, meet regulatory requirements, and reduce costs. Julien is an award winning electrical engineer who designed and developed the mechanics behind Senseware.

Nalini Venkatasubramanian is a professor in the School of Information and Computer Science at the University of California Irvine. She has significant research and industry experience in the areas of distributed systems, adaptive middleware, pervasive and mobile computing, cyberphysical systems, distributed multimedia and formal methods, She has more thn 200 publications in these areas.

Toward Semantic Interoperability in oneM2M Architecture

Prior standards, and also oneM2M, while focusing on achieving interoperability at the communication level, do not achieve interoperability at the semantic level. An expressive ontology for IoT called IoT-O is proposed, making best use of already defined ontologies in specific domains.

Mahdi Ben Alaya, Samir Medjiah, Thierry Monteil, and Khalil Drira

Abstract

The oneM2M standard is a global initiative led jointly by major standards organizations around the world in order to develop a unique architecture for M2M communications. Prior standards, and also oneM2M, while focusing on achieving interoperability at the communication level, do not achieve interoperability at the semantic level. An expressive ontology for IoT called IoT-O is proposed, making best use of already defined ontologies in specific domains such as sensor, observation, service, quantity kind, units, or time. IoT-O also defines some missing concepts relevant for IoT such as thing, node, actuator, and actuation. The extension of the oneM2M standard to support semantic data interoperability based on IoT-O is discussed. Finally, through comprehensive use cases, benefits of the extended standard are demonstrated, ranging from heterogeneous device interoperability to autonomic behavior achieved by automated reasoning.

Introduction

In the past few years, machine-to-machine (M2M) communications have witnessed the emergence of various and different standardization initiatives. Indeed, several applications sectors are pushing standards that are often targeting mainly a specific application domain such as smart meters standards developed by IEC or IEEE (EN 13757, IEEE 1888-2011, etc.). Different standardization defining organizations (SDOs) have tackled this problem by focusing on the definition of a horizontal service platform that fits different verticals. This work was consolidated later on into a global initiative called oneM2M.

It is worth noting that all these initiatives have focused on the communication interoperability between system entities (servers, devices, applications, etc.). Indeed, these standards have defined a horizontal service layer that enables seamless communication between heterogeneous entities independently of the underlying network and vendor-specific device technologies. It is thus possible to reach and deliver a message to any entity in the system. However, no standard has tackled the "meaning" of the message content being exchanged. Although the SmartM2M [1] standard has introduced some recommendations for supporting semantics [3], a generic data model has not been specified. This has been left to the appreciation of the service provider, system developer, or the system user. Such standards have achieved interoperability at the communication level only. It is important to investigate their extension to semantic interoperability. This will lead to efficient systems, where autonomic management could be achieved.

Semantic data is brought through the definition of a common set of ontologies that describe all system entities but also the data items produced, exchanged, and consumed by these entities. Various information models have been defined for the Internet of Things (IoT), ranging from specialized models such as the Zigbee or KNX data models to more general models such as W3C SSN [3]. These solutions suffer from two main issues: they are either too specialized and focus on a specific application domain, or they lack some concepts mainly related to actuators.

In this article we discuss and propose an ontology model called IoT-O that handles both sensing and actuating concepts of M2M devices, as well as concepts related to services. We then discuss the extension of the oneM2M standard to support semantic data based on the proposed ontology. Finally, through comprehensive use cases, we show the use of IoT-O following the oneM2M standard.

The oneM2M Standard

The oneM2M global initiative [4] is an international partnership project established in June 2009 by the seven most important SDOs in the world and various alliances and industries. The main goal is to define a globally agreed M2M service platform by consolidating currently isolated M2M service layer standards activities. The oneM2M standard is organized into five technical working groups focusing on M2M requirements, system architecture, protocols, security, and management, abstraction and semantics.

As described in Fig. 1, the oneM2M system architecture is composed of the following four functional entities: the application dedicated node (ADN); the application service node (ASN); the middle node (MN); and the infrastructure node (IN). Each node contains a common services entity (CSE), an application entity (AE), or both. An AE provides application logic, such as remote power monitoring, for end-to-end M2M solutions. A CSE comprises a set of service functions called common services functions (CSFs) that can be used by applications and other CSEs. CSFs includes registration, security, application, service, data and device management, etc.

The oneM2M standard adopted a RESTful architecture, thus all services are represented as resources to provide the defined functions. In the rest of the article we will focus on oneM2M, but the proposed solution also works for SmartM2M since the two standards share a similar architecture.

Mahdi Ben Alaya, Samir Medjiah, and Thierry Monteil are with CNRS and Univ de Toulouse.

Khalil Drira is with CNRS.

0163-6804/15/$25.00 © 2015 IEEE

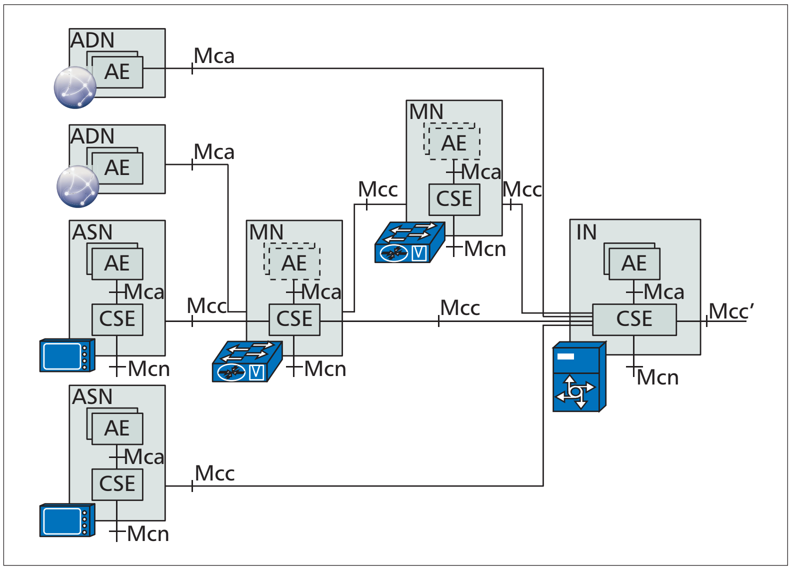


Figure 1. oneM2M system architecture [4].

Full Interoperability Challenge

Full interoperability is a desirable property to achieve in M2M systems, and will pave the way to the ultimate goal, i.e. autonomic management. Indeed, interoperability between heterogeneous devices and services is required to achieve self-configuration, self-healing, self-optimization, and self-repairing.

As introduced earlier, almost all standardization initiatives have not efficiently tackled the issue of full interoperability, i.e. considering both communication and data interoperability. Thanks to communication interoperability, M2M system entities already benefit from services such as discovery, monitoring, management, etc. Although such a service platform can be sufficient for the design and implementation of specific M2M systems, autonomic management using automated reasoning based on a knowledge oriented service platform cannot be achieved.

For example, using a service platform built upon oneM2M, an application can seamlessly discover new devices plugged into the system. This application can subscribe to the new device events, and will receive them as soon as they are triggered, even if the routing path implies crossing multiple entities and using heterogeneous communication protocols or network technologies at any segment of the communication path. This has been made possible thanks to the interoperability at the communication level. Now that device events have been successfully reported, the application does not have any means to understand the reports' content without prior conventions (data formats, encapsulation, and semantics) set up between the application and the device application developers.

IoT Ontology Principles

Ontologies have proven to be beneficial for intelligent information integration, information retrieval, and knowledge management. They enable the indexing of resources' content using semantic annotations that can result in the representation of explicit knowledge that cannot be assessed and managed because of their mess. Ontologies are very popular and useful to overcome challenges fixed in the proposed study because they provide an efficient way of cleverly structuring a domain, making use of semantic hierarchical and property/value relationships based on a vocabulary of concepts/instances [5].

In an M2M system, users and applications should be able to discover, monitor, and control sensors and actuators offering particular services and having particular properties with a high degree of automation. To reach this goal, an ontology for IoT shall represent a variety of concepts such as platform, deployment system, thing, device, node, service, sensor, actuator, sensing and actuating capabilities, observation, operation, time, unit, kind, and their relationships.

Since ontologies are designed to be reusable and extensible, it is possible to define a complete ontology for IoT by reusing existing ontologies. New concepts should be designed only when needed. This approach helps reduce the ambiguity of IoT terminology and makes it possible to converge quickly to a common vocabulary.

IoT-O: An Ontology for IoT

Since there is no single model that covers all IoT concepts, a set of well-defined ontologies were carefully selected to create an efficient ontology for IoT called IoT-O.

IoT-O Concepts and Relationships

IoT-O consists of five main parts: sensor, observation, actuator, actuation, and service models. Figure 2 shows how the selected ontologies are merged together to form this new ontology.

The DUL upper ontology is selected to describe very general concepts that are the same across all knowledge domains, and so facilitate reuse and interoperability. It is a lightweight foundational model for representing either physical or social contexts. The SSN ontology, which is aligned with DUL, is selected to represent sensors in terms of measurement capabilities and properties, observations, and other related concepts. However, it does not describe actuator devices.

Since currently there is no ontology that accurately describes actuators, we designed a new ontology called SAN, which is inspired by SSN and aligned with DUL, to describe actuators in terms of actuating capabilities and properties, actuation, and related concepts. The QUDV ontology was selected to represent quantities, units, dimensions, and values. The OWL-TIME ontology was selected to provide a vocabulary for expressing facts about topological relations among instants and intervals, together with information about duration, and about date time information.

Given that oneM2M aims to enable seamless interactions between business applications and services, it is important to represent how these services can be requested without any ambiguity in order to reduce the amount of manual effort required for discovering and using them. The MSM ontology was selected to describe services since it provides a common vocabulary based on existing web standards able to capture the core

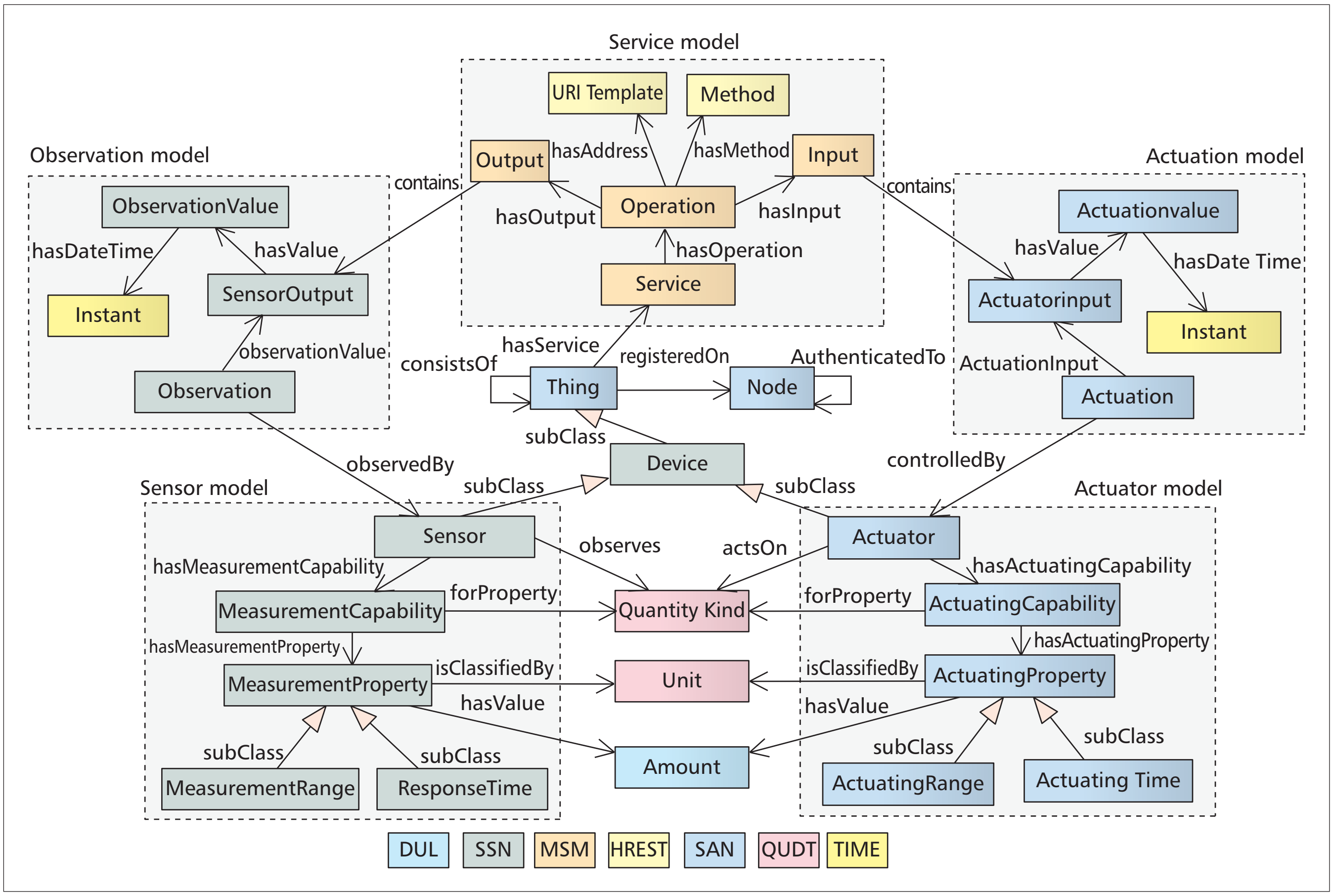


Figure 2. IoT-O ontology model.

semantics of both Web services and Web APIs in a common model. Each service is described using a number of operations that have address, method, input, and output message contents.

Actuator Model Instance According to IoT-O

Let's consider a concrete example representing a real actuator using the IoT-O ontology. The "HUELUX" actuator is a digitally dimmable wireless lighting bulb from Philips having power ranging from 0 Watt to 50 Watt. The luminosity level can be dimmed by requesting the required power value. The light bulb offers a web service to enable remote luminosity control. The luminosity can be dimmed instantaneously by sending a create request to the address "/HUELUX_APP/dimming" with a message body containing the required power. Figure 3 illustrates the corresponding ontology instance. It shows how the actuator, actuation, and service information are inserted into the IoT-O ontology. The actuator model represents the light bulb information and actuating capabilities, including power range. The actuation model represents the dimming command. The service model represents the light bulb web service, including the luminosity dimming operation, address, and method.

Semantics Extension to the oneM2M Standard

Through oneM2M working group 5, semantics is already envisioned for the oneM2M standard. However, as of its first release, semantics aspects are not yet tackled. In this subsection we will discuss a possible extension to oneM2M in order to support semantic interoperability. Two options are available.

The first option, which we called *reference-based integration*, uses the ontology reference attribute *ontologyRef* of *type xs:anyURI*, already defined in the oneM2M standard to include the semantic meta data. This attribute contains a unique reference of an ontology concept that describes the semantic meaning of a specific resource. It can be used to enhance the discovery mechanism to find a specific resource based on a semantic concept. This option is very limited as it informs only about a concept, and does not provide any semantic meta-data about the concept relationships, which is important for semantic interoperability.

The second option, which we called *inline integration*, consists of defining a new sub-resource called a "descriptor" to the oneM2M resource architecture. The "descriptor" sub-resource contains a complete semantic description of the resource defined using the resource description framework (RDF). This semantic description is exposed and shared across different applications, thus enabling semantic interoperability. This option requires the use of a triple store within a CSE to store the ontology instance.

We do not have to chose between the two options as they are complementary and can coexist in the same architecture. Such a solu-

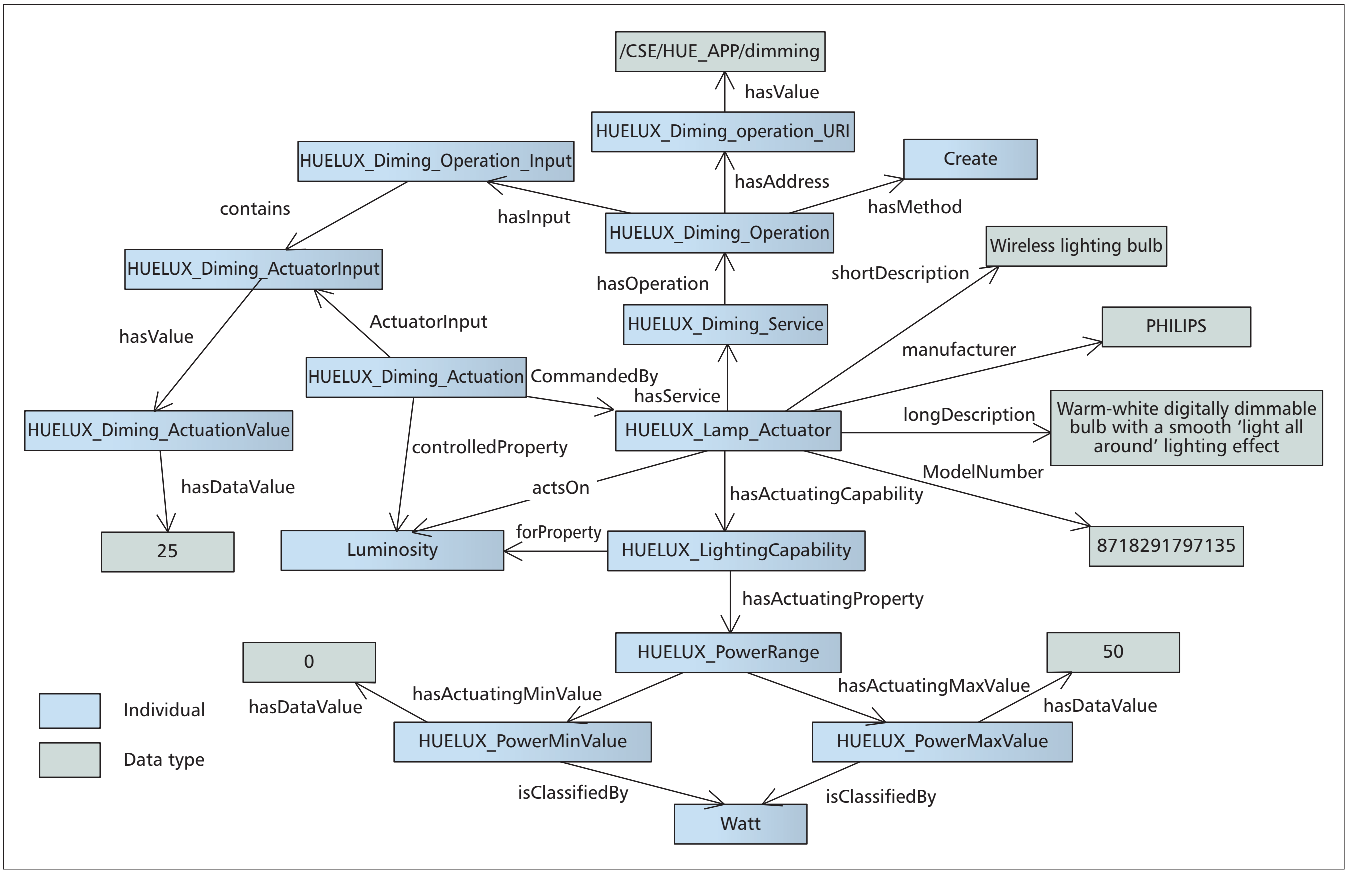


Figure 3. Actuator model instance according to IoT-O.

tion offers more flexibility to applications. In fact, depending on the scenario, an application may decide to read only the ontology reference attribute or go further and retrieve the full semantic description from the descriptor resource.

Semantic Data Interoperability

In this section, and through comprehensive use cases, we present the usability of the IoT-O ontology. The use case will also feature our living lab, i.e. the ADREAM smart building, and tackle the addressed challenge in a real scenario. Our OM2M [6] platform, which offers an open source implementation of the oneM2M standard, is used to validate the proposed approach. The OM2M high level architecture is described in Fig. 4.

LAAS Smart Building: ADREAM

ADREAM is the LAAS-CNRS smart experimental building. The main originality of this instrumented building compared to already existing buildings is that it is a "living lab" of 1700m^2, as it is both a research tool and a building with offices for the researchers.

The building includes 500m^2 of technical platforms (IoT, robotics, ambient intelligence, and energy) and 700m^2 of offices, with the remaining space devoted to a garden. It hosts our smart apartment equipped with various sensors and actuators connected using different networking technologies. The device set includes different sensors (temperature, humidity, luminescence, presence, etc.), as well as actuators such as electric plugs attached to different elements, e.g. lamps, fans, humidifier, etc., with all these devices gathered around different gateways. Each gateway is specialized in one or two networking technologies. Finally, these gateways are connected to one central server.

Seamless Device Discovery and Interaction

In order to demonstrate the interoperability achieved by the OM2M platform through its compliance with the oneM2M standards and its support of a generic data model IoT-O, we propose a simple scenario where the software platform is able to discover newly plugged devices such as sensors and actuators, browse the exposed attributes and methods, and finally interact with these devices by retrieving sensed data or triggering actions.

The scenario setup includes different devices attached to an M2M gateway through local network technologies that are either wireless, e.g. ZigBee and 6lowpan, or wired, e.g. KNX. The M2M gateway is connected to an M2M server. The M2M gateway entity is equipped with mapping modules that translate every communication with a specific networking technology into a generic communication protocol that is completely independent from the transport protocol or the network access. Thus, the support of new technologies or protocols is simply achieved through the implementation of the translation module, i.e. the interworking proxy unit (IPU).

When the IPU discovers a new device through the specific technology discovery mechanism, it will expose this device along with its attributes and methods to other entities in the M2M system. From the M2M system perspective, any data or action request is routed to this IPU in order to be translated to the specific technology operations. In this way, any application present in the M2M system can access the newly discovered resources (device, device's attributes, device's actions, etc.) using standardized restful operations. This can be achieved without any knowledge of the underlying network technology or its low-level mechanisms.

Furthermore, as all exchanged messages are augmented with semantics, an application not only has access to the data being generated by the device or the actions it exposes, but it can also understand the meaning of the data. Indeed, as the application has access to the ontology that defines the data, it can browse the ontology and map the received data elements into this ontology and perform the appropriate processing.

Toward Autonomic M2M Systems

In this section we demonstrate how the IoT-O ontology can be used to develop autonomic M2M systems [7, 8] capable of self-management to hide the intrinsic complexity from administrators and users. The main goal here is to use IoT-O ontology to dynamically reconfigure CSE resources to interconnect applications according to their semantic description.

Autonomic Service for Self-Configuring M2M Resources

In general, an application must manually perform a set of complex requests to discover relevant resources and perform several subscriptions to monitor the evolution of interesting resources. In addition, as an application has only a partial view of the M2M environment, it becomes difficult to find relevant services, especially in huge and highly distributed M2M environments.

An autonomic computing service is integrated into oneM2M as a new CSF is integrated into a CSE. A CSE control loop enables the dynamic reconfiguration of the oneM2M resource architecture when needed. The objective here is to help applications discover relevant devices and exchange data with the correct communication mode based on its description, role, and relationships.

To meet this goal a representative model of M2M system knowledge is required to assist the execution of the management process. Since the IoT-O ontology covers all required M2M concepts needed for this use case, it will be used as a knowledge model by the autonomic service.

ADREAM Smart Building Use Case

The ADREAM building contains two middle nodes, "MN-CSE-1" and "MN-CSE-2", registered to an infrastructure node "IN-CSE". A luminosity sensor is connected to the "MN-CSE-1", where it created the "LumSensorApp" application. A lamp actuator is connected locally to "MN-CSE-2", where it created the "LampActuatorApp" application.

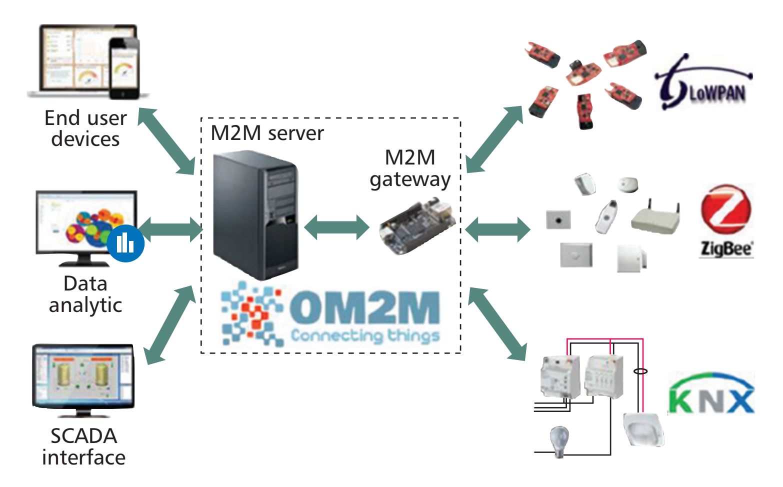


Figure 4. OM2M high level architecture.

The autonomic service discovers and monitors all registered applications, reasons on the IoT-O ontology model using inference rules to find relevant matching, and finally reconfigures the resource architecture accordingly to set up the required connections.

Monitoring the Smart Building System Entities

The autonomic service is responsible for detecting all existing nodes within the M2M system and reflecting the corresponding configuration into its knowledge base. It adds the infrastructure node as an individual of the "IN-CSE-1" node into the IoT-O instance. It then adds the two discovered middle nodes as individuals of the "MN-CSE" class. For each discovered node, the autonomic service retrieves the registered applications with their corresponding services and operations, and adds them to the ontology instance. Figure 5 shows the IoT-O instance as generated by the autonomic service.

The "LuminositySensor" individual is added as an instance of the "Sensor" class. It is registered to the "MN-CSE-1" node and is linked to the "Luminosity" quantity kind with the "observes" relationship. This sensor provides a service called "LuminosityService" that offers several operations to measure the luminosity level. In particular, the operation "RetrieveLuminosityOperation" informs about the address, method, and output required for retrieving data from to the sensor.

The "LampActuator" individual is added as an instance of the "Actuator" class. It is registered to the "MN-CSE-2" node and is linked to the "Luminosity" quantity Kind using the "actsOn" relationship. This actuator provides a service called "LampService" that offers several operations to configure the luminosity level. In this example the lamp is capable of adjusting its intensity according to ambient luminosity received as input. The operation "LampOperation" informs about the address, method, and input required to send to update the actuator state.

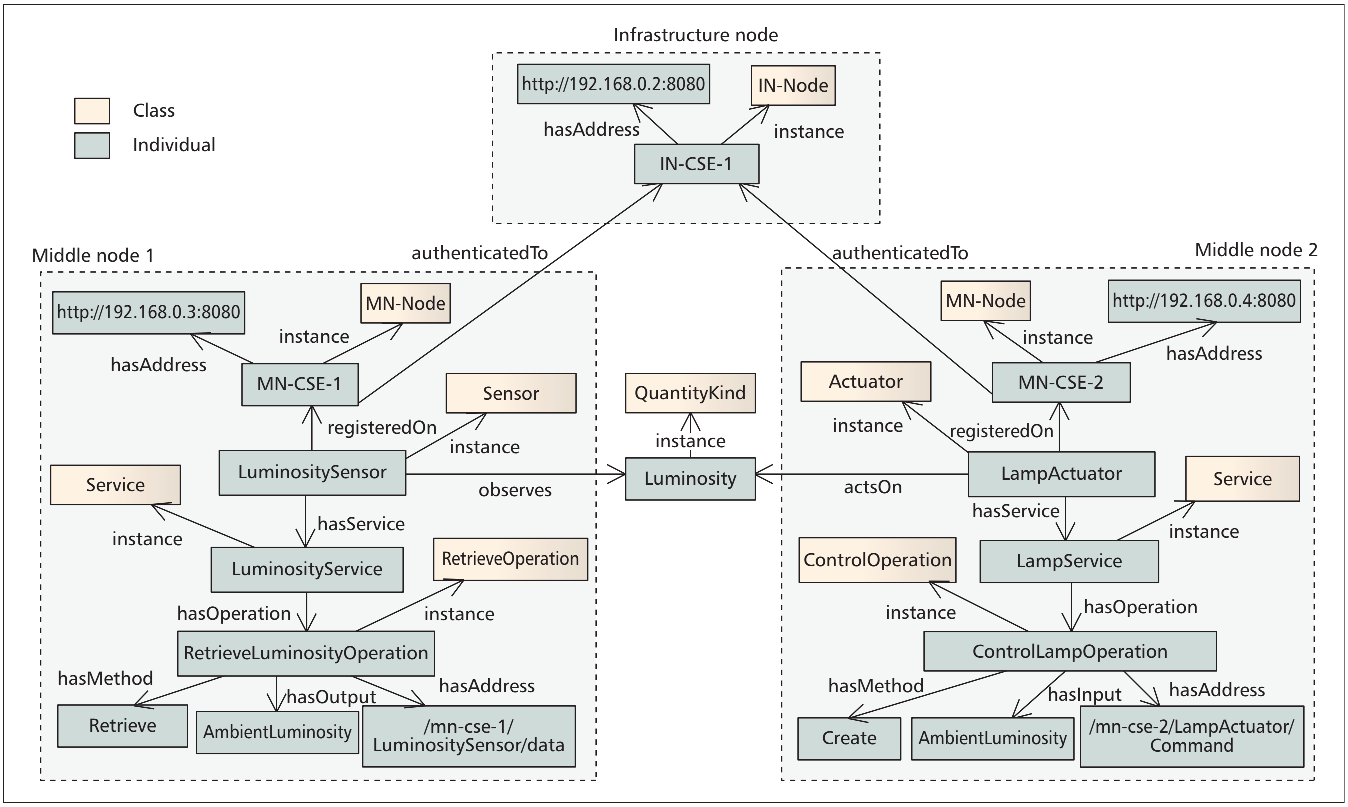


Figure 5. IoT-O smart building use case instance example.

Analyzing Applications Semantic Matching Using Inference Rules

Inference rules can be applied to infer new knowledge and so enrich the IoT-O instance with new individuals and relationships. This new knowledge is necessary to understand the role of each application in the M2M architecture. It allows each application to take maximum advantage of the services offered by other applications.

In this example the following SPARQL [9] rule is applied by the analyzer to find semantic matching between available things. It makes it possible to connect sensors with relevant actuators within the M2M system. For example, if there is a sensor that observes a particular quantity kind, and there also exists an actuator that acts on the same quantity kind, and if this sensor is not already matched to that actuator, then as a result a new "matches" relationship is inferred to link these two devices together. In our example, one semantic matching is detected: the luminosity sensor is matched with the lamp actuator:

```
CONSTRUCT {
        ?sensor iot-a:matches ?actuator
}
WHERE {
        ?actuator rdf:type san:Actuator.
        ?sensor rdf:type ssn:Sensor.
        ?quantityKind rdf:type qu:QuantityKind.
        ?actuator san:actsOn ?quantityKind.
        ?sensor ssn:observes ?quantityKind.
        FILTER EXISTS {
                ?sensor iot-a:matches ?actuator
        }
}
```

Using the same approach, more advanced rules can be applied including more constraints such as the device location, temporal requirements, and quality of services parameters.

Planning of Resource Reconfiguration Actions

The autonomic service plans the list of required actions to establish the communication between the matched things. Concretely, for each matching, a specific subscription action must be created with the correct parameters such as method, input, and also source and subscriber addresses. The autonomic service can extract these parameters from the IoT-O instance by processing the list of operations provided by each thing.

To configure the detected matching, the autonomic service extracts, from the operation "RetrieveLuminosityOperation", the method, input, and the source address where the subscription should be applied "/mn-cse-2/LuminositySensor/data". The operation "ControlLampOperation" makes it possible to extract the address of the subscriber (/mn-cse-1/LampActuator/Command). As a result, the autonomic service creates a subscription action to subscribe the lamp actuator to the luminosity sensor.

Establish the Communication between Matched Applications

Finally, the autonomic service converts each planned action to a Restful request, including all required parameters, using a specific communication protocol such as HTTP or CoAP. Then it sends each request to the IN-CSE. If a subscription request is executed successfully, the autonomic service adds the "manages" relationship between the corresponding actuator and sensor; otherwise, an alert including the error details is reported. As soon as all requests are executed and reported, a new control loop can start.

The lamp actuator is dynamically subscribed to the luminosity sensor. It receives luminosity notifications and updates its level accordingly. Using the same approach, the current lamp actuator subscription can be cancelled and new subscriptions can be dynamically created to connect to more adapted sensors according to changes in the M2M environment.

Conclusion

Current M2M standards aim to provide a horizontal service platform to enable communication interoperability between machines. However, semantic data interoperability is not achieved, which brings into question the horizontality of such a platform. To overcome this challenge, a dedicated ontology for IoT, called IoT-O, has been defined. IoT-O merges together a set of popular ontologies and is enriched with new relevant concepts and relationships. Two possible integrations with the oneM2M standard are discussed. The main concepts and relationships of IoT-O are described using different use cases. An autonomic service making use of IoT-O and inference rules for resource architecture dynamic reconfiguration was also explained.

In future work, we propose to calculate the overhead cost of the proposed solution and the resulting overload. We also propose to validate IoT-O in various vertical M2M domains such as e-health, transport, and smart grid. The use of semantics makes it possible for end users to perform advanced discovery requests based on a SPARQL endpoint. However, the current oneM2M security solutions are not sufficient to support such a capability, the authorization mechanism must be rethought. Terminology for cloud data management and service lifecycle management can also be used to extend IoT-O concepts. A set of contributions will be sent to oneM2M for integrating IoT-O concepts into the standard to move forward toward semantic data interoperability.

References

[1] D. Boswarthick, O. Elloumi, and O. Hersent, *M2M Communications: A Systems Approach,* John Wiley & Sons, 2012.

[2] ETSI, "TR 101 584–Study on Semantic support for M2M Data," ETSI M2M, 2013, v2.1.1, pp. 1–34.

[3] M. Compton *et al.*, "The SSN Ontology of the W3C Semantic Sensor Network Incubator Group," *Web Semantics: Science, Services and Agents on the World Wide Web*, 2012, vol. 17, pp. 25–32.

[4] J. Swetina *et al.*, "Toward a Standardized Common m2m Service Layer Platform: Introduction to onem2m," *IEEE Wireless Commun.*, 2014, vol. 21, no 3, pp. 20–26.

[5] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, 2001, vol. 284, no 5, p. 28–37.

[6] M. Ben Alaya *et al.*, "OM2M: Extensible ETSI-Compliant M2M Service Platform with Self-Configuration Capability," *Procedia Computer Science*, 2014, vol. 32, pp. 1079–86.

[7] J. O. Kephart and D. M. Chess. "The Vision of Autonomic Computing," *Computer*, 2003, vol. 36, no 1, pp. 41–50.

[8] M. Ben Alaya and T. Monteil. "Frameself: An Ontology-Based Framework for the Self-Management of M2M Systems," *Concurr. Comput. Pract. Exp*, 2006, vol. 18. no 6, pp. 1412–26.

[9] J. Perez, M. Arenas, and C. Gutierrez. "Semantics and Complexity of SPARQL," *ACM Trans. Database Systems (TODS)*, 2009, vol. 34, no. 3, pp. 1–45.

Biographies

Mahdi Ben Alaya (ben.alaya@laas.fr) is a computer science engineer and IoT solutions architect. He obtained a Ph.D. in networks, telecommunications, systems, and architecture from the University of Toulouse, France. He is vice-chairman of the oneM2M Testing Group. He is co-founder and technical manager of the open source project OM2M at the Eclipse Foundation. His research interests include M2M interoperability, autonomic computing, IoT semantic, and information centric networking.

Samir Medjiah (medjiah@laas.fr) received a Ph.D. (2012) in computer science from the University of Bordeaux, France. He is an associate professor at Paul Sabatier University in Toulouse (France) and a research scientist at the Laboratory for Analysis and Architecture of Systems (LAAS-CNRS). His main research interests include application-network co-optimization, software-defined networking, network virtualization, M2M communications, and IoT applications. He is actively working on R&D projects related to M2M/IoT.

Thierry Monteil (monteil@laas.fr) has been an associate professor in computer science since 1998 at INSA Toulouse and a researcher at LAAS-CNRS. He works on parallel computing, cloud resource management, autonomic middleware, and machine-to-machine and Internet of Things architecture. He is member of ETSI (European Telecommunication Standards Institute). He also represents CNRS in the Eclipse Foundation and co-leads the OM2M open source project. He has authored more than 50 regular and invited papers in conferences and journals.

Khalil Drira (drira@laas.fr) obtained the Ph.D. and HDR degrees in computer science from UPS, University Paul Sabatier Toulouse, in October 1992 and January 2005, respectively. From October 1992 to September 2010 he was Chargé de Recherche, and since October 2010 he has been Directeur de Recherche, a full-time research position, at the French National Center for Scientific Research (CNRS). His research interests include formal design, implementation, testing and provisioning of distributed communicating systems and cooperative networked services.

The authorization mechanism must be rethought. Terminology for cloud data management and service lifecycle management can also be used to extend IoT-O concepts. A set of contributions will be sent to oneM2M for integrating IoT-O concepts into the standard to move forward toward semantic data interoperability..

Toward Enhanced Data Exchange Capabilities for the oneM2M Service Platform

The Machine-to-Machine Service Platform is being standardized to enable the intercommunication of devices, which is the basis for smart environments and intelligent transport systems applications. In such environments, adapting the data exchange between devices and applications to the requirements of the application is a critical step in ensuring the functionality and reliability of the service.

Markus Glaab, Woldemar Fuhrmann, Joachim Wietzke, and Bogdan Ghita

Abstract

The machine-to-machine (M2M) service platform is being standardized to enable the intercommunication of devices, which is the basis for smart environments and intelligent transport systems (ITS) applications. In such environments, adapting the data exchange between devices and applications to the requirements of the application is a critical step in ensuring the functionality and reliability of the service. This article employs test-cases to analyze the data exchange of the oneM2M standard using an M2M-based automotive service delivery platform. Following the analysis, it proposes enhancements such as application-data-dependent criteria for data notification in combination with aggregation of different subscriptions to the same resource. Finally, the article discusses the proposed enhancements against the background of M2M design considerations and improved privacy.

Introduction

The machine-to-machine (M2M) service platform is being developed with the aim of overcoming existing vertical silo solutions and of building a standardized horizontal integration platform for manifold machines (also known as devices or things) and domains. This facilitates new use cases and concepts that are often referred to as 'smart' or 'intelligent', such as smart home, smart grid, smart cities, and intelligent vehicles as part of an intelligent transportation system (ITS).

The European Telecommunication Standards Institute (ETSI) M2M service architecture [1] was the first step toward a universal M2M platform, which already provided a good maturity level on unified communication capabilities that are abstracted from specific technologies and protocols. Currently, the oneM2M Global Initiative[1] works on a harmonized reference architecture, which integrates previous work such as the aforementioned by ETSI, as well as from other standardizing organizations, for example, the Open Mobile Alliance (OMA) and the Broad Band Forum (BBF). OneM2M released the first version of their service platform specification in April 2015 [2]. Future releases are expected to feature full semantic interoperability, which is necessary to facilitate inter-vendor and inter-domain applications without additional a-priori agreements [3].

From the software engineering perspective, M2M use cases consist of distributed inter-connected applications on top of the oneM2M service platform. This means, prior to data processing within an application on a device or server, input data must be acquired from other components, such as measurements from a large number of sensors, or preprocessed content from other machines. Therefore, having adequate data exchange capabilities is a key factor for the oneM2M service platform because it affects the service quality during operation. In particular for devices connected via a wireless access network, the constraints of the respective transport networks have to be reflected. Herein, the ability to tailor data acquisition to the requirements of a particular use case directly impacts resulting bandwidth requirements, which may conflict with the capabilities of the transport network. This is a deciding factor as to whether a particular functional split between applications and devices is feasible, including the respective costs.

This article focuses on the data exchange capabilities of the current oneM2M specification and introduces a series of enhancements to support application-data-dependent notification criteria including local aggregation. This can enable significant bandwidth saving for many distributed usage scenarios. The considerations arose in the context of research on the applicability of M2M as the enabler for future distributed automotive software platforms [4]. The article starts by introducing the developed architecture of an M2M-based automotive service delivery platform (ASDP) and two example use cases. We present the analysis of current data exchange capabilities and propose enhancements derived from a typical automotive scenario. We detail a number of enhancements and their implementation, and evaluate them against the introduced automotive scenario. We discuss the approach and limitations within the wider context of oneM2M standardization, and discuss future work. Finally, we summarize the contributions of this study.

An M2M-Based Automotive Service Delivery Platform: Background, Architecture, and Use Cases

This section provides the foundation for an M2M-based ASDP. It starts with a short introduction to the automotive software domain. Afterward, the architecture and two use case examples are described.

Background

The automotive domain is changing. More than 80 percent of vehicular innovations are related to electronics and software [5]. This is driven by

Markus Glaab is with University of Applied Sciences Darmstadt and Plymouth University.

Woldemar Fuhrmann is with University of Applied Sciences Darmstadt.

Joachim Wietzke is with Karlsruhe University of Applied Sciences.

Bogdan Ghita is with Plymouth University.

[1] http://www.onem2m.org

several factors. In-vehicle infotainment (IVI) systems must compete with developments in the area of consumer electronics (CE) products, in particular smartphones and tablets. Besides, advanced driver assistance systems (ADAS) aim to continuously increase traffic safety toward zero accidents through assistance and (semi-) autonomous driving capabilities. Finally, within the superior vision of ITS, vehicles are becoming an integral part of our connected world, aiming at further increases in traffic efficiency, safety, and comfort of its users [6].

Hence, the expectation is that future vehicles will be connected to the Internet and to neighboring peer vehicles and infrastructure. In this regard, the automotive domain is a prime example of the concept of an Internet of Things (IoT) and its inherent challenges. Car manufacturers or suppliers, in the following referred to as original equipment manufacturers (OEMs), are now facing the task of integrating applications and services from several platforms and domains. These applications differ in many ways, such as innovation and lifecycles, performance and real-time requirements, and criticality [7]. However, the OEMs have to integrate them into a homogeneous overall system that remains functional over the complete lifetime of the car [8]. This requires new automotive software architectures that are able to handle the heterogeneity of the future vehicular application landscape [5, 9].

The expectation is that future vehicles will be connected to the Internet and to neighboring peer vehicles and infrastructure. In this regard, the automotive domain is a prime example for the concept of an Internet of Things (IoT) and its inherent challenges.

Architecture

Connected vehicles raise new challenges for current software development, but they also enable new ways to address them. One promising approach for the software architecture of the next generation of automotive applications is an ASDP [4], aiming at the increased utilization of connectivity together with server infrastructure, thereby shifting the hub for the integration of automotive applications from the vehicle to a related OEM server. The offloading of existing automotive applications and the cloud-based implementation and integration of new OEM and third party functionality typically face less computational constraints than the traditional approach of integration on embedded automotive systems. Further, an intermediary OEM server between the vehicle and third party applications and domains increases mediation capabilities. Thus, the new approach is advantageous for many applications, in particular those that require connectivity [10].

Some manufacturers have already started to build proprietary cloud solutions. However, in our ASDP approach, we envisage the alternative of an open and standardized architecture, not limited to one vendor or the automotive domain. Further, a more extensive network integration of vehicles is intended, toward an "embedded Internet" [11]. In this regard, current developments within M2M architectures suit well the approach of an ASDP, and M2M has been selected as the underlying platform [10].

Following the concepts of the oneM2M service platform [2], a vehicle could be an application service node (ASN) or a middle node (MN) that integrates all vehicle-internal ASNs or non-oneM2M device nodes (NoDN). Within our ASDP, it has been decided that the vehicle is a oneM2M-compliant ASN, and the OEM server is an infrastructure node (IN). The ASN is located inside the field domain and is connected to the infrastructure domain, using wireless access networks.

The ASN and IN are divided into the application layer and the common services entity (CSE), with the goal to encapsulate essential functions for M2M application entities (AE). This reflects the objective of a universal, horizontal integration platform.

Figure 1 illustrates the compound functional architecture of an M2M-based ASDP with the reference points *Mca* (vertical interface for AE), and *Mcc* (horizontal between two CSE*s*). oneM2M-compliant third party servers are connected by use of the *Mcc'* interface with the OEM server, while other third party platforms could be connected through adaptor-AEs. For the sake of completeness, the possibility of a direct connection between the vehicle and other M2M-conformant INs, e.g. third party servers, via the *Mcc* interface is hinted. Against the background of mediation capabilities between a vehicle and third party server, the ASDP concept describes the value of an interposed OEM server. Hence, the direct connection to third party servers is currently not favored.

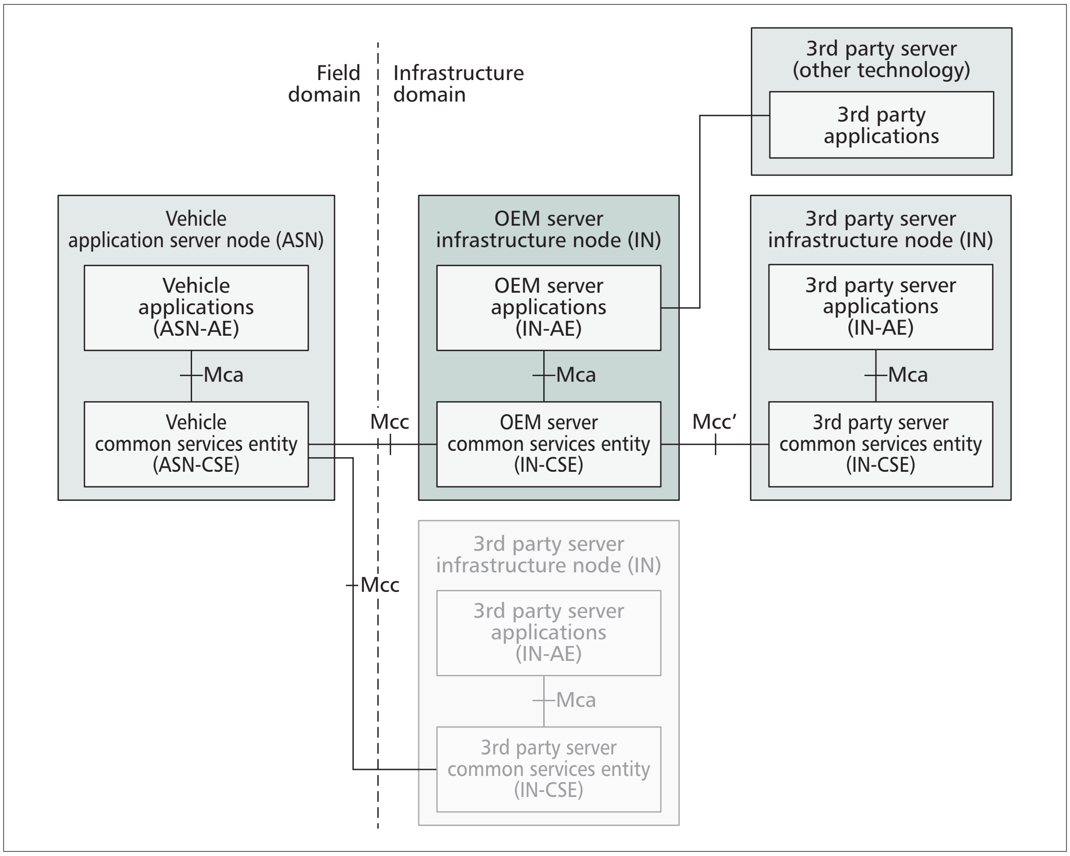
Use Cases

The ASDP concept with an OEM server as the hub for application integration facilitates many use cases, currently associated with connected vehicles. To evaluate the data exchange capabilities of oneM2M, this section introduces two basic use cases that are widely accepted.

Extended Floating Car Data: With extended floating car data (XFCD) [12], vehicles are used as driving traffic information sensors. They periodically report at a minimum their current location, together with the timestamp, to the OEM server. The trigger for these reports may be time-related, distance-related, or a combination of both. The OEM server, respectively a third party traffic management center, aggregates and analyzes the data and traffic models and can then detect traffic jams and calculate average trip times. This information can be used by navigation systems that are able to consider the current traffic situation. XFCD includes the transmission of additional sensor data to the OEM server, if it detects critical situations through the vehicular sensors. Triggers may vary from outside temperature or rain intensity to driving dynamic control interventions, such as an electronic stability control (ESP) intervention. The provision of these measurements, together with the position and speed, enables advanced inference regarding the momentary traffic safety conditions on a certain road section.

Vehicle Maintenance/Fleet Management: Modern vehicles have variable service intervals, depending on their usage, which is monitored over time to estimate when thresholds are exceeded and service has become necessary. Additionally, various vehicular sensors and check routines continuously monitor component status and individual component failures. These are currently only locally stored using a fault record-

By design-choice of the current oneM2M releases, the application data is opaque at CSE level. Hence, application-data-dependent notification criteria that are derived from the real use case requirements, tailoring the transmitted data to the actual needs, cannot be applied.



Field domain
Infrastructure domain
3rd party server (other technology)
3rd party applications
Vehicle application server node (ASN)
Vehicle applications (ASN-AE)
Mca
Vehicle common services entity (ASN-CSE)
OEM server infrastructure node (IN)
OEM server applications (IN-AE)
Mca
OEM server common services entity (IN-CSE)
3rd party server infrastructure node (IN)
3rd party server applications (IN-AE)
Mca
3rd party server common services entity (IN-CSE)
Mcc
Mcc'
Mcc
3rd party server infrastructure node (IN)
3rd party server applications (IN-AE)
Mca
3rd party server common services entity (IN-CSE)

Figure 1. Functional architecture of an M2M-based automotive service delivery platform.

er and manual readout at the garage. Connected vehicles enable use cases, where relevant data can be submitted to the OEM server periodically, or upon error occurrence. The gathered data may be subsequently used to initiate a separate business process of contacting the vehicle owner, discussing necessary service amounts, arranging workshop dates, or in a wider scope, it might be used for quality management and product improvements. Remaining fuel range might also be monitored, to trigger other use cases that may propose a cheap gas station on the route.

Analysis of Current Data Exchange Capabilities of the oneM2M Service Platform

The following section provides an overview of the current capabilities for data exchange, based on a selected ASDP scenario, and then follows up with a discussion on recommended enhancements.

Principles

At the core of oneM2M interworking between AEs and *Nodes* is a generic resource tree (RT), located inside each CSE. In addition to the structured storage of application data (such as sensor measurement), the RT facilitates essential functions, such as registration, discovery, deregistration, announcement, grouping, subscription, and notification management. From an implementation perspective, the RT is mapped to uniform resource identifiers (URI) and exposed through the standardized interfaces (*Mca*, *Mcc*, *Mcc*'), following the RESTful architectural style. Accordingly, the resources are manipulated using create, retrieve, update, delete plus notify methods (CRUD+N) [2], which are mapped to the applied application layer protocols, most likely Hypertext Transfer Protocol (HTTP), Constrained Application Protocol (CoAP) [13], or Message Queue Telemetry Transport (MQTT[2]).

According to the middleware approach of oneM2M, AEs exchange application data using the capabilities of the CSE(s), offered through the standardized interfaces. At first, the AE stores application data in the RT structure of their local or a remote CSE. For this it uses a *container* resource, which can contain, besides others, one or many contentInstances, according to specifiable memory constraints, such as *maxNrOfInstances*, *maxByteSize*, and *maxInstanceAge*. The actual application data is then stored within the attribute *content* of a *contentInstance*. The stored application data can be received from other AEs in various ways. For instance, other AEs can retrieve this data on demand. Optionally, to retrieve certain resources or subsets, conditions could be defined that are related, e.g. to time, size, state, label, number of matches, or content type of the resource.

In addition, oneM2M provides a subscribe/notify mechanism, whereby AEs can subscribe to

[2] http://www.mqtt.org

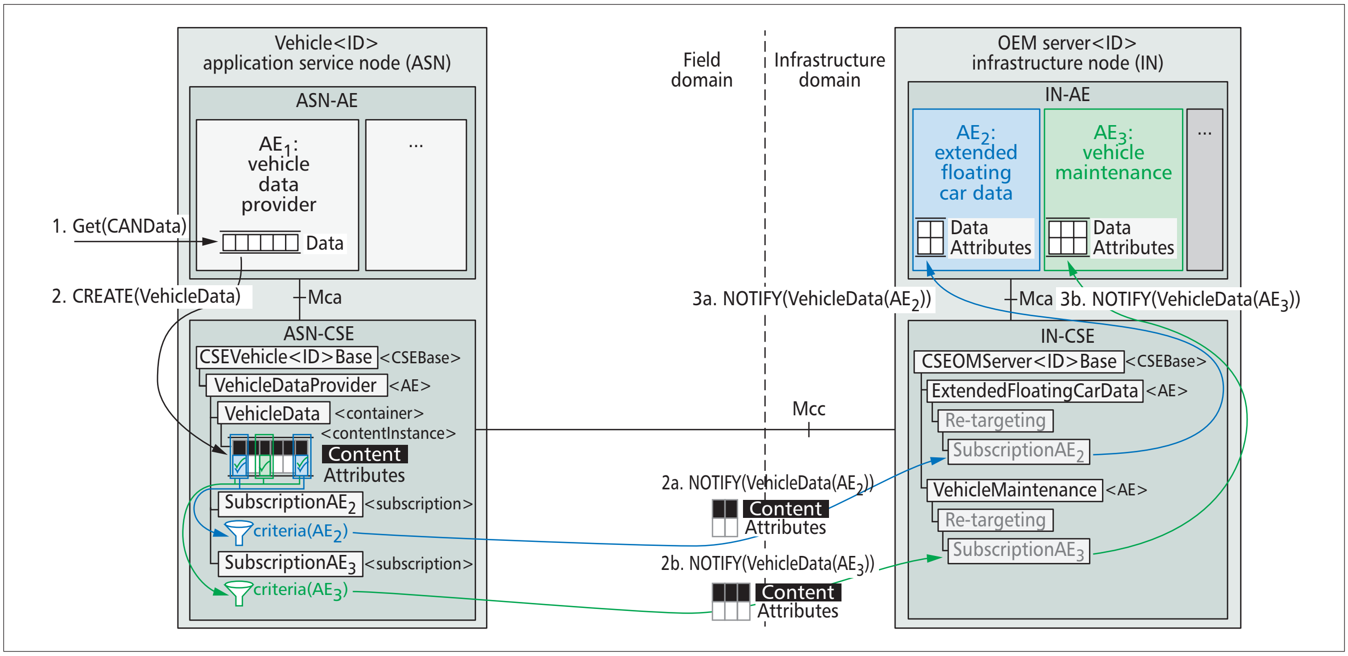


Figure 2. An automotive scenario with current M2M data exchange capabilities.

resource changes. Similar to the retrieve method, constraints can also be defined for the notification, by time, state, size, or status. Furthermore, the subscribe/notify mechanism can include communication-related constraints, from batch notification, rate limit, or priority, to comprehensive notification schedule policies. These capabilities can be used to improve the "network friendliness" of M2M traffic.

Since the subscribe/notify data exchange mechanism provides significant time and space decoupling of AEs, it is particularly suitable for distributed M2M use cases across different devices, vendors, and domains, and is therefore selected for the ASDP use cases.

Analysis

In the following, a simplified scenario derived from the two use cases with one vehicle and one OEM server *node*, and three AEs (Fig. 2), is used for analysis of current data exchange capabilities of oneM2M. Since vehicular sensor data is the foundation for many automotive-related applications, it is made available to all AEs within the ASDP. Accordingly, an *AE*$_1$ "Vehicle Data Provider" was introduced to exemplarily make input data, such as position (latitude, longitude, heading), speed, and ESP control intervention information available in the local CSE RT through appropriately structured *container* resources. The *AE*$_1$ proprietarily obtains the vehicle data from an external source, such as a controller area network fieldbus (CAN-bus) (see step 1). Position data is usually determined using an internal global positioning system (GPS) receiver, connected to the CAN-bus with a typical update rate of 1 Hz to 4 Hz, while other sensor values might be available at a much higher rate. In step 2 the *AE*$_1$ pushes the data, unaware of requirements of (future) AEs within the ASDP, with undiminished resolution to the *container* VehicleData, where it is stored within the attribute *content* of a *contentInstance* resource. The *AE*$_1$ additionally can specify a *contentInfo*, which is a composite attribute of an Internet media type and encoding information, e.g. base64 encoded string. The CSE further expands the *contentInstance* with the typical resource attributes, such as *contentSize* and *creationTime*.

The scenario assumes the existence of two AEs at the IN: an AE$_2$ "Extended Floating Car Data," and an AE$_3$ "Vehicle Maintenance," which implement the respective application logic. Both AEs in this scenario subscribe to the VehicleData of the ASN with appropriate notification criteria. According to the existing oneM2M capabilities, time-related schedules are defined: AE$_3$ configures the scheduleElement, for example, to receive notifications with the latest representation of the content resource at maximum every five seconds; AE$_2$ configures every 10 seconds. These subscriptions are sent through the local IN-CSE, where they are both re-targeted to the target ASN-CSE. For the opposite notifications, local callbacks (within the IN-CSE) are created to route them to the AEs. (These steps are by-passed in Fig. 2). In the given example, the criteria(AE$_2$) according to its configuration selects two contentInstance resources out of six, hence the ASN-CSE performs two notify operations that contain the latest representation of the resource including the VehicleData content (Step 2a). Similarly, the criteria(AE$_3$) selects three contentInstance resources. Thus, the ASN-CSE performs three notifications that contain the latest representation of the resource (Step 2b). Both are in each case re-targeted at the IN-CSE to their originators AE$_2$ (Step 3a) and AE$_3$ (Step 3b).

This example reveals the following two drawbacks of current oneM2M data exchange capabilities.

No Aggregation of Subscriptions: Subscrip-

Description	EPL statement

Notification, if remaining fuel range is smaller than 100 km.	SELECT * FROM VehicleData
	WHERE fuelRange < 100

Notification, if heavy rain is detected.	SELECT * FROM VehicleEnvironmentData
	WHERE rainSensor > 4

Notification, if there is a risk of freezing rain.	SELECT * FROM VehicleEnvironmentData
	WHERE temperature < 3 AND rainSensor > 0

Notification, if there is an electronic stability control intervention.	SELECT * FROM VehicleData
	WHERE ESP=true

Notification, if a strong deceleration of greater than $6 m/s^2$ is detected (calculated on speed delta [km/h] and time delta [s]).	SELECT * FROM pattern[a=Position ->
	b=Position((b.speed -a.speed)/
	((b.timestamp-a.timestamp)*3.6) < -6)]

Table 1. Examples of automotive application-data-dependent notification criteria and their EPL statement representation.

tions are not aggregated or harmonized at the local or transit CSE(s). This potentially leads to redundant data transmissions as the example, illustrated in Fig. 2, shows. Here, five notification messages with the respective latest representation of the VehicleData container are transmitted from the vehicle to the OEM server, whereas only three of these represent different contentInstances; two are redundant.

No Application-Data-Dependent Criteria for Notification: By design-choice of the current oneM2M releases, the *contentInfo* attribute is not used at the CSE level. Accordingly, the application data (stored within the attribute *content* of the *contentInstance* resource) becomes opaque. This has a significant impact on the available notification criteria for the subscribe/notify mechanism, i.e. the *eventNotificationCriteria* conditions. Currently, these cannot refer to the *content*, but can only refer to other attributes of the *contentInstance* resource. For example, *eventNotificationCriteria* can relate to the *creationTime* of the *contentInstance* (with *createdBefore*, *createdAfter* conditions), or to the contentSize (with *sizeAbove*, *sizeBelow* conditions). Application-data-dependent notification criteria that are derived from the real use case requirements, tailoring the transmitted data to the actual needs, cannot be applied.

In the absence of such criteria, it has to be assumed that the inaccurately selected data transmitted needs to be further filtered at the receiving AE. Hence, it is very likely that the network bandwidth consumption of the distributed M2M applications is above the effective requirements of the use case.

Finally, the opaque *content* prevents the specification and detection of application-data-dependent events for notification. This, for instance, allows a short ESP intervention, which is only reflected within the opaque structure of a related *content*, to be missed, unless this value is provided within a separate container resource (on which an "on change" subscription is sufficient). However, the detection of an application-data-dependent threshold exceedance is still not possible.

Proposed Enhancements for Data Exchange Capabilities of the oneM2M Service Platform

This section describes the proposed enhancements for data exchange capabilities of oneM2M. It starts with an overview of its objectives, used to derive a set of requirements, followed by a description of a potential implementation. The discussion concludes with an evaluation on the basis of the introduced scenario.

Objectives and Requirements

For the enhancements of oneM2M standards, we propose to include the following capabilities.

Subscription Aggregation:

- Different subscriptions to the same remote resource shall be aggregated at the local CSE.
- Different subscriptions to the same remote resource should be aggregated at transit CSE(s).

Enhanced Notification:

- Applications shall provide their application data in a standardized way that enables application-data-dependent notification criteria for subscribe/notify mechanisms.
- A comprehensive language shall be provided to enable the standardized description of gainful application-data-dependent notification criteria, including basic arithmetic and logical operations on common data types.

The left column of Table 1 names some advantageous automotive notification criteria with respect to the introduced use cases.

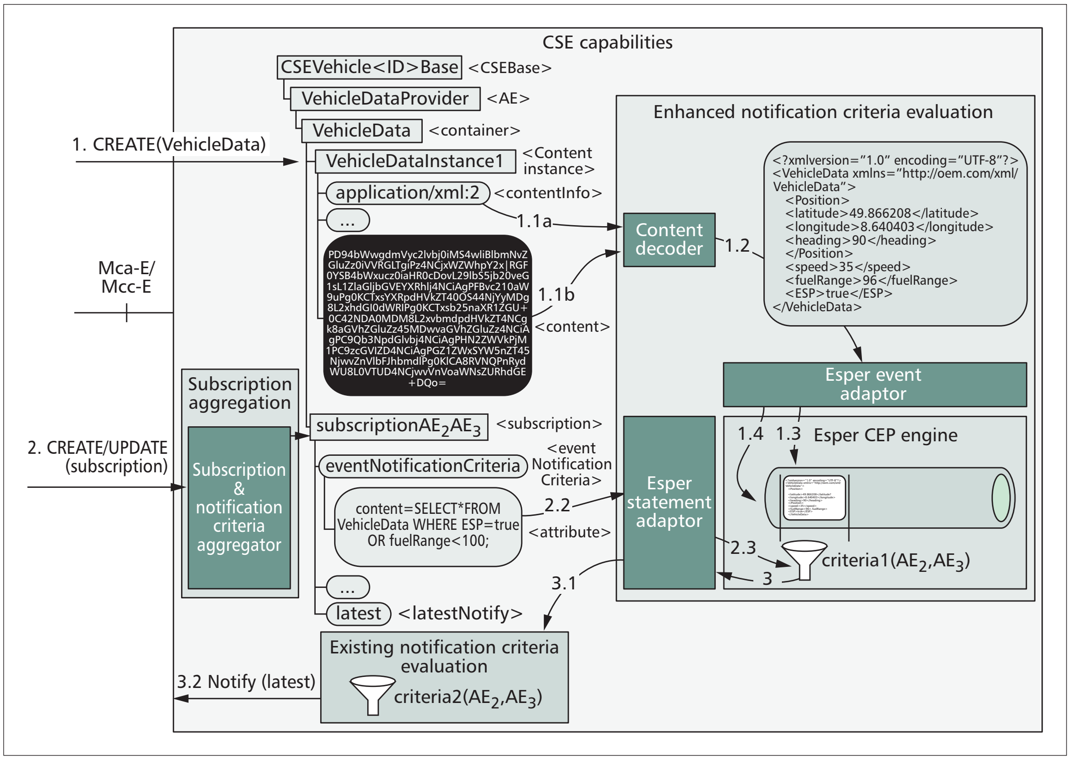
Implementation

The applicability of the proposed enhancements was validated by means of a prototype implementation. The eclipse OM2M project[3] was used for basic functionality of the oneM2M service platform. Comprehensive capabilities for the continuous analysis of data streams are available within the field of complex event processing (CEP) [14]. In contrast to other approaches of CEP for M2M, such as [15], which focus on the analysis of machine-generated data at a server, we use CEP mechanisms at each M2M node to analyze (and filter) the data at its source and tailor the data transmitted. We have integrated the open-source Java Esper CEP[4] component into OM2M to add enhanced data stream processing capabilities to the CSE layer. Figure 3 illustrates the prototype and the interaction of enhancements with existing CSE capabilities within the given scenario.

The starting point is the creation of a new *contentInstance* VehicleDataInstance1 within the *container* VehicleData (step 1). If *contentInfo* is set to "application/xml:2" (which indicates a base64 encoded string of an XML document), this encoding information is forwarded to a content decoder (step 1.1a) together with the related content (step 1.1b) and then, in step 1.2, passed to the Esper event adaptor. The XML document can now be verified against the included link to at least one XML schema definition (XSD), in this example "http://oem.com/xml/VehicleData," which is therefore downloaded. If the Esper event

[3] http://www.eclipse.org/om2m/

[4] http://www.espertech.com/esper/



PD94bWwgdmVyc2lvbj0iMS4wliBlbmNvZ
GluZz0iVVRGLTgiPz4NCjxWZWhpY2x|RGF
0YSB4bWxucz0iaHR0cDovL29lbS5jb20veG
1sL1ZlaGljbGVEYXRhlj4NCiAgPFBvc210aW
9uPg0KCTxsYXRpdHVkZT40OS44NjYyMDg
8L2xhdGl0dWRlPg0KCTxsb25naXR1ZGU+
0C42NDA0MDM8L2xvbmdpdHVkZT4NCg
k8aGVhZGluZz45MDwvaGVhZGluZz4NCiA
gPC9Qb3NpdGlvbj4NCiAgPHN2ZWVkPjM
1PC9zcGVlZD4NCiAgPGZ1ZWxSYW5nZT45
NjwvZnVlbFJhbmdlPg0KlCA8RVNQPnRyd
WU8L0VTUD4NCjwvVnVoaWNsZURhdGE
+DQo=

<?xmlversion="1.0" encoding="UTF-8"?>
<VehicleData xmlns="http://oem.com/xml/
VehicleData">
  <Position>
  <latitude>49.866208</latitude>
  <longitude>8.640403</longitude>
  <heading>90</heading>
  </Position>
  <speed>35</speed>
  <fuelRange>96</fuelRange>
  <ESP>true</ESP>
</VehicleData>

Figure 3. Implementation of application-data-dependent notification criteria within the CSE.

> The applicability of the proposed enhancements was validated by means of a prototype implementation. The eclipse OM2M project was used for basic functionality of the oneM2M service platform. Comprehensive capabilities for the continuous analysis of data streams are available within the field of complex event processing.

adaptor retrieves an XSD file for the first time, this is passed to the Esper CEP engine to register an associated event type (step 1.3). Afterward, in step 1.4, the content XML is sent to Esper as a new event.

The second part of the enhancements refers to the subscription create/update (step 2). Prior to adding the subscription to the remote *container* resource, it is aggregated with existing ones to the same resource at the local CSE, as illustrated in Fig. 4. For the description of application-data-dependent *eventNotificationCriteria*, the Esper Event Processing Language (EPL) is used, whose semantic and syntax is close to the Structured Query Language (SQL). The EPL statements to trigger the notification are enclosed in the *attribute* condition tag, referred to the content. Figure 3 shows the aggregated application-data-dependent EPL statement that facilitates both AE_2 and AE_3 criteria. The overall eventNotificationCriteria consists of the application-data-dependent part, criteria1(AE_2,AE_3), and may have a second part, criteria2(AE_2,AE_3), related to the existing notification criteria on other *contentInstance* attributes. The EPL statement is extracted by the Esper statement adaptor and added to the Esper CEP engine (steps 2.2 and 2.3).

If a new *contentInstance* is created and a related subscription that includes an EPL statement exists, the *content* is forwarded to the Esper CEP engine (steps 1.1a, 1.1b, 1.2). If the EPL statement (criteria1) is fulfilled, the continuation of notification criteria evaluation (criteria2) is triggered (step 3.1). If these are also fulfilled, a notify is sent according to the subscription (step 3.2).

The enhancements were implemented aiming to have a minimal modification on the reference points. But, to provide the enhancements to the AEs and CSEs, the reference points had to be enhanced, which is indicated through the adapted naming *-E. Nevertheless, the application-data-dependent notification criteria remain optional, and AEs can still provide opaque *content* that is transferred according to existing capabilities.

Evaluation

Figure 5 illustrates how the proposed enhancements positively affect the scenario of Fig. 2. It is assumed that initially the AE_3 creates a subscription with criteria(AE_3) on the VehicleData *container*. According to the new capabilities, a criteria derived from the use case "extended floating car data" is used instead of an unspecific time constraint. In this example, this criteria is the interference of the ESP. Afterward, the AE_2 creates a subscription also to the VehicleData with the criteria(AE_2) that requests notifications, if the remaining fuel range is below 100 km. At the time of the second subscription create, the local CSE detects that a subscription to the same remote resource already exists and aggregates the two criteria (denoted criteria(AE_2,AE_3)). Figure 3 shows the resulting aggregated EPL statement. Other EPL examples are listed at the right column of Table 1.

The example illustrated in Fig. 5 assumes that the ESP intervention is true at the first and the last *contentInstance*. It further assumes that the remaining fuel range at the last *contentInstance* is the first time below 100. This leads to a total of two notifications, which are further distributed at the IN-CSE. The AE2 receives two notifications (including respective *latest contentInstance*); the AE_3 receives one notification.

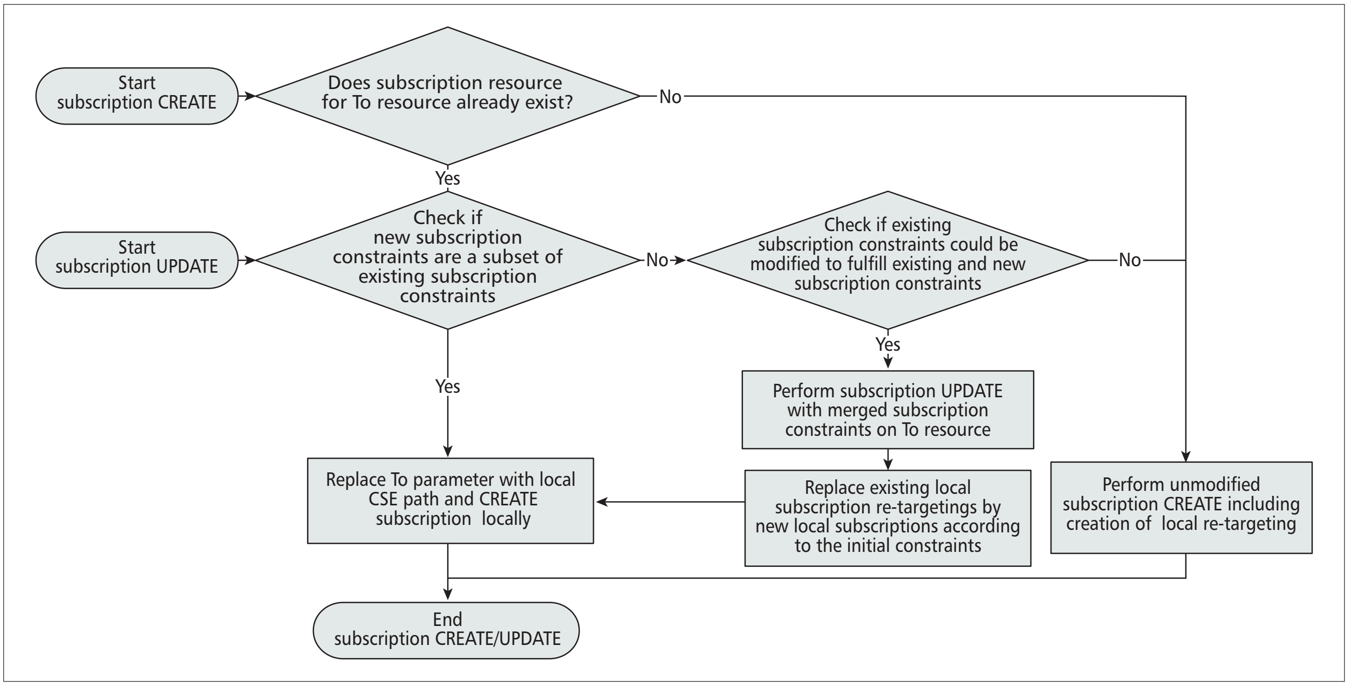

Start
subscription CREATE
Does subscription resource
for To resource already exist?
No
Yes
Start
subscription UPDATE
Check if
new subscription
constraints are a subset of
existing subscription
constraints
No
Check if existing
subscription constraints could be
modified to fulfill existing and new
subscription constraints
No
Yes
Yes
Perform subscription UPDATE
with merged subscription
constraints on To resource
Replace To parameter with local
CSE path and CREATE
subscription locally
Replace existing local
subscription re-targetings by
new local subscriptions according
to the initial constraints
Perform unmodified
subscription CREATE including
creation of local re-targeting
End
subscription CREATE/UPDATE

Figure 4. Flowchart for aggregation of subscription constraints at local CSE.

Discussion

The development of an M2M service platform is a trade-off between an overly application-specific or domain-specific platform (which is just another silo solution) and an overly common platform with too little capabilities, which might be inefficient and again causes silos, i.e. the applications on top. In this regard, any (additional) functionality of the CSE could be controversial.

The existing data exchange capabilities of oneM2M might be sufficient for scenarios where limited sensors provide their measurements at low frequency (and hence have low bandwidth consumption). They are also appropriate, when the whole range of data should be acquired, for example, in the context of big data, where data analysis is performed later offline.

Our proposed enhancements are particularly beneficial for, but not limited to, domains such as automotive, where sensors or nodes possibly provide a high amount of data, of which only certain subsets are required for a use case. In such a scenario, it is neither feasible nor reasonable to transfer all sensor measurements with respect to the large number of vehicles and resulting bandwidth requirements to wireless access networks. The a priori clipping of sensor data available within the oneM2M service platform is not a suitable solution, since the car manufacturer can hardly estimate future use cases and related data requirements. Additionally, they could possibly come from different vendors and domains. Here, as indicated, application-data-dependent notification criteria, together with local aggregation of subscriptions, enable significant bandwidth savings for many use-cases and may make certain functional splits between distributed applications possible.

The capability to detect application-data-dependent events at the CSE level supports oneM2M use cases, related for example to distributed control system applications.

The proposed enhancements of application-data-dependent notification criteria might also improve privacy. On one hand, the capability to better tailor data acquisition to the actual use case requirements can prevent applications from receiving more data than necessary, only due to the lack of appropriate notification criteria capabilities within the CSE. On the other hand, it could be assumed that a binary decision, whether an AE is or is not allowed to perform a certain CRUD+N operation on a resource, will no longer be sufficient. In our opinion, applications may require more detailed access specifications. In this regard, transparent content structures at the CSE level could facilitate future oneM2M enhancements toward application-data-dependent *accessControlPolicies*, similar to the proposed enhancements for notification criteria. For example, enhanced access policies could prohibit applications to:

- Create subscriptions with notification criteria that analyze the vehicle speed.
- Subscribe to position updates of the vehicle with an update interval smaller than every 15 minutes.
- Use the most accurate vehicle position data available.

This, in turn, could enable further differentiation between applications and groups, such as OEM applications, third party applications, safety-related applications, or consumer applications. If users are empowered to configure such enhanced access policies for certain applications, for instance through mechanisms of the ASDP, this could finally be one aspect to better address the “right to privacy” or improve “informational self-determination.”

Our enhancements at this time only use a subset of the overall CEP possibilities. Similarly

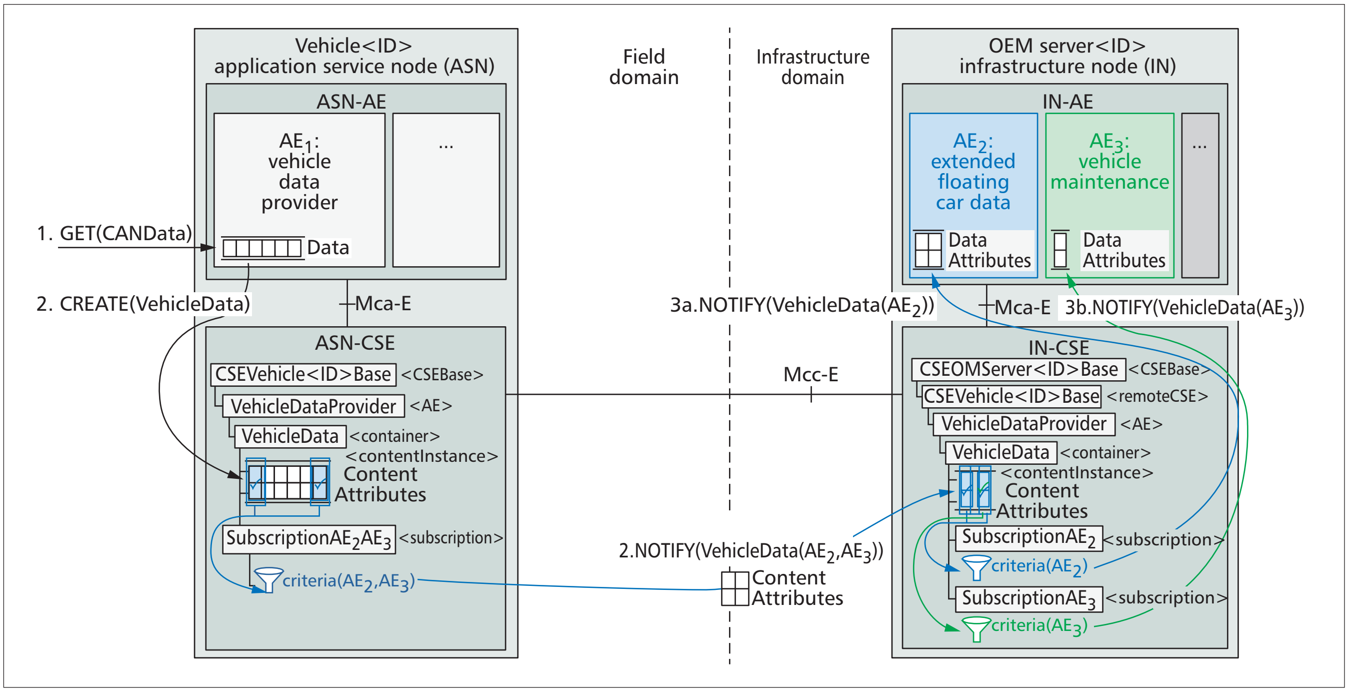


Figure 5. An automotive scenario with enhanced M2M data exchange capabilities.

to the existing oneM2M capabilities, so far the enhanced notification criteria are also limited to the respective resource, where the subscription is placed. This means that criteria across several *containers* are currently not supported. Furthermore, the notification *content* is currently not configurable, which offers possibilities for future enhancements. In this context, ongoing oneM2M standardization activities toward full semantic interoperability are also beneficial. Virtual resources, dynamically created according to the requirements of an AE through semantic mashup, could be one solution to the aforementioned limitations.

Summary

The oneM2M service platform is currently developed as the standardized horizontal enabling platform for smart homes, smart cities, and intelligent transportation systems, all implemented as distributed inter-connected applications. In this context, data exchange capabilities represent a key functionality with respect to network efficiency and possible functional splits between applications and nodes.

This article presented an analysis of data exchange capabilities of current oneM2M service platform specifications by means of vehicular use cases, which are implemented with an M2M-based automotive service delivery platform. We found that the absence of local aggregation of different subscriptions to the same resource can cause redundant data transmissions. Besides, the opaque handling of application data at the common service entity prevents the definition of application-data-dependent notification criteria for the subscribe/notify mechanism. This reduces the capabilities to tailor the data acquisition to the actual requirements of the use case. Furthermore, the detection of application-data-dependent events, such as a threshold exceedance, cannot be used as a notification trigger.

Our proposed enhancements use existing oneM2M attributes, available for technical interworking of different oneM2M applications, to decode the opaque application data at the common service entity level. The introduction of a complex event processing engine facilitates the specification of complex statements, which enable application-data-dependent criteria for notifications, including the detection of events. This enables significant bandwidth savings for many oneM2M usage scenarios, not limited to the automotive domain.

References

[1] ETSI, "Machine-to-Machine Communications (M2M); Functional Architecture," TS 102 690, V.2.1.1, Oct. 2013.

[2] oneM2M, "Functional Architecture," TS-0001, V.1.6.1, Jan. 2015.

[3] oneM2M, "Study of Abstraction and Semantics Enablements," TR-0007, V.2.5.1, July 2015.

[4] M. Glaab *et al.*, "A M2M-based Automotive Service Delivery Platform for Distributed Vehicular Applications," *Proc. 10th Int'l. Network Conf.*, Plymouth, UK, 2014, pp. 35–45.

[5] M. Broy *et al.*, "Engineering Automotive Software," *Proc. IEEE*, vol. 95, no. 2, Feb. 2007, pp. 356–73.

[6] T. Kosch *et al.*, "Communication Architecture for Cooperative Systems in Europe," *IEEE Commun. Mag.*, vol. 47, no. 5, May 2009, pp. 116–25.

[7] M. Broy, "Challenges in Automotive Software Engineering," *Proc. 28th Int'l. Conf. Software Engineering*, 2006, pp. 33–42.

[8] S. Bauer, "Das vernetzte Fahrzeug — Herausforderungen für die IT," *Informatik Spektrum*, vol. 34, no. 1, Dec. 2010, pp. 38–41.

[9] A. Pretschner *et al.*, "Software Engineering for Automotive Systems: A Roadmap," *Proc. 2007 Future of Software Engineering*, May 2007, pp. 55–71.

[10] M. Glaab, W. Fuhrmann, and J. Wietzke, "Entscheidungskriterien für die Verteilung zukünftiger automotiver Anwendungen im Kontext vernetzter Fahrzeuge," *Proc. Mobilkommunikation 2011 — Technologien und Anwendungen — 16. ITG-Fachtagung*, Osnabrück, DE, 2011, pp. 149–54.

[11] G. Wu *et al.*, "M2M: From Mobile to Embedded Internet," *IEEE Commun. Mag.*, vol. 49, no. 4, Apr. 2011, pp. 36–43.

[12] W. Huber, M. Lädke, and R. Ogger, "Extended Floating-Car Data for the Acquisition of Traffic Information," *Proc. 6th World Congr. Intelligent Transportation Systems*, 1999, pp. 1–9.

[13] C. Bormann, A. P. Castellani, and Z. Shelby, "CoAP: An Application Protocol for Billions of Tiny Internet Nodes," *IEEE Internet Comp.*, vol. 16, no. 2, 2012, pp. 62–67.

[14] D. C. Luckham, *The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems, 5th Edition*, Boston, USA: Addison-Wesley Professional, 2002.

[15] R. Bruns *et al.*, "Intelligent M2M: Complex Event Processing for Machine-to-Machine Communication," *Expert Systems with Applications*, vol. 42, no. 3, Feb. 2015, pp. 1235–46.

Biographies

Markus Glaab (markus.glaab@h-da.de) is currently a Ph.D. student at the Centre for Security, Communications and Network Research at Plymouth University, UK. He was also a researcher with the In Car Multimedia Labs at the University of Applied Sciences Darmstadt, DE, where he received his M.Sc. in computer science in 2009. Prior to academia, he gained professional experience as a software developer in the area of Car2X communication. His research interests include automotive software engineering, connected vehicles, intelligent transportation systems, and machine-to-machine communication.

Woldemar Fuhrmann received his diploma degree in electrical engineering in 1974 and a Ph.D. degree in 1978, both from the University of Erlangen. He has a long professional background in telecommunications with Philips, Siemens, Detecon, and Deutsche Telekom. In 1990 he joined the University of Applied Sciences in Dieburg, DE as a professor in telecommunications, and since 2000 he has been a professor at the University of Applied Sciences in Darmstadt, DE. His interests are in wireless communication, software-defined networks, and service delivery platforms.

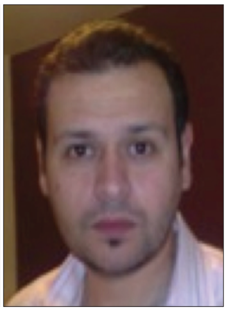
Joachim Wietzke is currently a professor for embedded systems on the Faculty of Mechanical Engineering and Mechatronics at the Karlsruhe University of Applied Sciences, DE. Formarly he was a professor for informatics and headed the In-Car Multimedia Labs at the University of Applied Sciences Darmstadt, DE. Prof. Wietzke has been involved in a number of industrial projects related to software project task forces as well as next generation ICM systems. Prior to his professorship he gained extensive industrial experience in leading positions of software development units at Bosch, Blaupunkt, and Harman/Becker. His research interests include next generation ICM systems and HMIs.

Bogdan Ghita received his Ph.D. in 2005 from Plymouth University, UK. He is an associate professor at Plymouth University and leads the networking area within the Centre for Security, Communications, and Network research. His research interests include computer networking and security, focusing on the areas of network performance modelling and optimization, wireless and mobile networking, and security. He has been principal investigator in a number of industry-led, national, and EU research projects. He was a TPC member for more than 40 international conference events, as well as a reviewer for *IEEE Communications Letters*, computer communications, and future generation computer systems journals, and he is the chair of the International Networking Conference series.

Research and Standards: Advanced Cloud and Virtualization Techniques for 5G Networks (Part II)

Kan Zheng

Tarik Taleb



Adlen Ksentini

Chih-Lin I

Thomas Magedanz

Mehmet Ulema

The evolution of mobile network architecture is an essential part in the development process of the fifth generation (5G) of cellular mobile systems, and that is through the incorporation of advanced cloud technologies and network function virtualization techniques. The new network architecture needs to support a wide range of high data rate applications and services, offering capacities of up to multiple gigabits per second, yet meeting extremely stringent latency and reliability requirements under a diverse variety of scenarios. Thus, the automation of network control and overall system management to achieve such an ambitious set of performance targets became crucial. Significant global effort for the necessary new technologies has been initiated.

Emerging paradigms, such as and not limited to, Software-Defined Networking (SDN) and Network Function Virtualization (NFV), represent a first concrete step toward this direction, catalyzing the idea of decoupling the software defined control plane from the hardware driven data plane, and thus the virtualization of network functions on general purpose hardware. There is an increasing trend toward implementing more and more functions of mobile communications systems in software, e.g. for signal and protocol processing. This will significantly influence the future development of 5G technologies and architectures. In addition to savings in both operational and capital expenditures, the introduction of logically centralized controllers enables the employment of various intelligent control algorithms. However, how to leverage the benefits of NFV in both the Radio Access Network (RAN) and the mobile Core Network (CN) is yet to be investigated fully.

The aim of this special issue is to highlight the 5G requirements and how they can be met through novel mobile network architecture designs. The special issue mainly focuses on new technologies being researched that have great potential to impact standards. Original contributions were solicited on topics of relevance to the evolution of mobile network architectures, and particularly on SDN, NFV and cloud computing.

In Part I of the special issue [1], the main reported research contributions were: Computation Offloading at Ad Hoc Cloudlet [2], Software Defined Networking (SDN): Architecture for Virtualized Wireless Access for Cellular Network Security [3–5], 5G visions: Network Architecture, Virutal RATs and Flexible Tailored Radio Access Network, Cloud Assisted HetNets, Content Distribution over Content Centric Mobile Social Networks [6–9]. In Part II of the special issue, the following set of papers have been accepted.

In "An Effective Approach to 5G: Wireless Network Virtualization," Feng *et al.* present a layered model of wireless network virtualization. Correspondingly, they design a hierarchical control scheme to support their proposed model. Finally, two use cases have been analyzed to show the efficiency of their schemes. Their methods might be a feasible solution to realize wireless network virtualization in 5G systems.

In order to guide the deployment strategy of virtualized mobile networks, the authors of "Cost Analysis of Initial Deployment Strategies for a Virtualized Mobile Core Network Functions" compared two constraint-based heuristic approaches for the deployment of virtualized Evolved Packet Core (vEPC) and Virtualized Network Functions Component (VNFC). Also, the impacts of these deployment strategies on the cost are analyzed.

Meanwhile, another new way to improve the efficiency of 5G networks consists in device-to-device (D2D) communications. In the third article, "Buffer-Aided Device-to-Device Communication: Opportunities and Challenges" Zhang *et. al.*, propose a dynamic graph model for D2D communication underlaying cellular networks. Then they analyze the system constraints to form a weighed directional graph optimization. The influence of the helper to subscriber ratio is discussed as well as the allocation of the system bandwidth to cellular and D2D communications on the achievable system performance.

In the fourth article, "Benefits and Challenges of Virtualization in 5G Radio Access Networks," the authors present their views on the benefits, challenges, and limitations that accompany virtualization in 5G radio access networks (RANs). The implementation requirements and the cost are also analyzed. They discuss their impact on standardization such as in 3GPP.

In the last article, "XG-FAST: The 5th Generation Broadband," experts from Alcatel-Lucent, Bell Labs give us the new picture to use the software defined networking (SDN) technique. In this article, they first discuss the role of XG-FAST in future mobile network architectures. Then, the XG-FAST system concept is introduced with the improved techniques. The measurements are performed on several short copper cables to shown the effectiveness of the XG-FAST technology. Finally, they present the next step to put it into practice.

As Guest Editors, we would like to thank all the authors for their submissions to this Feature Topic. The interest and quality of submissions were beyond our imagination. We are also grateful to the reviewers for the timely responses and their valuable comments to improve the quality of the articles. We appreciate the support from both Mr. Glenn Parsons, current Editor-in-Chief of *IEEE Communications Magazine* Supplement on Communications Standards, and Dr. Osman S. Gebizlioglu, Editor-in Chief of *IEEE Communications Magazine*. We also appreciate the help of Joseph Milizzo, Jennifer Porcello, and Charis Scoggins throughout the publication process. Finally, our hope is that the readers of *IEEE Communications Magazine* Supplement on Communications Standards enjoy the articles of this Feature Topic, and would consider contributing to future editions.

References:

[1] K. Zheng *et al.*, "Research & Standards: Advanced Cloud & Virtualization Techniques for 5G Networks," Guest Editorial, Communications Standards Supplement, *IEEE Commun. Mag.*, vol. 53, no. 6, June 2015, pp. 16–17.

[2] M. Chen *et al.*, "On the Computation Offloading at Ad Hoc Cloudlet: Architecture and Service Modes," Communications Standards Supplement, *IEEE Commun. Mag.*, vol. 53, no. 6, June 2015, pp. 18–24.

[3] F. Granelli *et al.*, "Software Defined and Virtualized Wireless Access in Future Wireless Networks: Scenarios and Standards," Communications Standards Supplement, *IEEE Commun. Mag.*, vol. 53, no. 6, June 2015, pp. 26–34.

[4] A. Bradai *et al.*, "Cellular Software Defined Networking: A Framework," Communications Standards Supplement, *IEEE Commun. Mag.*, vol. 53, no. 6, June 2015, pp. 36–43.

[5] M. Dabbagh *et al.*, "Software-Defined Networking Security: Pros and Cons," Communications Standards Supplement, *IEEE Commun. Mag.* , vol. 53, no. 6, June 2015, pp. 73–79.

[6] J. Wang *et al.*, "i-Net: New Network Architecture for 5G Networks," Communications Standards Supplement, *IEEE Commun. Mag.*, vol. 53, no. 6, June 2015, pp. 44–51.

[7] S. Chen *et al.*, "Virtual RATs and a Flexible and Tailored Radio Access Network Evolving to 5G," Communications Standards Supplement, *IEEE Commun. Mag.*, vol. 53, no. 6, June 2015, pp. 52–58.

[8] N. Zhang *et al.*, "Cloud Assisted HetNets Toward 5G Wireless Networks," Communications Standards Supplement, *IEEE Commun. Mag.*, vol. 53, no. 6, June 2015, pp. 59–65.

[9] Z. Su and Q. Xu, "Content Distribution over Content Centric Mobile Social Networks in 5G," Communications Standards Supplement, *IEEE Commun. Mag.*, vol. 53, no. 6, June 2015, pp. 66–72.

Biographies

Kan Zheng [SM] (zkan@bupt.edu.cn) is currently a professor at Beijing University of Posts &Telecommunications (BUPT), China. He received the B.S., M.S., and Ph.D. degrees from BUPT, China, in 1996, 2000, and 2005, respectively. He has extensive industry experience in the standardization of the new emerging technologies. He is the author of more than 200 journal articles and conference papers in the field of resource optimization in wireless networks, M2M networks, VANET, and so on. He holds editorial board positions for several international journals. He has organized several special issues in famous journals including *IEEE Communications Surveys & Tutorials* and *Transactions on Emerging Telecommunications Technologies* (ETT). He was the general Vice-Chair of Mobiquitous 2012, and workshop co-chair on QoE in Energy-Efficient Wireless Networks in IEEE ISCIT' 2012. Also, he was the TPC Co-Chair of IEEE PIMRC 2013, and TPC Track Chair of IEEE WiMob 2015 and IEEE SmartGridComm 2015. He has served as a TPC member of IEEE conferences including INFOCOM, ICC, Globecom, and VTC.

Tarik Taleb (Tarik.Taleb@neclab.eu) is an IEEE Communications Society (ComSoc) Distinguished Lecturer and a senior member of IEEE. He is currently a professor at the School of Engineering, Aalto University, Finland. Prior to his current position, he was working as a senior researcher and 3GPP standards expert at NEC Europe Ltd, Heidelberg, Germany. He was then leading the NEC Europe Labs Team working on R&D projects on carrier cloud platforms. Before his work at NEC and until Mar. 2009, he worked as an assistant professor at the Graduate School of Information Sciences, Tohoku University, Japan, in a lab fully funded by KDDI, the second largest network operator in Japan. From October 2005 to March 2006 he was working as research fellow with the Intelligent Cosmos Research Institute, Sendai, Japan. He received his B.E. degree in information engineering with distinction, and M.Sc. and Ph.D. degrees in information sciences from GSIS, Tohoku Univ., in 2001, 2003, and 2005, respectively. His research interests are in the field of architectural enhancements to mobile core networks (particularly 3GPP's), mobile cloud networking, mobile multimedia streaming, congestion control protocols, handoff and mobility management, inter-vehicular communications, and social media networking. He has also been directly engaged in the development and standardization of the Evolved Packet System as a member of 3GPP's System Architecture Working Group. He is a board member of the IEEE Communications Society Standardization Program Development Board. In an attempt to bridge the gap between academia and industry, he has founded and has been the general chair of the "IEEE Workshop on Telecommunications Standards: from Research to Standards," a successful event that was given the "best workshop award" by the IEEE Communication Society (ComSoC). Based on the success of this workshop, he has also founded and has been the steering committee chair of the IEEE Conference on Standards for Communications and Networking (IEEE CSCN). He is/was on the editorial board of the *IEEE Transactions on Wireless Communications*, *IEEE Wireless Communications Magazine*, *IEEE Transactions on Vehicular Technology*, *IEEE Communications Surveys & Tutorials*, and a number of Wiley journals. He is serving as chair of the Wireless Communications Technical Committee, the largest in IEEE ComSoC. He also served as Secretary and then as Vice Chair of the Satellite and Space Communications Technical Committee of IEEE ComSoc (2006–2010). He has been on the technical program committee of different IEEE conferences, including Globecom, ICC, and WCNC, and chaired some of their symposia. He is the recipient of the 2009 IEEE ComSoc Asia-Pacific Best Young Researcher Award (June 2009), the 2008 TELECOM System Technology Award from the Telecommunications Advancement Foundation (March 2008), the 2007 Funai Foundation Science Promotion Award (April 2007), the 2006 IEEE Computer Society Japan Chapter Young Author Award (December 2006), the Niwa Yasujirou Memorial Award (February 2005), and the Young Researcher's Encouragement Award from the Japan chapter of the IEEE Vehicular Technology Society (VTS) (October 2003). Some of his research work has also been awarded best paper awards at prestigious conferences.

Adlen Ksentini [SM] (adlen.ksentini@irisa.fr) is an associate professor at the University of Rennes 1, France. He is a member of the INRIA Rennes team Dionysos. He received an M.Sc. in telecommunication and multimedia networking from the University of Versailles. He obtained his Ph.D. degree in computer science from the University of Cergy-Pontoise in 2005, with a dissertation on QoS provisioning in IEEE 802.11-based networks. His other interests include: future Internet networks, mobile networks, QoS, QoE, performance evaluation, and multimedia transmission. He is involved in several national and European projects on QoS and QoE support in future wireless and mobile networks. Dr. Ksentini is a co-author of more than 60 technical journal and international conference papers. He received Best Paper Awards from IEEE ICC 2012 and ACM MSWiM 2005. He is TPC Chair of the IEEE Third Workshop on Standards on Telecommunication (collocated with Globecom 2014), and workshop chair of the ACM/IEEE QShine 2014. He is a guest editor for the *IEEE Wireless Communication Magazine* Special Issue on Research & Standards: Leading the Evolution of Telecom Network Architectures. He has been on the technical program committee of major IEEE ComSoc conferences, including ICC/Globecom, ICME, WCNC, and PIMRC.

Chih-Lin I (icl@chinamobile.com) is the China Mobile Chief Scientist of Wireless Technologies, in charge of advanced wireless communication R&D efforts at China Mobile Research Institute (CMRI). She established the Green Communications Research Center of China Mobile, spearheading major initiatives including 5G key technologies R&D; high energy efficiency system architecture, technologies, and devices; green energy; C-RAN and soft base station. She received her Ph.D. degree in electrical engineering from Stanford University. She has almost 30 years experience in wireless communication area. She has worked in various world-class companies and research institutes, including the Wireless Communication Fundamental Research Department at AT&T Bell Labs; headquarters of AT&T, as Director of Wireless Communications Infrastructure and Access Technology; ITRI of Taiwan, as Director of Wireless Communication Technology; Hong Kong ASTRI, as its VP and the Founding GD of its Communications Technology Domain. She received the *IEEE Transactions on Communications* Stephen Rice Best Paper Award, and is a winner of CCCP "National 1000 talent" program. She was an elected Board Member of IEEE ComSoc, Chair of the ComSoc Meeting and Conference Board, and the Founding Chair of the IEEE WCNC Steering Committee. She is currently the Chiar of FuTURE Forum 5G SIG, a Steering Board Member of WWRF, an Executive Board Member of GreenTouch, and a Network Operator Council Member of ETSI NFV.

Thomas Magedanz, Ph.D. (thomas.magedanz@fokus.fraunhofer.de) is a professor of the chair for Next Generation Networks (AV — Architektur der Vermittlungsknoten in German) on the electrical engineering and computer sciences faculty of the Technische Universität Berlin, Germany, where he is educating Master and Ph.D. students in the converging fields of SDN-based control platforms for multimedia and M2M/IOT applications on top of converging fixed and mobile broadband networks. In addition, he leads the Next Generation Network Infrastructures (NGNI) Competence Center at Fraunhofer Institute FOKUS in Berlin, Germany, where he is responsible for the performance of major international R&D co-operations and related academic and industry projects in the context of future seamless communication infrastructure prototyping. In this context he is a globally recognized pioneer in the development and delivery of advanced network and service technology software tools, known as the OpenXXX toolkits, and related testbeds, known as the FOKUS playgrounds, in the fields of Mobile Next Generation Networks. Well known examples include the OpenIMSCore, OpenEPC OpenMTC and the new Open5GCore, as well as the Open IMS Playground, the FUSECO-Playground and the new Open 5G Playground being part of the 5GBerlin testbed.

Mehmet Ulema (mehmet.ulema@manhattan.edu) is a professor in the Computer Information Systems Department at Manhattan College, New York. Previously, he held management and technical positions in Daewoo Telecom, Bellcore (now called Telcordia), AT&T Bell Laboratories, and Hazeltine Corporations. He is an active member of IEEE. Currently, he is a ComSoc Director of Standards Development. He served as the chair and co-founder of the IEEE Communications Society's Information Infrastructure Technical Committee. He is involved in numerous IEEE conferences. He was a General Co-Chair of IEEE BlackseaCom 2014. He was the Technical Program Chair for the IEEE Global Communications (Globecom) Conference) in 2009. He was the General Co-Chair of IEEE Network Operations and Management Symposium (NOMS) in 2008, the Program Chair of the IEEE International Communications Conference (ICC) in 2006, and the IEEE Consumer Communications and Networking Conference in 2004. He received M.S. & Ph.D. degrees in computer science at Polytechnic University (now called Polytechnic Institute of New York University), Brooklyn, New York, U.S.A. He also received B.S. & M.S. degrees from Istanbul Technical University, Turkey.

An Effective Approach to 5G: Wireless Network Virtualization

The explosively growing demands for mobile service bring both challenges and opportunities to wireless networks, giving birth to fifth generation (5G) mobile networks. The features and requirements of different services are diverse in 5G. The management and coordination among heterogeneous networks, applications, and user demands need the 5G network to be open and flexible to ensure that network resources are allocated with high efficiency.

Zhiyong Feng, Chen Qiu, Zebing Feng, Zhiqing Wei, Wei Li, and Ping Zhang

Zhiyong Feng, Chen Qiu, Zebing Feng, Zhiqing Wei, and Ping Zhang are with the Beijing University of Posts and Telecommunications.

Zhiyong Feng is the corresponding author.

Wei Li is with the University of Victoria.

This work was supported by the National High-Tech R&D Program (863 Program 2015AA01A705), the Funds for Creative Research Groups of China (61421061), the Beijing Municipal Science and Technology Commission research fund project No. D151100000115002 and the National Natural Science Foundation of China (61201152).

Abstract

Nowadays, the explosively growing demands for mobile service bring both challenges and opportunities to wireless networks, giving birth to fifth generation (5G) mobile networks. The features and requirements of different services are diverse in 5G. The management and coordination among heterogeneous networks, applications, and user demands need the 5G network to be open and flexible to ensure that network resources are allocated with high efficiency. To fulfill these requirements, wireless network virtualization is used to integrate heterogeneous wireless networks and coordinate network resources. In this article we propose a model of wireless network virtualization consisting of three planes: the data plane, the cognitive plane, and the control plane. A novel control signaling scheme has also been designed to support the proposed model. From the implementation perspective of network virtualization, a hierarchical control scheme based on cell-clustering has been used with dynamically optimized efficiency of resource utilization. Two use cases have been analyzed to demonstrate how the schemes work under the proposed model to improve resource efficiency and the user experience.

COMMUNICATIONS STANDARDS

Introduction

It is estimated that by the year 2020 the traffic carried by wireless networks will be one thousand times greater than in 2010 [1]. Hence, the 5G wireless communication system is expected to offer comparable network capacity, and at the same time support a large number of simultaneously connected devices with diverse mobile services [2]. Research and development efforts from both the academic and industrial communities, such as IMT-2020 in China, METIS in Europe, and the 5G Forum in Korea, have been devoted to the upcoming 5G wireless communication system.

Spatial densification is an effective approach to achieve the required performance under a complex communication background with heterogeneous network deployments of macrocells, picocells, femtocells, relays, and device-to-device (D2D) communication [3, 4], among others. 5G networks are expected to feature great heterogeneity in network deployment, supported services, traffic features, and QoS requirements. In network deployments, 5G networks will confront the coexistence of macrocells, small cells, and various access points in wireless local area networks (WLANs), relay stations, and so on. There are various services such as voice, data, online gaming, social network services, and machine-to-machine (M2M) services in 5G systems. The traffic features and requirements of different services are diverse, e.g. different levels of QoS or security requirements. This presents great challenges for 5G network design in a heterogeneous environment. 5G systems will require flexibility in the management and coordination among of heterogeneous networks, applications, and user demands by allocating network resource in an optimized approach. To provide users with a smooth service experience, wireless network virtualization is introduced to integrate and coordinate heterogeneous wireless networks and resources.

Network virtualization can be regarded as a process of sharing the entire network system with an optimized approach [5]. Therefore, wireless network virtualization can go beyond the physical network borders and realize the convergence of heterogeneous wireless networks. By utilizing virtualization technology, we can integrate different radio access networks (RANs) to provide unified services to users. Therefore, it is feasible to optimize the allocation of resources while guaranteeing the required performance. In addition, network management can be simplified through the abstraction of network functions and centralized control. Moreover, capital expenditure (CAPEX) and operating expense (OPEX) for operators can be reduced by network virtualization [5].

In wired networks, virtualization has been applied since the introduction of virtual private networks (VPNs) over wide area networks (WANs) and virtual local area networks (VLANs) in enterprise networks. Recently, software defined networking (SDN) has been widely studied as a promising technology for network virtualization [6]. Those technologies shed some light on what we can do for 5G network virtualization. However, because wireless networks have unique characteristics, the approach of existing virtualization can not be applied directly.

In commercial wireless networks, some initial approaches and the promotion of network sharing have been developed by standardization organizations. For example, 3GPP [7] has defined two architectures for network sharing: multi-operator core network (MOCN) and gateway core network (GWCN). In MOCN, only RANs with corresponding resource are shared. GWCN allows the sharing of the whole network including the core network. Meanwhile, there are five scenarios defined in [8] that fall into three categories. The first category is RAN sharing. Different operators may share common RANs instead of the radio frequency. It is noticed that RANs owned by different operators may cover different geographical areas. The second catego-

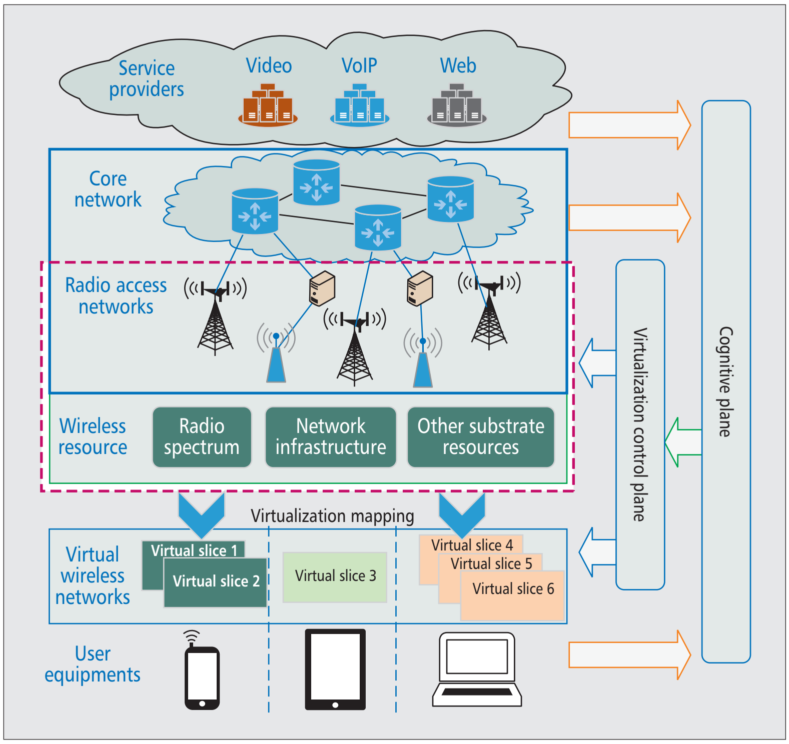


Figure 1. Model of wireless network virtualization.

ry is spectrum sharing, where the licensed spectrum can be shared by operators, probably by pooling the spectrum. The last category is core network sharing. Multiple RANs can belong to different network operators. Due to operators' deployment, different nodes or part of the common core network can be shared. Wireless virtualization can provide an integrated solution for the first two categories.

In this article we propose a model of wireless network virtualization consisting of three planes: the data plane, the cognitive plane, and the control plane. In addition, we propose a cell clustering based hierarchical control scheme for wireless network virtualization that is designed to coordinate the virtualized networks. Furthermore, we design the establishment and adjustment of signaling in wireless network virtualization. Finally, two use cases are provided to demonstrate how the proposed scheme works and improves resource efficiency and the user experience.

Model of Wireless Network Virtualization

In this section we address 5G heterogeneous wireless networks with the ability to be virtualized. Therefore, wireless resources, such as spectrum and infrastructures, should be virtualized to enable flexible access by users. This means that wireless resources can be dynamically and properly allocated to user equipment (UE) to meet specific service requirements. In the virtualization of 5G heterogeneous wireless networks, a new design of wireless resource management and the corresponding control signaling is required.

The available resource for a specific UE can be denoted as $R = RAN$, Capacity, where *RAN* denotes the available radio access networks for users to access, e.g. LTE, 3G, or WLAN. The achieved capacity for each RAN is measured by bps (bits per second) which is related to the design and implementation of RANs. Thus, any UE have access to proper wireless networks with high probability that they will be allocated with the desired capacity for their QoS requirements.

Meanwhile, to achieve high efficiency, network resources should be dynamically allocated among different base stations (BSs) and access points (APs) to provide UEs with on-demand service. Each RAN owns radio resources in the multi-domain of space, time, frequency, and so on. These resources need to be reasonably assigned in different RANs to provide sufficient capacity. Besides these, computing and storage resources can be taken into consideration in wireless network virtualization. High speed and efficient processors can accelerate the data computing process, e.g. fast Fourier transformation (FFT) in orthogonal frequency division multiplexing (OFDM). Faster allocation algorithm operation and response can also facilitate network resource management. Storage devices such as RAM or a hard disk embedded at the BS or UE can provide a buffer for mobile services. Users' requested services can be stored at the buffer of the BS, which provides a low latency response and high rate transmission for wireless links. Utilization of computing and storage resources may bring extra capacity and decrease transmission latency. A combined utilization of radio, computing, and storage resources can be one of the key features of future 5G wireless networks.

Radio resources can be split into slices for allocation to satisfy varying user requirements. Note that appropriate slice granularity needs to be considered since excessively small resource slices increase management complexity, and resources could possibly be wasted with excessively large resource slices. Both situations decrease resource utilization efficiency. Furthermore, it is crucial to prescribe the geographic scope where specific resources can be utilized to avoid interference from nearby UEs and raise resource reuse opportunities among UEs that are far away from each other.

Wireless network virtualization can match resources with user requirements by dynamically and flexibly slicing the infrastructure and resources into virtual networks to achieve global optimized resource utilization [9]. As illustrated in Fig. 1, the model of wireless network virtualization consists of three planes: the data plane, the cognitive plane, and the control plane. The functions of each plane are described as follows.

Data Plane: The data plane is responsible for the transmission of user traffic via virtual networks. Heterogeneous radio access networks such as LTE, 3G-like access networks, WLAN, etc., and corresponding wireless resources, are sliced into virtual networks, which carry user traffic. To ensure that virtual networks operate normally, wireless resources, including spectrum and infrastructure, are allocated.

Cognitive Plane: The cognitive functions in this plane capture the state of services, network resources, and user requirements. The state information is collected from UEs, RANs, and service providers, and then transmitted to the

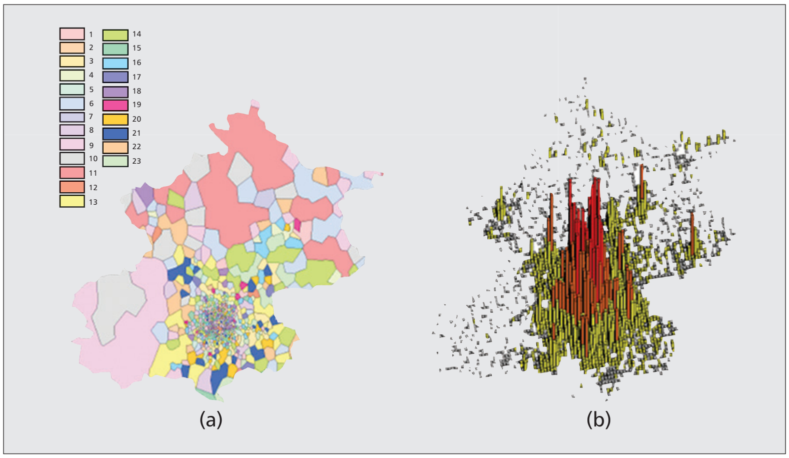


Figure 2. Clustering result in the Beijing area: a) cell clustering; b) traffic distribution.

control plane, which can be helpful to properly slice and allocate the infrastructure and resources to allocate them to the UEs [10].

Control Plane: The control plane is in charge of integrating all the available RANs and resources together and slicing them into virtual networks. All virtual networks are independent and there is no interference and conflict between them. Each of them consists of network infrastructure and wireless resources to be allocated to UEs.

The self-organization of wireless network virtualization can be realized in this model. RANs and resources are virtualized as virtual networks, and users can achieve their traffic requirements without detailed knowledge, such as which RAN to access and which resources they are getting. Virtual networks are isolated to guarantee that there is no serious interference among users.

It is obvious that not all control functions can be integrated into the virtualization control plane. To deal with the time varying fluctuations in wireless channels, many control signaling messages in low layers are designed to ensure successful transmission. However, such control signaling is highly coupled with a particular transmission and is quite sensitive to delay. Thus, the low-layer control functions cannot be implemented in the virtualization control plane. Take the Long Term Evolution (LTE) system as an example. The physical (PHY), media access control (MAC), radio link control (RLC), and Packet Data Convergence Protocol (PDCP) layers are highly coupled with wireless transmission, and thus should not be replaced by the virtualization control plane. The non-access stratum (NAS) and radio resource control (RRC) are responsible for different layers of resource allocation and thus can be implemented in the virtualization control plane to achieve high resource utilization.

Hierarchical Virtualization Control Scheme

Cloud radio access networks have a typical centralized control architecture, where the control functions are implemented in a cloud computing plane [11, 12]. In contrast, our proposed hierarchical virtualization control scheme organizes classified cells into communities based on their traffic features. The control functions inside one community have a certain degree of autonomy. However, there is high-level centralized control to provide inter-community coordination among cell communities, making it possible to provide efficient control for ultra-dense 5G networks. In the following sections of this article we first analyze the clustering character of mobile cellular systems as a basic guideline of community classification. Then the cell clustering based hierarchical virtualization control scheme is proposed with control signaling for virtualization in details.

Cell Clustering

Mobile cellular networks are comprised of fluctuating resources. First, different kinds of cells (e.g. macrocell, microcell, picocell, femtocell) are configured with different wireless transmission parameters, e.g. frequency band, power, location, etc. On the other hand, a network snapshot at any time shows various resource utilization conditions, e.g. traffic load, available data channels, etc. Cell resource utilization conditions are related to mobile users' traffic, which generates different quantities of data traffic and occupies certain amounts of radio resources. Thus, we have analyzed several gigabytes of BS data collected from some operators and observed that certain number of cells show similar characteristics for resource utilization. For example, most cells in a residential area reveal a tidal effect in the traffic volume, with similar trends in commercial areas or college areas.

We have performed cell clustering with the data of all the BSs in Beijing, which contains the BSs' latitude and longitude information, cell types (macrocell, microcell, picocell), site number, transmit power, downlink and uplink data channel utilization (one week), and so on. The clustering results are illustrated in Fig. 2a and Fig. 2b, which have regionalization cell community features. Such results are useful for hierarchical control, because the virtualization control plane can reasonably adjust resource slices among intra-communities and inter-communities according to their various traffic requirements. For example, assume the cells in one community show regular traffic delivering, i.e. a large amount of traffic is generated at supper time, the virtualization control plane can transfer idle wireless resources from other communities with light load to the communities with high traffic requirements to satisfy user throughput demands. After peak time, radio resources can be returned to the resource pool.

Hierarchical Virtualization Control Scheme

Cell clustering has identified the areas with similar traffic features, which can be used as a guideline for virtualized wireless resource allocation. Different cell communities are constructed based on this guideline and then the wireless network virtualization management scheme for 5G systems can be proposed for cell communities. The basic idea of this scheme is to design a hierarchical control scheme that has a centralized control plane and self-governing cell communities.

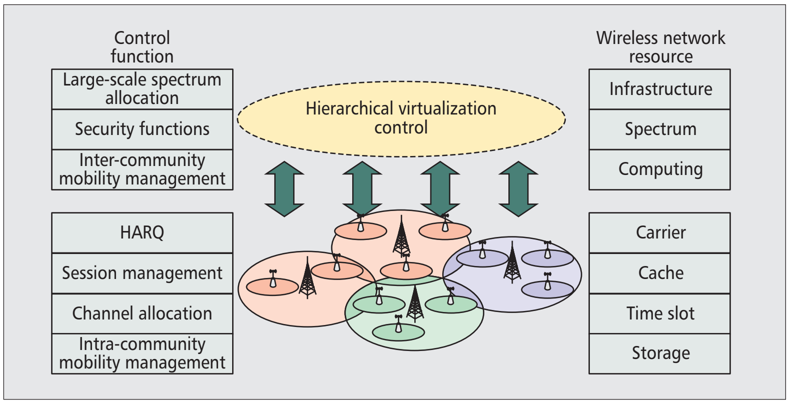

Control function
Large-scale spectrum allocation
Security functions
Inter-community mobility management
HARQ
Session management
Channel allocation
Intra-community mobility management
Hierarchical virtualization control
Wireless network resource
Infrastructure
Spectrum
Computing
Carrier
Cache
Time slot
Storage

Figure 3. Hierarchical virtualization control.

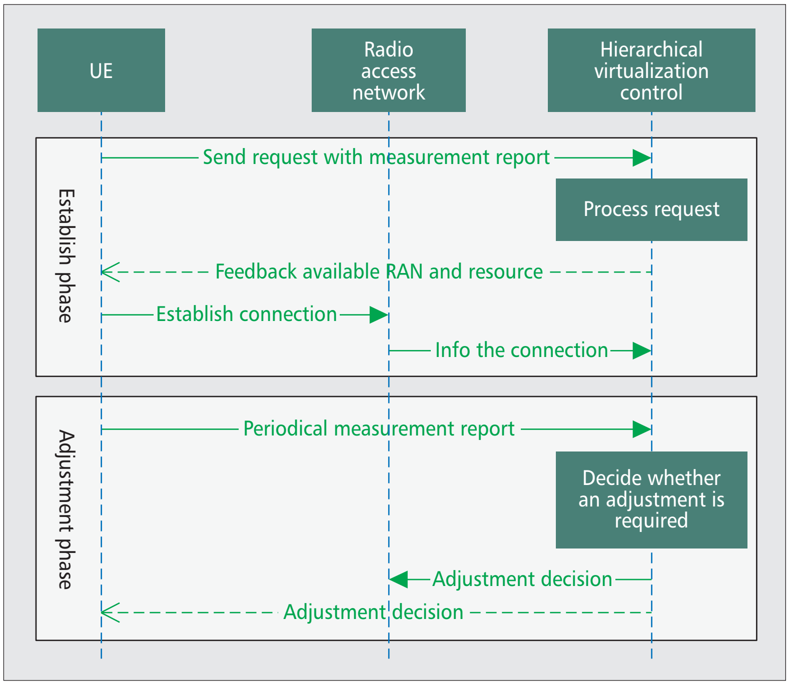

UE
Radio access network
Hierarchical virtualization control
Establish phase
Send request with measurement report
Process request
Feedback available RAN and resource
Establish connection
Info the connection
Adjustment phase
Periodical measurement report
Decide whether an adjustment is required
Adjustment decision
Adjustment decision

Figure 4. Example of establishment and adjustment signaling in the hierarchical virtualization control scheme.

Figure 3 illustrates the hierarchical virtualization control scheme, which is designed with the functions of the control plane and the cognitive plane in Fig. 1. There are three different communities as shown in Fig. 1, illustrated by different colors. The control function in a community has a certain degree of self-governing. The high-level hierarchical control functions are implemented to provide inter-community coordination among cell communities.

Small-scale intra-community control can perform real-time coordination among BSs inside one community. For instance, highly efficient resource allocation for UEs can be achieved by adapting wireless resources to the users' service characteristics. The intra-community control can take advantage of the locally centralized control and allocate all kinds of resources such as time slot, carrier, spreading code etc., while avoiding interference among UEs. High-level inter-community control can perform coordination among communities. However, such control among large scale inter-communities has drawbacks, such as large latency. This inter-community control can be used to perform large time-scaled coordination among communities, such as security functions and large-scale spectrum allocation. This scheme can achieve high resource utilization by exploiting the resource usage characteristics of different communities.

Control Signaling for Virtualization

To optimize the use of wireless resources, it is essential to have virtualization control among different RANs. Under the virtualization scheme, wireless resources are allocated dynamically. A virtualization control plane is necessary to adjust the resource allocation among RANs, and the control signaling also needs to be designed correspondingly.

In order to provide services to users, the virtualized wireless network and UEs need to be informed of the resource allocation and channel condition to establish the wireless data transmission connection. Then the connection is adjusted when the transmission environment have changed during the transmission procedure, e.g. the channel condition changes due to users' mobility. This adjustment is necessary to provide a smooth service experience to users. An example of establishment and adjustment in a hierarchical virtualization control scheme is illustrated in Fig. 4.

In the realization of virtualization control, the control signaling can be generated and transmitted in both in-band and out-of-band approaches [13]. In-band control does not need any specified resource allocation for the control channel, but may cause interference, while for out-of-band control, a dedicated band is used for control signaling transmission, thus making the operation simple and reliable. However, allocation of resources for the control channel is required, which decreases resource utilization in wireless virtualization.

Wireless Network Virtualization Use Cases

Heterogeneous wireless network deployment is one of the key features of 5G networks. To efficiently utilize the multiple resources of heterogeneous wireless networks, e.g. spectrum, infrastructure, and so on, resource sharing and virtualization management must be used. To test our proposed hierarchical virtualization control scheme, we developed the testbed to verify resource utilization improvement in spectrum-level and flow-level virtualization.

Spectrum Virtualization

Spectrum virtualization or slicing can be viewed as an extension of dynamic spectrum access or sharing. Since spectrum resources can be owned by an infrastructure provider (InP) or mobile virtual network operator (MVNO), spectrum sharing requires all the operators to contribute their licensed spectrum into a spectrum pool [8]. According to the capacity requirements of the traffic service, the spectrum is reasonably assigned to either the MVNO or service provider (SP) by the virtualization control plane. Spectrum sharing can be conducted among mobile network operators (MNOs) or MVNOs, as well as different radio access networks in terms of cognitive radio technology.

In our hierarchical virtualization control scheme, spectrum is virtualized within a community, and spectrum utilization entities (e.g. MVNOs and InPs) share the same spectrum pool. Because of the dynamic distribution of mobile users and traffic requests, the capacity density in each cell is quite different. To meet the fluctuations of traffic in each cell, dynamic spectrum sharing is adopted to efficiently allocate spectrum slices to cells carrying different traffic load.

We have developed an advanced spectrum management (ASM) demonstration to verify the spectrum efficiency improvement by virtual spectrum sharing among cells in a heterogeneous network environment [14]. The ASM server is embedded in the hierarchical virtualization control, equipped with functions of periodically estimating the traffic requirements of each cells and dynamically adjusting the spectrum slices to each cell. Spectrum allocation is based on the statistical average of the fluctuations in the cells' traffic requirements and resource utilization conditions.

Flow-Level Resource Virtualization

With the cognitive plane, it is possible to build a virtualized network by integrating different RANs. Within the unified wireless network, different RANs cooperate to provide the transmission to UEs simultaneously. Flow-level resource virtualization requires different sets of flows to be allocated to the appropriate virtual network for a successful transmission. The wide application of multi-mode terminals (MMTs) guarantees UEs' simultaneous access to different RANs, which makes it suitable to access virtualized networks for multi-flow transmission. MMTs receive the multi-flows carried on different virtual networks and combine them appropriately. In addition, flexibly adjusting the size of the flow slice according to network load conditions in different virtual networks may also efficiently utilize wireless resources. Next we discuss flow-level resource virtulization based on our research [15]. Notice that the coordination between different RANs is the extension of RAN sharing [8].

Virtualization Management for Flow Division: A major challenge faced by multi-flow transmission is how to achieve accurate synchronization. Service flows from different virtual networks must be recombined at the MMT. Therefore, the packets transmission delay of each flow should be kept low enough to guarantee the QoS requirements. To realize multi-flow transmission, the effective scheduling scheme needs to be adopted and the resource virtualization acts as a proper measure. Traffic flows can be divided into packet blocks in the virtualization control plane, which are labelled with the virtual resources being carried and packet blocks sequence. Without knowing which virtual resources are allocated and how the virtual networks transmit, UEs equipped with MMTs reconstruct these packet blocks according to the sequence label and merge them into the original service flow.

A flow scheduling server is in charge of dividing the service flow into multiple dedicated slices for different virtual networks to carry, according to the virtual resource quality and traffic load of the virtual networks. The functions of the flow scheduling server are embedded in the hierarchical virtualization control, owned by a MVNO or a SP. Hierarchical virtualization control can judge the access network resource conditions and reasonably adjust its flow slice proportion based on the feedback information by the MMT. Then the virtualized access network resources are allocated to users, thus improving the performance of the multi-flow transmission system while increasing throughput.

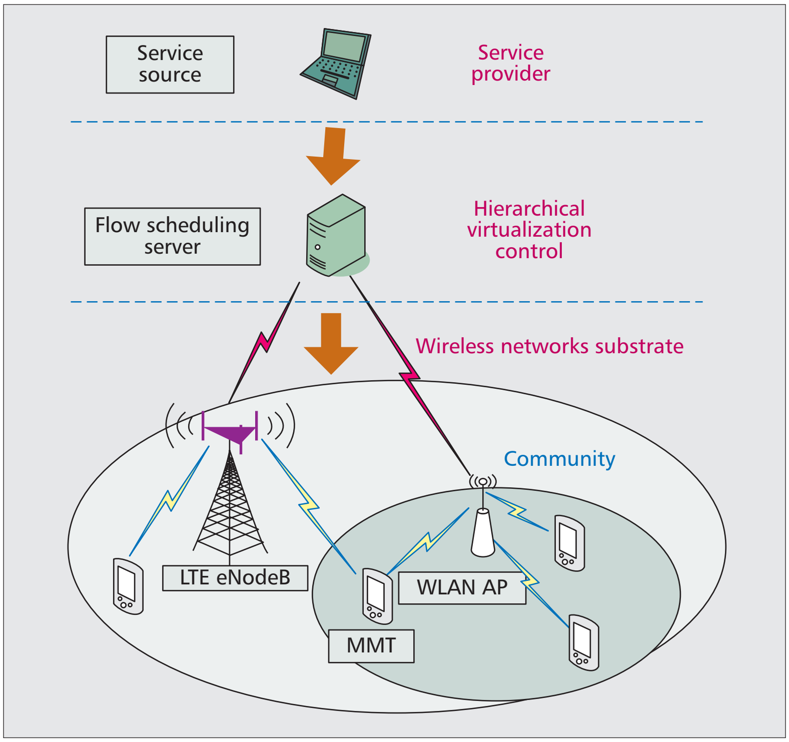


Figure 5. Flow-level virtualization architecture in hierarchical virtualization control scheme.

Multi-Flow Transmission: We have already implemented the LTE and WLAN multi-flow transmission in the laboratory testbed as shown in Fig. 5. The multi-flow transmission system is composed of service resources, multi-mode mobile terminal, and flow scheduling server. As shown in Fig. 5, LTE and WLAN access networks are virtualized as a single virtual network in charge of traffic flow delivery. The flow scheduling server acts as the virtualized control entity to collocate all the available RANs. After determining the proper proportion for two flows, it allocates certain RANs to carry the traffic and transmit to the mobile terminals. MMTs reconstruct the flow to form an integrated flow and restore the original video The entire process runs automatically without intervention from users.

Evaluation Results

Spectrum Virtualization: In the spectrum virtualization testbed, we deployed two cells, each having four UEs. The ASM server obtains the available spectrum resource information from the cognitive plane and updates the latest spec-

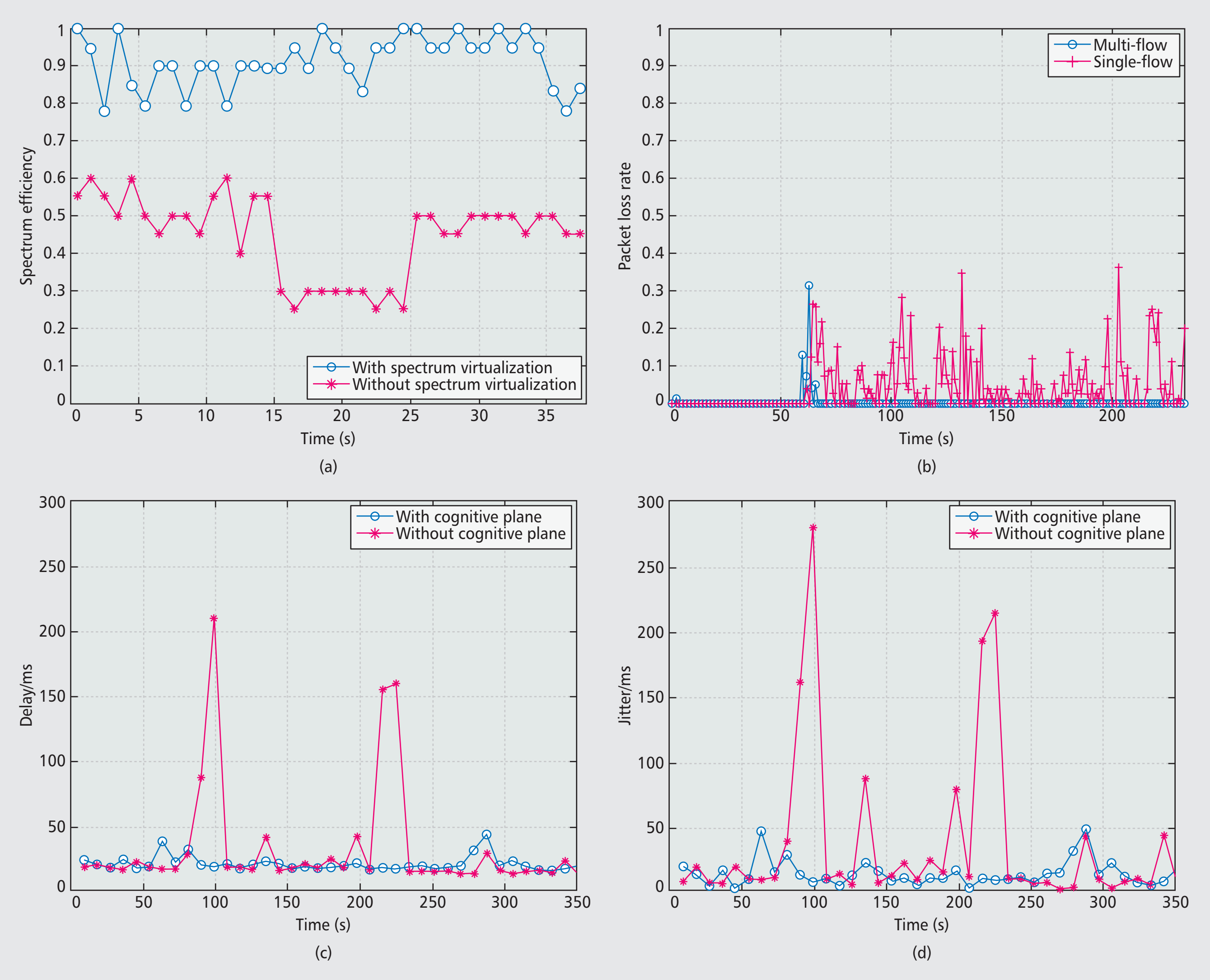


Figure 6. Evaluation results: a) spectrum virtualization; b) packet loss rate; c) multi-flow delay; d) multi-flow jitter.

trum usage information from the eNB. Then the ASM is responsible for spectrum reallocation and coordination among different eNBs. As shown in Fig. 6a, after spectrum virtualization, the total spectral utilization percentage has increased by nearly 30 percent compared to the case without spectrum virtualization, which demonstrates the flexibility of spectrum virtualization and the efficient utilization of spectrum.

Flow-Level Virtualization: To test the reliability of the multi-flow transmission, online video playback is chosen as an example of a high throughput real-time application. Since the quality of video is very sensitive to packet delay and jitter, those two parameters are chosen to illustrate the quality of video playback. We build a heterogeneous network with LTE and WLAN, where the throughput of LTE and WLAN is about 600 kb/s and 800 kb/s, respectively. During the experiment, two test users play back a video with code rate between 300 kb/s~60 kb/s and 600 kb/s~100 kb/s.

As illustrated in Fig. 6b, the multi-flow transmission has a lower package loss rate (PLR) than the single-flow transmission. This demostrates that multi-flow transmission has a much better QoS by combining the capacity of multiple RANs. Since multi-flow transmission can deliver data packets to UEs in time, a smaller number of packets are dropped by the eNB or AP due to imperfect channel conditions and the overflow of embedded cache.

Our results also proved that multi-flow transmission with the cognitive plane works better. When the total throughput required from users exceeds the capacity of the LTE or WLAN network, the MMT offered smoother playback by providing the transmission service through different RANs simultaneously. The cognitive plane is a key enabler for smooth transmission. As shown in Fig. 6c, the delay with the cognitive plane is lower and more steady than the one without the cognitive plane, because the cognitive plane acts as a self-adaptive controller, which can handle feedback information and allocate reasonable bandwidth to adapt the dynamic wireless links. A similar observation can be found in the jitter comparison shown in Fig. 6d.

Conclusion

In this article we proposed a general model of wireless network virtualization consisting of the data plane, the cognitive plane, and the control plane, which can dynamically and flexibly slice the infrastructure and resources into virtual networks to achieve global optimized resource utilization. In addition, we proposed a hierarchical control scheme based on cell clustering to improve the management efficiency of wireless network virtualization. The control signaling for this scheme is also designed. Finally, by use case analysis we have shown that the proposed scheme can be used to improve resource utilization. Spectrum virtualization and flow level virtualization have proved to be efficient virtualization methods for futute 5G networks, where the convergence of heterogeneous RANs is already a vital issue. As for the cost, our proposed wireless network virtualization model can be implemented in a software platform, thus the investment of infrastructure is protected. We believe our proposed wireless network virtualization method is one of the effective approaches to realize future 5G networks.

> Our proposed wireless network virtualization model can be implemented in a software platform, thus the investment of infrastructure is protected. We believe our proposed wireless network virtualization method is one of the effective approaches to realize future 5G networks.

References

[1] Qualcomm, “The 1000x Data Challenge,” 2013, available:https://www.qualcomm.com/invention/technologies/1000x.

[2] K. Zheng ***et al.***, “10 Gb/s Hetsnets with Millimeter-Wave Communications: Access and Networking — Challenges and Protocols,” *IEEE Commun. Mag.*, vol. 53, no. 1, 2015, pp. 222–31.

[3] F. Boccardi ***et al.***, “Five Disruptive Technology Directions for 5G,” *IEEE Commun. Mag.*, vol. 52, no. 2, 2014, pp. 74–80.

[4] P. Demestichas ***et al.***, “5G on the Horizon: Key Challenges for the Radio-Access Network,” *IEEE Vehic. Tech. Mag.*, vol. 8, no. 3, 2013, pp. 47–53.

[5] C. Liang and F. Yu, “Wireless Network Virtualization: A Survey, Some Research Issues and Challenges,” *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 1, 2015, pp. 358–80.

[6] S. Sezer ***et al.***, “Are We Ready for SDN? Implementation Challenges for Software-Defined Networks,” *IEEE Commun. Mag.*, vol. 51, no. 7, 2013, pp. 36–43.

[7] 3GPP, “Technical Specification Group Services and System Aspects; Network Sharing; Architecture and Functional Description,” 3GPP, TS 23.251, Mar. 2015.

[8] 3GPP, “Technical Specification Group Services and System Aspects; Service Aspects and Requirements for Network Sharing,” 3GPP, TR 22.951, Oct. 2014.

[9] X. Costa-Perez ***et al.***, “Radio Access Network Virtualization for Future Mobile Carrier Networks,” *IEEE Commun. Mag.*, vol. 51, no. 7, 2013, pp. 27–35.

[10] W. Xu ***et al.***, “Cognition Flow in Cognitive Radio Networks,” *China Commun.*, vol. 10, no. 10, pp. 74–90, 2013.

[11] C.-L. I ***et al.***, “Toward Green and Soft: A 5G Perspective,” *IEEE Commun. Mag.*, vol. 52, no. 2, 2014, pp. 66–73.

[12] D. Wubben ***et al.***, “Benefits and Impact of Cloud Computing on 5G Signal Processing: Flexible Centralization Through Cloud-RAN,” *IEEE Signal Proc. Mag.*, vol. 31, no. 6, 2014, pp. 35–44.

[13] Q. Zhang, Z. Feng, and G. Zhang, “A Novel Homogeneous Mesh Grouping Scheme for Broadcast Cognitive Pilot Channel in Cognitive Wireless Networks,” *IEEE Int’l. Conf. Commun. (ICC)*, 2010, pp. 1–6.

[14] P. Zhang ***et al.***, “Intelligent and Efficient Development of Wireless Networks: A Review of Cognitive Radio Networks,” *Chinese Science Bulletin*, vol. 57, no. 28–29, 2012, pp. 3662–76.

[15] H. Lian ***et al.***, “Match-Degree Based Bandwidth Allocation Scheme in Heterogeneous Networks,” *IEEE Int’l. Conf. Commun. (ICC)*, 2014, pp. 1242–47.

Biographies

Zhiyong Feng (fengzy@bupt.edu.cn) received B.S., M.S., and Ph.D. degrees from Beijing University of Posts and Telecommunications (BUPT), China. She is a professor at BUPT, and the director of the Key Laboratory of Universal Wireless Communications, Ministry of Education, P. R. China. She is a senior member of IEEE and active in standards development such as ITU-R WP5A/WP5D, IEEE 1900, ETSI and CCSA. Her main research interests include wireless network architecture design and radio resource management in 5th generation mobile networks (5G), spectrum sensing and dynamic spectrum management in cognitive wireless networks, universal signal detection and identification, network information theory, etc.

Chen Qiu (jp092983@bupt.edu.cn) received B.S. degrees from both Beijing University of Posts and Telecommunications (BUPT) and Queen Mary University of London (QMUL) in 2013. He is currently a Ph.D. student at BUPT. His research interests are mobile data analysis and 5G wireless network.

Zebing Feng (fengzebing@bupt.edu.cn) received a B.S. degree from Nanjing University of Science and Technology (NJUST) in 2011. He is currently studying toward the Ph.D. degree in communication and information systems at the Key Laboratory of Universal Wireless Communications Ministry of Education of Beijing University of Posts and Telecommunications (BUPT). His research interests include the convergence of heterogeneous wireless networks, dynamic spectrum management, machine type communications, and cognitive radio technology.

Zhiqing Wei (weizhiqing@bupt.edu.cn) received B.S. and Ph.D. degrees from Beijing University of Posts and Telecommunications (BUPT) in 2010 and 2015, respectively. He is a lecturer at BUPT. His research interests are capacity and delay analysis of cognitive radio networks and 5G wireless networks.

Wei Li (weili@ece.uvic.ca) received the Ph.D. degree in electrical engineering from the University of Victoria, Canada in 2004. In 2005 he joined the France Telecom San Francisco Lab as a research scientist, where he was a member of international standard bodies for wireless networks. He is currently an adjunct professor at the University of Victoria, Canada. His research interests are in information theory, wireless networks, smart grid, and heterogeneous networks.

Ping Zhang (pzhang@bupt.edu.cn) received the M.S. degree in electrical engineering from Northwestern Polytechnical University, Xi’an, China, in 1986, and the Ph.D. degree in electric circuits and systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1990. He is currently a professor at BUPT and the director of the State Key Laboratory of Networking and Switching Technology, China. His research interests include cognitive wireless networks, 4th generation mobile communication (4G), 5th generation mobile networks (5G), communications factory test instruments, universal wireless signal detection instruments, mobile Internet, etc. He is the executive associate editor-in-chief on information sciences of the *Chinese Science Bulletin*, a guest editor of *IEEE Wireless Communications Magazine*, and an editor of *China Communications*. He was a recipient of the First and Second Prizes of the National Technology Invention and Technological Progress Awards, as well as the First Prize of the Outstanding Achievement Award of Scientific Research in College.

Cost Analysis of Initial Deployment Strategies for Virtualized Mobile Core Network Functions

The authors analyze the cost incurred by two "constraint-based heuristically applied" initial VNF/VNFC deployment strategies with reference to a virtualized mobile network infrastructure providing EPCaaS (evolved packet core as a service) while taking into consideration functional and administrative constraints. The cost of deployment is measured in terms of the utilization of data center infrastructure resources such as compute and networking.

Faqir Zarrar Yousaf, Paulo Loureiro, Frank Zdarsky, Tarik Taleb, and Marco Liebsch

Faqir Zarrar Yousaf, Paulo Loureiro, and Marco Liebsch are with NEC Laboratories Europe.

Frank Zdarsky is with Red Hat GmbH.

Tarik Taleb is with Aalto University.

Abstract

A virtualized network infrastructure is composed of multiple virtualized network functions (VNF) interconnected by well defined interfaces, and thus forming a VNF-graph. The initial deployment of such a VNF-graph inside a data center (DC) is a complex task with multidimensional aspects compared to deploying a single VNF that may represent a single network function. The problem space becomes more complex when each VNF is further decomposed into multiple VNF components (VNFC), where each VNFC embodies a subset of network functions attributed to the respective VNF. The challenge is to ensure that the deployment strategy meets the intra-functional constraints between the multiple VNFCs constituting the VNF-graph while ensuring service, performance and operational integrity, and also ensures optimal utilization of the underlying resources of the DC infrastructure (DCI).

In this article we analyze the cost incurred by two "constraint-based heuristically applied" initial VNF/VNFC deployment strategies with reference to a virtualized mobile network infrastructure providing EPCaaS (evolved packet core as a service) while taking into consideration functional and administrative constraints. The cost of deployment is measured in terms of the utilization of DC infrastructure resources such as compute and networking. We also present the discussion in view of the ETSI NFV MANO framework, undergoing standardization, that is responsible for management and orchestration of NFV systems including VNF deployment.

COMMUNICATIONS STANDARDS

Introduction

Background

Network function virtualization (NFV) is fast emerging as a promising technology that leverages the concept of cloud technology and virtualization techniques into the realm of telecommunication networks. Mobile network operators are especially interested in exploring the potential of adopting this technology to enhance their competitiveness and reduce capital and operational costs. At present the network function entities are developed and ported on customized hardware platforms designed and tested for meeting the functional and operational requirements for a specific function or set of functions. Such a rigid infrastructure makes network scalability difficult and expensive and increases the total cost of ownership (TCO). Additionally it also locks the operator into specific hardware and/or software vendors, while constraining the operator from rolling out new services as per demand, thereby impacting revenue.

NFV technology has the potential of offsetting the above issues while providing a highly scalable, flexible, and elastic network infrastructure. NFV involves the virtualization of network node functions and hosting these virtualized network functions (VNF) on virtual machines (VM), which in turn are deployed on commodity servers (i.e. COTS servers). These VNFs are then interconnected across the servers to provide the intended network services (NS). For example a mobile core network, such as evolved packet core (EPC), is composed of several functional entities interconnected via standardized interfaces. In order to virtualize an EPC for providing EPC as a service (EPCaaS), the functional elements are characterized by multiple VNFs, where the respective VNFs are interconnected via well-defined interfaces forming a VNF-graph. For truly resilient and elastic performance, the VNFs can also be decomposed into multiple inter-connected VNF components (VNFC), where each VNFC instance is hosted on a single VM.

ETSI NFV MANO System

The inherent advantages offered by NFV introduces the challenge of the management and orchestration of the multitude of distributed VNF(C)s deployed across multiple servers in a network functions virtualized infrastructure (NFVI), for example a data center (DC), to provide carrier-grade service. To this effect an ETSI ISG has been formed to standardize the various aspects of NFV enabled networks, including the NFV management and orchestration (MANO) framework [1]. The proposed MANO framework architecture is shown in Fig. 1, which is composed mainly of three functional blocks, namely the virtualized infrastructure manager (VIM), the VNF manager (VNFM), and the NFV orchestrator (NFVO), interconnected over specific reference points. There are additional data repositories that may contain necessary information about NS, VNF, NFV, and NFVI that will enable the NFVO to perform its tasks. The MANO architecture also defines reference points for interfacing the MANO system with external entities like NFVI, OSS/BSS, VNFs, and element managers (EM) for delivering unified management and orchestration of a VNF system.

An interfaces and architecture (IFA) WG has been formed under the ETSI NFV that has the mandate to develop specifications for the MANO framework. In this respect, the IFA WG at present is in the process of specifying interfaces, requirements, and operations for the reference points in view of the functional/operational scope of the NFVO, NFVI, and VIM, as described in [1]. Besides traditional FCAPS management, the MANO framework focuses on newer management aspects introduced by NFV, such as the creation and life-cycle management (LCM) of the virtualized resources for the VNF, and collectively referred to as VNF management [1]. There are several VNF management tasks such as VNF scaling, migrating, and updating, to name a few, but the deployment/instantiation of VNF(C)s in a DC (i.e. in the NFVI) is the main focus of this article.

It is a challenging task to initially deploy VNF(C)s on a NFVI, owing to the intra-functional dependencies and constraints among the various VNF(C)s. Thus, during deployment the MANO system must take into consideration the intricate (anti)affinity between the various VNF(C)s that constitute a complex NS such as a virtualized evolved packet core (vEPC). Following are the main aspects that the MANO system must take into consideration when forming a VNF-graph and making deployment decisions when realizing complex NS such as the vEPC:

- Networks by themselves offer a complex, well connected, and well defined ecosystems, composed of multiple complex, yet well defined and well specified functions and with strict relationships.
- The network functions are interconnected to each other via well-defined interfaces and communicate with each other using well-defined and specified protocols.
- These network functions work in a coordinated manner to ensure end-to-end service integrity and connectivity.
- Each network function has a different set of system and resource requirements.
- Each network function has a well-defined functional scope of operation as stipulated by the relevant standards.
- Achieving carrier-grade performance from the deployed VNF-graph is still the number one priority for many mobile operators.

In this article we address and analyze the issue of initial deployment of a virtual mobile network platform, represented by a VNF-graph, within a DCI (i.e. NFVI). For our analysis we have adopted a simplistic architecture of a vEPC network [2] as a reference for our analysis. The main objective and motivation behind this article is to compare two constraint-based heuristic approaches of initial deployment of vEPC VNFCs over a DCI and analyze the impact of the two deployment strategies on the cost of deployment.

The rest of the article is organized as follows. The next section provides some related research work, by which we will provide a conceptual and functional overview of EPCaaS with reference to the vEPC network. This is followed by a description of the evaluation framework and method that includes the modeling of the DC and vEPC network. We describe the two proposed deployment strategies, and we present performance analysis. The article then concludes.

Related Work

Several pioneering research works have been conducted to enable the creation and runtime management of mobile networks over the cloud, studying different implementation options [3] and devising an entire framework for the creation of end-to-end mobile services, including mobile transport networks, on the cloud [4]. For a successful creation of mobile core networks on the cloud, algorithms for optimal placements of VNFs on a federated cloud and within the same DC are of crucial importance.

In traditional mobile core networks, mechanisms and algorithms have been devised to select, for mobile users, optimal data anchor gateways from within a fixed range of geographically static gateways for the sake of communication efficiency.

However, in cloud-based mobile core network, gateways are realized as VNFs, which are not only created on-demand, but operators have more flexibility in deciding where to place VNFs of gateways, rather than just selecting gateways from within a fixed set of static gateways. Such flexibility helps mobile operators to dynamically dimension, plan, and re-plan their mobile networks whenever there is a need for that and as per the changing behavior of mobile users, the features of the provisioned services, and according to other metrics relevant to the mobile network performance. Regarding the latter, the authors in [5] proposed a VNF placement method, particularly for creating mobile gateway functionalities (serving gateway (S-GW)) and their placement in federated clouds so that the frequency of S-GW relocation occurrences is minimized. In [5], the aim was to conduct an efficient planning of service areas (SAs) retrieving a trade-off between minimizing the user equipment (UE) handoff between SAs, and minimizing the number of created instances of the virtual S-GWs. In [6] the focus was on VNF placement and instantiation of another mobile network functionality, namely data anchoring or PDN-GW (P-GW) creation/selection. That work argued the need for adopting application type and service requirements as metrics for creating VNF instances of PDN-GW and selecting adequate virtual P-GWs for UEs receiving specific application

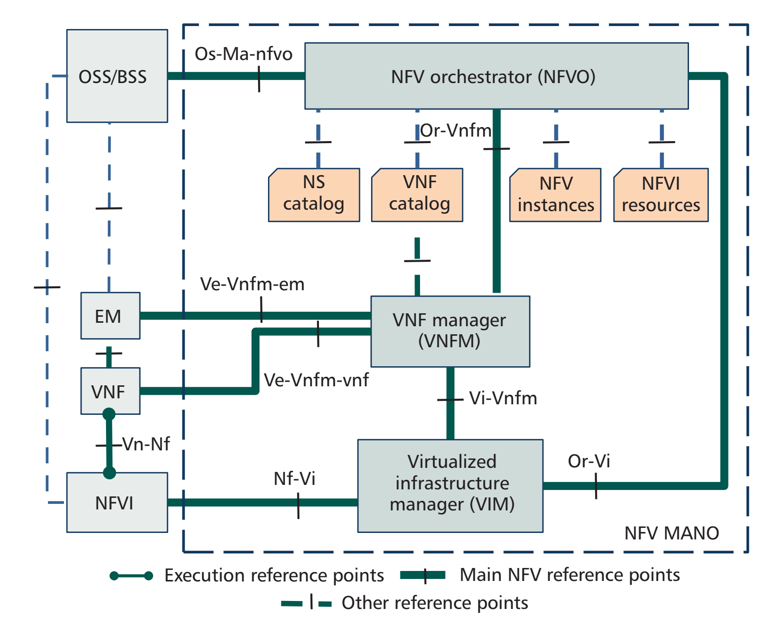


Figure 1. ETSI NFV management and orchestration (MANO) framework overview [1].

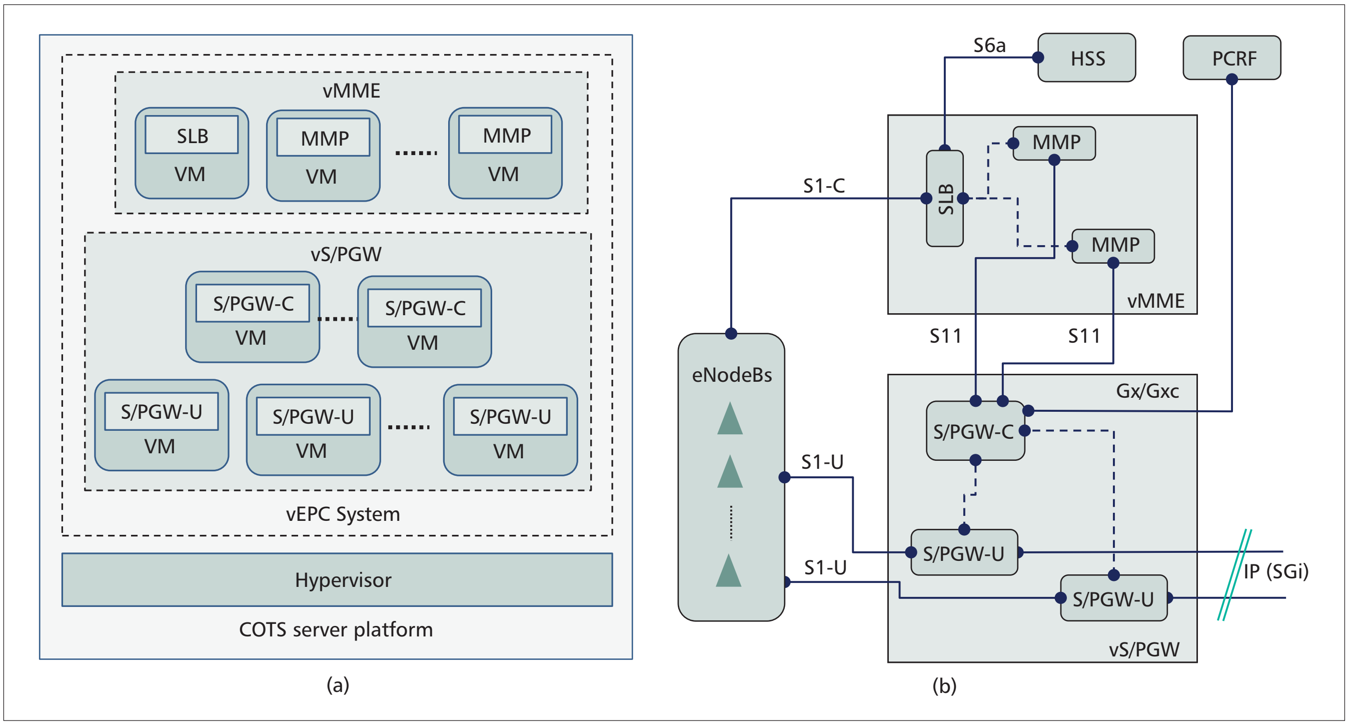


Figure 2. Virtualized evolved packet core (vEPC) system overview: a) functional overview; b) interfaces between vEPCs VNFCs.

types. The placement of P-GW VNFs was modeled through a nonlinear optimization problem whose solution is NP-hard. Three heuristics were then proposed to deal with this limitation. In [7] the authors proposed a softEPC framework for flexible and dynamic instantiation of EPC VNFs with reference to the actual traffic demand at appropriate topological locations.

While the above research works considered the problem of VNF placement across federated clouds, the present article will be looking into the VNF placement problem within the same DC. In this context research work has been conducted for decisions on VM placement within the same DC, having as the objective cost savings thanks to better utilization of computing resources and less frequent overload situations. In [8] performance isolation (e.g. CPU, memory, storage, and network bandwidth), resource contention properties (among VMs on the same physical host), and VMs' behavioral usage patterns are taken into account in decisions on VM placement, VM migration, and cloud resource allocations.

In other research works, optimal placement of VMs takes into consideration electricity-related costs as well as transient cooling effects [9]. Others do autonomic placement of VMs as per policies specified by the data center providers and/or users [10]. Other VM placement strategies consider maximizing the profit under a particular service level agreement and a predetermined power budget [11].

In [12] the authors take into consideration the reduction in control-signaling traffic and congestion in the data plane of a vEPC system. However, the authors in [12] take a different view, where instead of focusing on the placement of individual VMs, they propose to group multiple vEPC functions, for example SGW and PGW, in one VM, and then interconnecting the different VM segments via GTP to achieve the desired objective.

Thus instead of taking a one dimensional view of VM placement and focusing on the single optimization factor, the VNF placement problem, addressed in this article, is more complex. This is so because a virtualized network infrastructure is composed of multiple VNFs, which are interconnected by well-defined interfaces, thereby forming a VNF-graph. The problem space becomes more complex when a single VNF gets further decomposed into multiple interconnected VNFCs, and there is a strict functional relationship between the various VNFCs and performance constraints that makes the deployment process more complex.

Unfortunately, there is not much information available that may analyze the impact of deployment strategy during the initial deployment of a virtualized network infrastructure (represented by a VNF-graph) in a DC. In this regard, this article analyzes the impact on the cost of DC resources, such as networking and computing, by comparing the impact of two "constraint-based and heuristically-derived" deployment strategies, namely vertical serial deployment (VSD) and horizontal serial deployment (HSD) strategies adopted for the deployment of a virtualized mobile core network, referred to as a vEPC.

EPCaaS: Conceptual and Functional Overview

The objective of EPCaaS is to virtualize the EPC infrastructure in order to extend the advantages of the cloud system to mobile network operators.

This is done by instantiating the EPC system's functional entities (e.g. mobility management entity (MME), S-GW, P-GW) as VNFs on VMs over COTS servers instead of a specialized mission-specific, custom tuned, expensive hardware platform. To provide the service and design concept of EPCaaS we use a simplistic architecture of a vEPC network [2] as a reference use case. The overview of this architecture is depicted in Fig. 2a, where the MME and S/P-GW VNFs are referred to as vMME and vS/P-GW, respectively.

The following four possible architectural reference models for EPCaaS have been specified in [3, 13] based on how the VNFs are mapped on the VMs:

- 1:1 mapping, where each EPC VNF is implemented on a separate VM.
- 1:N mapping, where each EPC VNF is decomposed into sub-functional entities (i.e. VNFC) and each VNFC is implemented on a separate VM.
- N:1 mapping, where the complete EPC system is implemented on a single VM.
- N:2 mapping, which is similar to N:1 except that it separates the control plane (CP), user plane (UP), and database services of the EPC onto three separate interconnected VMs.

The vEPC system falls in the category of 1:N mapping, where the respective VNFs of the vMME and vS/PGW functions are decomposed into separate CP and UP VNFCs in order to render enhanced agility and elasticity in view of different traffic and application types. Thus the vS/PGW is divided into two VNFCs, namely vS/PGW-C and vS/PGW-U, with the former VNFC processing the CP load and the latter processing the UP load. Similarly, the vMME functionality is embedded in the combination of signaling load balancer (SLB) and mobility management processor (MMP) VNFCs, where the MMP performs the processing task of the MME. The combination of SLB and MMP will allow the scaling of vMME by using SLB and by adding/deleting MMPs. Each functional entity (i.e. SLB, MMP, vS/PGW-C, and vS/PGW-U) is realized on a separate VM, and the inter-connectivity between these VNFCs is based on standardized interfaces (Fig. 2b).

Evaluation Framework and Methodology

For cost analysis, we have developed an evaluation framework in C++. The evaluation framework has been designed with reference to the functional requirements of the MANO system [1]. This framework is composed of a DCI model, a vEPC system model, and a deployment model. For a specific CP/UP input load, the vEPC system model determines the required number of VNFCs and their respective resource requirements that will support the incident traffic load. The deployment model, based on a specific deployment strategy, will then deploy the respective vEPC's VNF-graph, including the respective VNFCs, on the servers of the underlying DCI model while taking into account the resource requirements of individual VNFCs and the vEPC system internal bounds and constraints. The framework then computes and determines the cost incurred by the respective deployment strategy in terms of DC networking and computes resource consumption for the incident CP/UP traffic load. Our evaluation framework can be scaled to any size DC and to any size vEPC system, depending on the load on the operator's network. The overview of the DCI model, the vEPC system model, and the deployment strategies, are discussed in the following sub-sections.

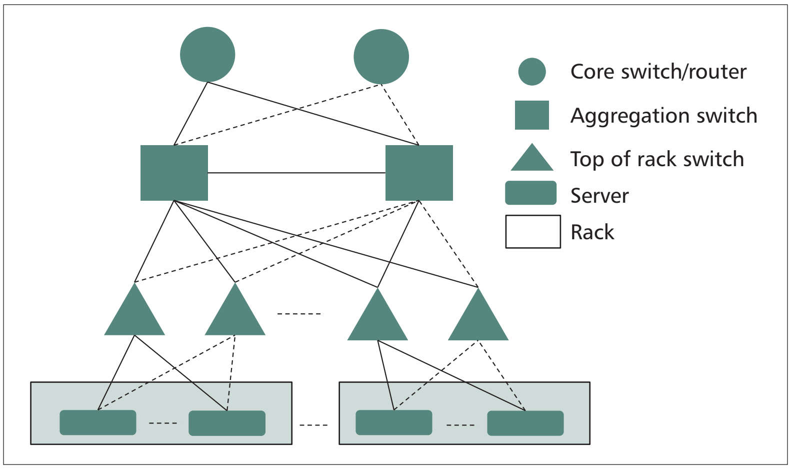


Figure 3. Three-layer date center infrastructure model.

Data-Center Infrastructure Model

For the analysis, we have modeled the traditional hierarchical three-tier DC architecture composed of:

- The core layer
- The aggregation layer
- The access layer [14]

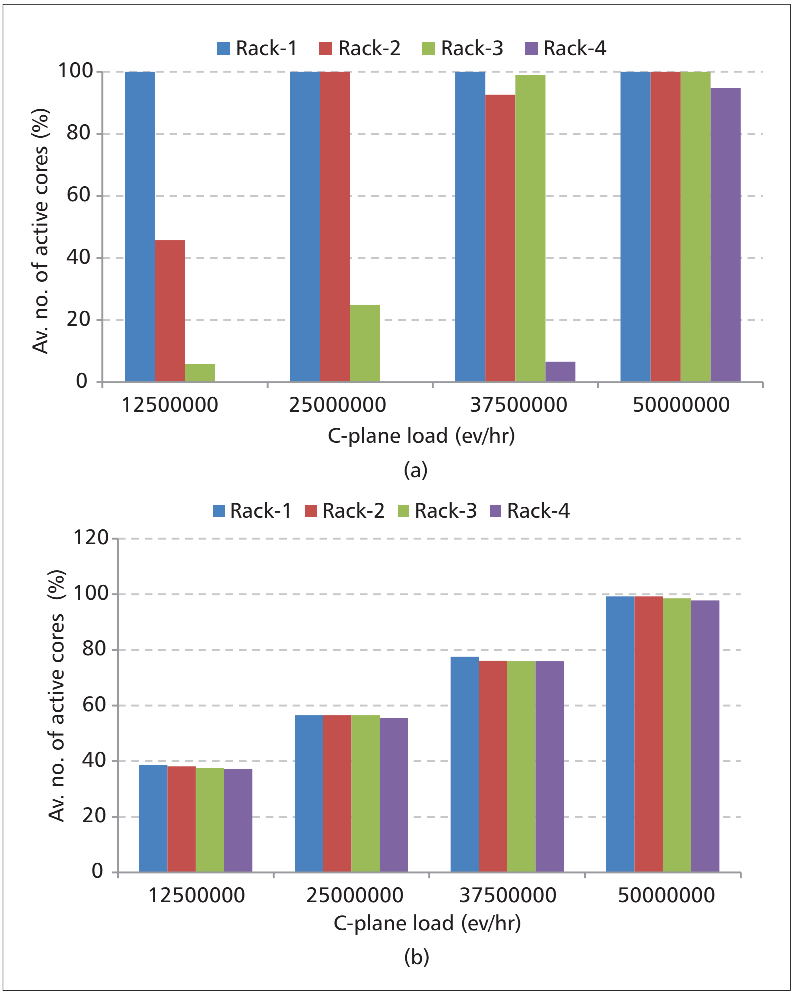
At the lowest level is the access layer, which contains pools of servers housed in racks, where each server is connected to one (or two for redundancy) top-of-rack (TOR) L2 switch. Each TOR switch is, in turn, connected to one (or two for redundancy) high capacity L2/L3 switch at the aggregation layer. The aggregation switches are then connected to the top-level switches/routers, forming the core layer. Such a fat-tree topology can be scaled up, in turn, by scaling up each individual switch. Figure 3 illustrates the DCI topology that we have modeled, where the dotted lines indicate redundant links, thereby connected to the redundant/backup node. For our analysis, we do not consider failure scenarios and hence the redundant links/nodes are not utilized. The access layer is modeled as an $m \times n$ matrix where m is the number of racks and n is the number of servers per rack. For our analysis, we consider a homogenous access system where all racks are of the same size and all servers are of the same configuration and form-factor. The servers are modeled having x number of CPU cores and *xGbps* aggregate network bandwidth. On the other hand, the switches/routers are modeled considering *xGbps* aggregate bandwidths.

vEPC System Model

The vEPC system is modeled by characterizing the individual VNFCs (SLB, MMP, S/PGW-C, and S/PGW-U) in terms of the CP/UP load that they process. The model also captures the interfaces between the different relevant VNFCs, as depicted in Fig. 2b. Figure 2b illustrates the interconnected VNFCs constituting the vEPC system with relevant interfaces. The model is able to determine

Parameter	Notation	Value
Total CPU cores per VNFC	N_{core}^{VNFC}	4
eNB per SLB	N_{eNB}^{SLBi}	100
Number of S/PGW-U per SPGW-C	N_{SPGWU}^{SPGWCi}	6
Maximum S1C load per MMP	$L_{S1C,max}^{MMPi}$	500,000 ev/hr
Maximum S11 load per SPGW-C	$L_{S11,max}^{SPGWCi}$	1,000,000 ev/hr
Maximum S11 load per SPGW-U	$L_{S11,max}^{SPGWUi}$	166666.7 ev/hr
CP load demand (ev/hr)	C_{cp}	$x * L_{S1C,max}^{MMPi}$ where x = 0.25, 0.50, 0.75, 1.0
Number of eNBs, where each value corresponds to the respective value of C_{cp}	N_{eNB}	[1500, 2000, 2500, 3000]
Average CP packet size (in bytes)		192
Average number of messages per CP event		6
UP load demand (Gb/s)	C_{up}	64, 128, 256, 512
UP packet size (in bytes)		512

Table 1. Simulation parameters.



(a)

(b)

Figure 4. Average number of active cores per rack: a) VSD; b) HSD.

not only the number of relevant VNFCs required to handle a particular CP/UP load profile, but also determines the resource requirements of individual VNFCs in terms of CPU cores and network bandwidth. This information is then used to analyze the deployment cost of the vEPC system in a DCI, thereby enabling the operators to dimension the resources of their respective DCI for specific load conditions and service requirements.

This model is expected to provide insight into the resource requirement of every VNFC and thus the size of the overall vEPC system, in response to external inputs. The load models for vMME and vSPGW are derived with reference to Fig. 2b, and summarized below.

Load Model for vMME: As stated earlier, the vMME is composed of the SLB and MMP virtual instances. The load from the eNBs is balanced by SLB among multiple MMP instances. We assume SLB to be balancing the load between MMPs in an equally weighted round-robin manner. The total S1C load on a single SLB instance from eNBs is the aggregate S1C loads from all eNBs associated with the SLB.

Thus, the load on a single MMP instance is the total S1C load on a single SLB instance divided by the number of MMPs that a single SLB can serve. The ratio between the number of eNBs per SLB and the number of MMPs per SLB depends on the load balancing capability of the SLB as well as the maximum load that an MMP instance can handle.

Load Model for vS/PGW: The vS/PGW is modeled by characterizing the S/PGW-C and S/PGW-U VNFCs in terms of the CP and UP load that the respective VNFCs process. With reference to Fig. 2b, the total CP load incident on a single S/PGW-C instance is the sum of the CP loads from the policy and charging rules function (PCRF) and a proportion of the total S11 load from the associated MMPs.

Similarly, the total load processed by a single S/PGW-U instance is the proportion of the S11 load (i.e. CP load) from the S/PGW-C and the S1U load (i.e. UP load) from the associated eNBs (Fig. 2b).

Deployment Strategies

Following are the two constraint-based and heuristically derived deployment strategies:

- Vertical serial deployment (VSD) strategy
- Horizontal serial deployment (HSD) strategy

Both strategies deploy VNFCs serially such that the vMME VNFCs (i.e. SLBs and MMPs) are deployed first, followed by the vS/PGW VNFCs (i.e. S/PGW-C and S/PGW-U). In VSD, VNFCs are deployed from top to bottom on servers of one rack, and when no more resources are available in the rack, VNFCs are deployed on to the servers in the next rack. In HSD, VNFCs are deployed on the first available server in a rack, then moving on to the next available server on the next rack, and so on until all VNFCs are deployed. In other words, considering the access layer as an $m \times n$ matrix, in VSD, VNFCs are deployed column-wise, whereas in HSD the VNFCs are deployed row-wise.

While deploying, the VSD/HSD deployment strategy will take into account the (anti)affinity between respective VNFCs, system reliability, server resources in terms of available CPU cores, and the network resources such as the capacity of the network interfaces on the servers and of the links in DCI. The following constraints are also taken into consideration during deployment:

- For reliability, a single server may not have more than one instance of the S/PGW-U belonging to the same logical vS/PGW.
- Each time a VNFC is instantiated, the associated standby VNFC will also be instantiated.
- A single server shall not host the active and standby instances of a particular VNFC.
- A VNFC is deployed only if the server has the CPU cores required by the target VNFC.

In both VSD and HSD, any server that may not have the resources required for a particular VNFC or does not offer affinity with any of the previously installed VNFCs is skipped over. For our analysis, and for the sake of simplicity, we assume that the servers are all dedicated for vEPC system deployment and no other third party services are running on them.

Performance Evaluation

In order to compare and analyze the cost impact of the VSD and HSD deployment strategies on the DCI computing and networking resources, we perform experiments on our evaluation framework using CP load (C_{cp}) and UP load (C_{up}) values based on conservative estimates during a busy hour period. According to [15], a MME can experience a sustained signaling load of 500 to 800 messages per UE during busy hours, and up to 1500 messages per UE per hour under adverse conditions. Furthermore, according to [16] the chattiest applications can generate up to 2400 signaling events per hour. Based on these observations, we assume 90 users per eNB that generate the bulk of traffic events. For our scenario, we assume 5 percent of users generating 2400 events/hr, 25 percent producing 800 events/hr, and 70 percent producing 500 events/hr during busy hours. Thus during busy hours a vEPC system will encounter 60300 events/hr from a single eNB. Based on the incident load, the vEPC system model with the help of equations 1-6 will compute the required number of VNFCs. These VNFCs are then deployed by the respective deployment strategy (i.e. HSD and VSD) on the DCI model in view of the constraints and affinity between the relevant VNFCs. The access layer of the DCI is modeled as a 4 × 45 matrix, and all of the 180 servers have 16 cores each.

For simplicity, we assume all VNFCs are assigned four CPU cores during deployment. Our evaluation framework also provides the standby VNFCs based on 1+N redundancy, but as we are not considering a failure scenario, we will not consider the standby VNFCs and corresponding links for throughput calculations. We also ignore the L_{PCRF} during calculations. The rest of the parameters used in our simulation framework are listed in Table 1, which are derived from equations 1-6 while based on assumptions described above.

The performance of two deployment strategies (i.e. VSD and HSD) are measured with respect to the average number of active cores utilized per rack (Fig. 4) and the average throughput per rack (Fig. 5) for four C_{cp} values.

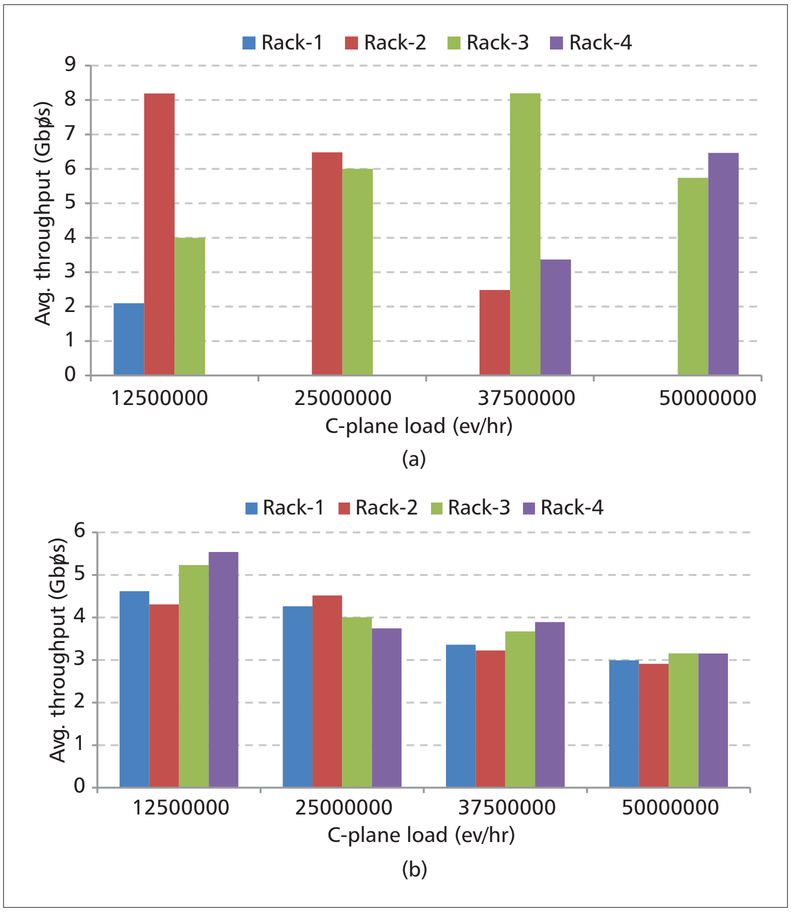


Figure 5. Average throughput per active server for UP load = 512 Gb/s: a) VSD; b) HSD.

As is evident, the deployment strategy has a marked and substantial effect on the distribution of the number of active cores on a per rack basis. With VSD (Fig. 4a), 100 percent of all cores (and hence all servers) are utilized in rack-1, while the cores in other racks become sequentially active with increasing load. As a result, there are load conditions where a rack (and hence the servers in it) may remain completely inactive and un-utilized. For example, the servers in rack-4 remain un-utilized for C_{cp} = 12.5 × 10^6 *ev*/*hr* and C_{cp} = 25 × 10^6 *ev*/*hr*. This will cause uneven load distribution over the access links and hence the ToR switches, where one link or switch may become overloaded, while the others may remain un/under-utilized. This is evident from Fig. 4a, where the load is unevenly distributed among servers in the four racks. For example, for C_{cp} = 50 × 10^6 ev/hr, all the load is on servers on racks 3 and 4, whereas racks 1 and 2 have no load on them.

In contrast to VSD, HSD deploys VNFCs evenly across the racks, resulting in even and optimum utilization of the computing and networking resources under all load conditions. This can be observed from Fig. 4b, where under all load conditions, VNFCs are evenly deployed across the racks and thus the CPU core assignment is even. This will also ensure even distribution of load over the access links, and hence the ToR switches, as evident from Fig. 5b.

Thus ,in contrast to VSD, the HSD strategy results in the optimal utilization of the DCI

At present a scalable NFV deployment planning and auto-evaluation tool is being developed based on the work presented in this article. Such a tool is expected to aid the DC operators with a quick analysis of their deployment policy.

resources without overloading any particular set of servers, and the resources scale evenly with an increase in input load. This marked difference in performance is due to the fact that in HSD, VNFCs are deployed horizontally on available servers across different racks, as opposed to VSD, where all resources of one rack need to be allocated before moving to the next rack.

Conclusion

In this article we propose and analyze two deployment strategies, namely HSD and VSD, for initial deployment of multiple VNF(C)s constituting a VNI on the operator's DCI. The performance is analyzed in terms of the cost incurred by the respective deployment strategy, where the cost is measured in terms of the utilization of a DC's computing and networking resources. The analysis is presented with reference to deploying a vEPC NS for providing EPCaaS.

As is evident from the results, for specific load profile, the total number of active servers and active cores are the same for both HSD and VSD. However, HSD delivers the best performance in terms of even distribution of load over all servers, access links, and hence the ToR switches.

In fact, for higher load profiles, HSD will result in reduced average throughput per active server as the load is evenly distributed across all racks while the number of active servers increases. In contrast to HSD, VSD is not efficient as it causes uneven distribution of VNFCs and hence load on particular servers. This may make specific racks, and hence the servers therein and associated links, to be 100 percent utilized while some other racks with servers may remain underutilized, or not utilized at all.

At present a scalable NFV deployment planning and auto-evaluation tool is being developed based on the work presented in this article. Such a tool is expected to aid DC operators with a quick analysis of their deployment policy, thus enabling them to appropriately dimension their respective DCIs to meet the expected peak traffic demands while optimizing the utilization of available resources. In the near future this tool will be integrated as part of an NFV DevOps solution that is in the planning stage.

Acknowledgment

The research work presented in this article is conducted as part of the Mobile Cloud Networking project, funded by the European Union Seventh Framework Program under grant agreement no. [318109].

References

[1] ETSI NFV GS, "Network Function Virtualization (NFV) Management and Orchestration," NFV-MAN 001 v0.8.1, Nov 2014.

[2] White paper TE-524262, "NEC Virtualized Evolved Packet Core — vEPC; Design Concept and Benefits," 2015.

[3] T. Taleb ***et al.***, "EASE: EPC as a Service to Ease Mobile Core Network Deployment over Cloud," *IEEE Network Mag.*, vol. 29, no. 2, Apr. 2015, pp. 78–88.

[4] T. Taleb, "Towards Carrier Cloud: Potential, Challenges, & Solutions," *IEEE Wireless Commun.*, vol. 21, no. 3, June 2014, pp. 80–91.

[5] T. Taleb and A. Ksentini, "Gateway Relocation Avoidance-Aware Network Function Placement in Carrier Cloud," *ACM Int'l. Conf. Modeling, Analysis & Simulation of Wireless and Mobile Systems (MSWIM'13)*, Nov. 2013.

[6] M. Bagaa, T. Taleb, and A. Ksentini, "Service-Aware Network Function Placement for Efficient Traffic Handling in Carrier Cloud," *IEEE Wireless Commun. and Net. Conf. (WCNC '14)*, Apr. 2014.

[7] F. Z. Yousaf ***et al.***, "SoftEPC: A Dynamic Instantiation of Mobile Core Network Entities for Efficient Resource Utilization," *IEEE Int'l. Conf. Commun. (ICC'13)*, June 2013.

[8] G. Somani, P. Khandelwal, and K. Phatnani, "VUPIC Virtual Machine Usage Based Placement in IaaS Cloud," *Computing Research Repository (CoRR)*, vol. abs/1212.0085, Dec. 2012.

[9] K. Le ***et al.***, "Reducing Electricity Cost Through Virtual Machine Placement in High Performance Computing Clouds," *ACM International Conference on High Performance Computing, Networking, Storage and Analysis (SC'11)*, article 22, 2011, 12 pages.

[10] C. Hyser ***et al.***, "Autonomic Virtual Machine Placement in the Data Center," HP Labs technical report, 2007.

[11] W. Shi and B. Hong, "Towards Profitable Virtual Machine Placement in the Data Center," *4th IEEE Int'l. Conf. Utility and Cloud Computing (UCC'11)*, Dec. 2011.

[12] A. Shami ; M. Mirahmadi, and R. Asal, "NFV: State of the Art, Challenges, and Implementation in Next Generation Mobile Networks (vEPC)," *IEEE Network Mag.*, vol. 28 , no. 6, Dec. 2014, pp. 18–26.

[13] "D4.1: Mobile Network Cloud Component Design," *EU Mobile Cloud Networking Project*, Nov. 2013.

[14] Cisco, "Cisco Data Center Infrastructure 2.5 Design Guide," Nov. 2, 2011.

[15] D. Nowoswiat, "Managing LTE Core Network Signaling Traffic," Alcatel-Lucent, TechBlog, July 30, 2013.

[16] T. Parker, "Chatty Smartphones Stressing Networks," *Fierce Wireless Tech*, Apr. 27, 2012.

Biographies

Faqir Zarrar Yousaf (zarrar.yousaf@neclab.eu) is a senior researcher at NEC Laboratories Europe in Heidelberg, Germany. His current research interest includes NFV/SDN related technologies and its application to the emerging 5G network architecture. He has made several contributions to the ETSI NFV standardization body, and has more than 30 publications in international conferences/journals. He received a Ph.D. from TU Dortmund, Germany in 2010, and has two masters degrees from George Washington University, USA (2001) and the University of Engineering and Technology, Peshawar, Pakistan (1999).

Paulo Loureiro (paulo.loureiro@neclab.eu) is a senior researcher at NEC Laboratories in Heidelberg, Germany. His current research interests include network based mobility protocols and extensions, with active contributions to standards (IETF and 3GPP), European projects, and internal business unit projects. Prototype implementation of the designed protocols is also one of his main activities.

Frank Zdarsky (fzdarsky@redhat.com) is a principal software engineer at Red Hat, responsible for the NFV/SDN technology and standards strategy. He has been active in ETSI NFV and OPNFV from their inception, and received the ETSI NFV Excellence Award for his contributions to the field. Prior to Red Hat, he was the head of NEC's European mobile network research, working on radio access and backhaul networks, and mobile core to service delivery platforms. He holds a Dr.-Ing. degree in computer science and a joint master's equivalent (Dipl. Wirtsch.-Ing.) in business administration and electrical engineering. He has more than 30 publications and an h-index of 13.

Marco Liebsch (marco.liebsch@neclab.eu) is currently working as a senior researcher at NEC Laboratories Europe in the area of mobility management, mobile content distribution, mobile cloud networking, and software defined networking. He has worked in different EU research projects and is contributing to standards in the IETF and 3GPP. For his thesis on paging and power saving in IP-based mobile communication networks, he received a Ph.D. from the University of Karlsruhe, Germany, in 2007. He has a long record of IETF contributions as well as RFC, journal, and conference publications.

Tarik Taleb [S'04, M'05, SM'10] (tarik.taleb@aalto.fi) received the B.E. degree (with distinction) in information engineering and the M.Sc. and Ph.D. degrees in information science from Tohoku University, Sendai, Japan, in 2001, 2003, and 2005, respectively. Prof. Taleb is a professor at the school of electrical engineering, Aalto University, Finland. He was a senior researcher and 3GPP standardization expert with NEC Europe Ltd. He was then leading the NEC Europe Labs Team, working on research and development projects on carrier cloud platforms. Prior to his work at NEC, he worked as an assistant professor at the Graduate School of Information Sciences, Tohoku University. His current research interests include architectural enhancements to mobile core networks, mobile cloud networking, mobile multimedia streaming, and social media networking. He has also been directly engaged in the development and standardization of the evolved packet system as a member of 3GPP's System Architecture Working Group. He is an IEEE Communications Society (ComSoc) Distinguished Lecturer. He is a board member of the IEEE ComSoc Standardization Program Development Board. He is serving as the Chair of the Wireless Communications Technical Committee, the largest in IEEE ComSoC. He founded and has been the General Chair of the IEEE Workshop on Telecommunications Standards: From Research to Standards, which is a successful event that received the "Best Workshop Award" from IEEE ComSoC. He is/was on the editorial board of *IEEE Transactions on Wireless Communications*, *IEEE Wireless Communications Magazine*, *IEEE Transactions on Vehicular Technology*, *IEEE Communications Surveys and Tutorials*, and a number of Wiley journals. He has received many awards, including the IEEE ComSoc Asia Pacific Best Young Researcher award in June 2009. Some of his research work has also received Best Paper Awards at prestigious conferences.

Buffer-Aided Device-to-Device Communication: Opportunities and Challenges

Although buffer-aided protocols may provide significant throughput gains in wireless networks, the opportunities and challenges of buffer-aided D2D communications are not yet fully understood. Differing from most existing works that focus on investigating buffering policy, the authors analyze the fundamental impact of the constrained buffers on the D2D communication underlaying cellular system by an optimization framework.

Haoming Zhang, Yong Li, Depeng Jin, Mohammad Mehedi Hassan, Abdulhameed Alelaiwi, and Sheng Chen

Haoming Zhang, Yong Li, and Depeng Jin are with Tsinghua University. Yong Li is the corresponding author.

Mohammad M. Hassan and Abdulhameed Alelaiwi are with King Saud University.

Sheng Chen is with the University of Southampton, and also with King Abdulaziz University..

The authors would like to extend their sincere appreciation to the Deanship of Scientific Research at King Saud University for its funding of this research through the International Research Group project No. IRG-17.

Abstract

To meet the increasing demands for popular content downloading services in next-generation cellular networks, device-to-device (D2D) communication was proposed to enable user equipments (UEs) to communicate directly over the D2D links in addition to traditional cellular operation by base stations (BSs), which is capable of utilizing the available cellular network's resource more efficiently to enhance content downloading performance. Although buffer-aided protocols may provide significant throughput gains in wireless networks, the opportunities and challenges of buffer-aided D2D communications are not yet fully understood. Differing from most existing works that focus on investigating buffering policy, we analyze the fundamental impact of the constrained buffers on the D2D communication underlaying cellular system by an optimization framework. Our study quantitatively reveals the positive correlation between the buffer sizes of BSs and UEs and the overall system performance, as well as further revealing the opportunities created by buffer-aided D2D communications for bandwidth conservation. In addition, we discuss practical challenges inherent in buffer-limited D2D communication underlaying next generation cellular networks, including increased transmission delay and optimal bandwidth allocation.

Introduction

As an underlay to LTE-A and fifth generation (5G) cellular networks, device-to-device (D2D) communication was introduced in Proximity Services (ProSe) in LTE Release-12 issued by 3GPP. D2D communications enhance many proximity-related services and applications, including content sharing and social networks. For local area services of popular content downloading, a few contents may be requested by a large number of users. Meeting this type of content downloading demand by cellular direct transmissions is extremely costly [1]. D2D communication enables user equipments (UEs) to communicate with each other directly on cellular resources [2], and may offer a high bit-rate and low power-consumption alternative. Specifically, D2D communications, in which UEs remain under the control of base stations (BSs) [2], take advantage of the physical proximity of communicating devices [1] and good channel conditions between them to better utilize the available resources.

Although D2D communication may enhance the performance of content-downloading systems, it can only take place when a UE is within the communication range of another UE or a BS that has the desired content, which indicates that the helper UE or BS must have stored the contents in its buffer in order to participate in D2D content downloading [3]. Therefore, the buffer sizes of both the BSs and UEs that serve as "relays" in the content-downloading paths play significant roles in the system performance and user experience, simply because the popular content must be stored in their limited storages so that the content can be transmitted to other UEs on appropriate occasions.

Nonetheless, existing studies [4–6] have not focused on the impact of limited buffer, a natural and indispensable attribute of UEs such as mobile phones, on the overall system performance. For example, in [4] D2D discovery processes are classified as either evolved packet core (EPC) network assisted discovery or direct discovery, and an energy-efficient D2D direct discovery is proposed, which facilitates D2D communications. Furthermore, current works fail to consider large-scale systems with hundreds of UEs [7, 8], and quantitative observations and conclusions are often reached under the unrealistic assumption that BSs and UEs have infinite storage. Thus, aiming to reveal the fundamental impact of the buffer on D2D communication underlaying cellular networks, we propose an optimization framework, a dynamic graph model that facilitates the analysis of system performance under optimal storage resource allocation and transmission control [9]. Based on this framework, we carry out the investigation under a practical network scenario with hundreds of UEs and multiple BSs. In addition to variable but limited buffer sizes of BSs and helpers, we also modulate the ratio of helpers to subscribers, and the allocation of the system bandwidth for cellular and D2D communications, which influence system performance as well. From the results and analysis, we draw conclusions regarding the opportunities and challenges created by the buffer, including boosting system performance, conserving bandwidth resources, and increasing transmission delay.

This article is structured as follows. We first provide an overview of D2D communication underlaying cellular networks. Then we propose a dynamic graph model and analyze the system constraints to form a weighed directional graph optimization model. With this optimization framework, we present our simulation results and analyze the positive impacts of the enlarged buffer, focusing on its theoretical performance bound. Next, we quantitatively analyze the boost-

0163-6804/15/$25.00 © 2015 IEEE

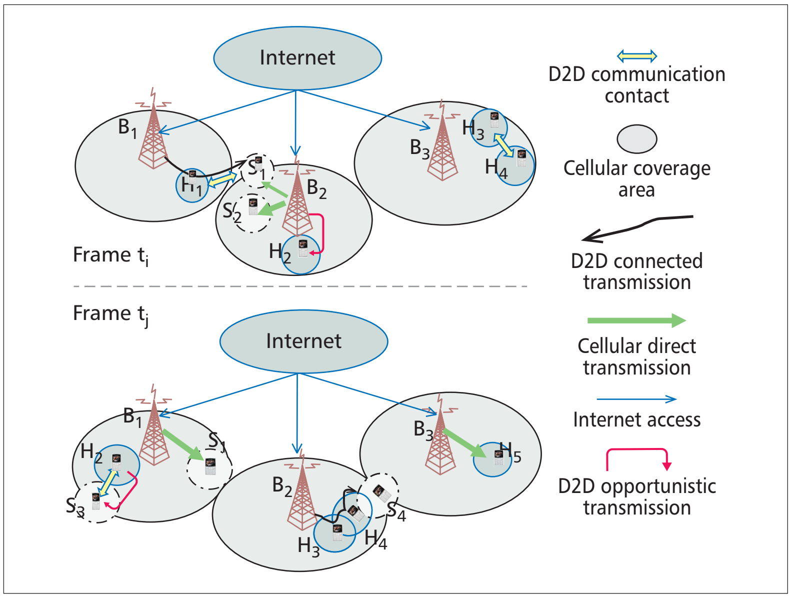

Internet
D2D communication contact
Cellular coverage area
D2D connected transmission
Cellular direct transmission
Internet access
D2D opportunistic transmission
Frame t_i
Frame t_j
Internet
B1
B2
B3
H1
H2
H3
H4
H5
S1
S2
S3
S4

Figure 1. Illustration of D2D communication underlaying a cellular system, where UEs gain access to the cellular BSs or establish D2D communication.

ed system performance and the conserved bandwidth as well as the increased delay brought by enlarging the buffer. We also analyze the influence of the helper to subscriber ratio and the allocation of the system bandwidth to cellular and D2D communications on the achievable system performance. Finally, the article is concluded and further works are pointed out.

System Overview

A typical scenario of D2D underlaying content downloading cellular networks is illustrated in Fig. 1, where the BSs, whose coverage areas and buffers are circumscribed, are connected to the Internet to provide service to UEs. The buffer-constrained UEs are mobile nodes whose positions and access states change over time. Therefore, at different time frames, their physical locations and access relations are different. Here, a time frame is loosely used to mark a system time period during which access and physical relationships remain unchanged. For example, two different time frames, t_i and t_j $(i < j)$, are indicated in Fig. 1. In content sharing systems, UEs are naturally divided into two different groups in the time frames considered: the UEs that are requesting and downloading content are called subscribers, while other UEs that currently are not retrieving content for themselves are referred to as helpers. Helpers may participate in data transmission by receiving some content, storing it in their buffer, and then transmitting it to the relevant subscribers via D2D communication. For the example depicted in Fig. 1, there are three BSs denoted by B_1 to B_3, five helpers denoted by H_1 to H_5, and four subscribers denoted by S_1 to S_4, whose requested content is delivered to them from BSs and helpers. The dotted thin circles denote the communication ranges of subscribers, while those of helpers are denoted by solid thin circles. Apart from the original way of cellular direct transmission trough BSs, UEs can also receive data from helpers in the two D2D transmission modes defined below.

D2D Connected Transmission: Utilizing the physical proximity of user devices, connected transmission paths from BSs via some helpers to subscribers can be established. In Fig. 1, S_1 and H_1 have established D2D communication contact, and a connected path from B_1 via H_1 to S_1 is established so that B_1 is able to transmit content to S_1 with the aid of H_1, during time frame t_i. Similarly, B_2 is transmitting content to S_4 via the D2D connected path $B_2 \rightarrow H_3 \rightarrow H_4 \rightarrow S_4$, during time frame t_j. D2D connected communication is also known as relay assisted communication.

D2D Opportunistic Transmission: As UEs are naturally mobile, a D2D connected path is prone to be broken and the channel conditions always fluctuate. Nevertheless, a helper is able to store some content in its finite buffer and wait for the opportunity to transmit the data to a subscriber when it establishes a communication contact with the subscriber under good channel conditions. For example, in Fig. 1, H_2 has received data from B_2 during time frame t_i and has stored the data in its buffer. During time frame t_j, when H_2 establishes a contact with S_3, it transmits the data to S_3. D2D opportunistic communication is based on the store-carry-forward mechanism that exploits opportunistic connectivity and UE mobility.

In the system, the content is available at the initial period to the BSs, and the BSs, whose buffers are also far from unconstrained, are able to store the data temporally and deliver the content to the subscribers under appropriate circumstances, by means of cellular direct transmission, D2D connected transmission, and/or D2D opportunistic transmission. In order to model this sophisticated scenario and to analyze the impact of buffering on the D2D underlaying cellular network, we develop an optimization framework for evaluating the theoretical performance bound of the D2D underlaying content downloading cellular network.

Model and Analysis Framework

Graph Model and Objective

There are five types of network events (start of cellular accessing, start of D2D contact, end of cellular accessing, end of D2D contact, and change in link quality) that may affect the access relationship and D2D contact in the network. Consequently, continuous time can be divided into n time periods, and within each time period the access states of all the network participating nodes remain unchanged. In other words, during a time period between two successive events, called a time frame, neither any contact event nor any change in link quality occurs.

Clearly, we can acquire a static graph similar to Fig. 1 for every time frame. In order to include all potential transmission modes, the graph model should include all BSs and UEs in the network. Assume that there are b BSs labeled by the set of $\mathcal{B} = \{B_1, B_2, \cdots B_b\}$, h helpers labeled as $\mathcal{H} = \{H_1, H_2, \cdots, H_h\}$, and s subscribers labeled as $\mathcal{S} = \{S_1, S_2, \cdots, S_s\}$. We can use a node to represent a BS or a UE in a given time frame. Then a static graph model of each time frame includes $b + h + s$ nodes. The data flows between nodes (BSs, helpers, and subscribers) within the time frame can be represented by directed edges, among which the edges of D2D opportunistic transmissions are from helpers to subscribers (or other helpers) and the edges of D2D connected transmissions are from BSs via some helpers to subscribers, while those of cellular direct transmissions are directly from BSs to subscribers.

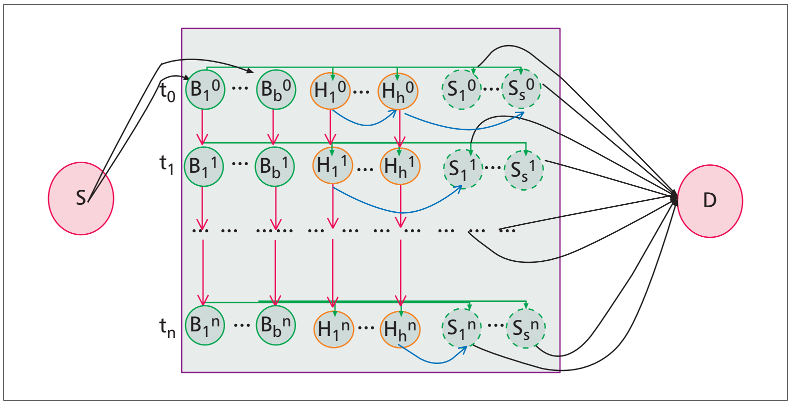


Figure 2. Static weighed directional graph model of the buffer-limited D2D communication underlaying cellular system.

Because the buffer-aided D2D mechanism enables helpers and BSs to store the content in their local buffers at certain time frames and then transmit it in the coming frames, this mechanism based on finite data buffering enables data flows across time frames and accordingly makes it possible for us to model the time evolution of this time-varying system by static graphs. When we take n time frames into consideration, we can first put the n graphs of a single time frame together and then use directed edges across time frames to represent data flows in buffers. In other words, because data flows can transmit across time frames (but only from a time frame to its successive time frame), the static graph becomes a connected digraph. For example, in Fig. 2 all the possible transmission modes, cellular direct transmissions, D2D connected transmissions, and D2D opportunistic transmissions, are included. In Fig. 2, BSs and UEs are represented by vertices, and directed edges are added to UE vertices to represent the data flows by the cellular direct transmission and/or the D2D communication.

Next we can model data flows by attributing weights to the directed edges and make the connected digraph weighted. For instance, each directed edge in the same row, green arrows for direct cellular transmission and blue arrows for D2D transmissions in Fig. 2, is associated with a positive value representing the data flow transmitted within this time frame, whose upper bound is the product of the temporal link transmission rate and the time-frame duration. It should be emphasized that the directed edges from BSs and helpers to themselves between two successive time frames (red arrows) represent the data buffering of these nodes across the two successive time frames, and the positive weights associated with these directed edges correspond to their finite buffer capacities, i.e. the finite amounts of the data stored.

Furthermore, to model the Internet access of BSs, all the content is distributed to the BSs by the Internet source, denoted as $\boldsymbol{S}$ in Fig. 2, at the initial period before time frame t_0, which represents the content downloaded from the Internet during the time period considered. Similarly, the total amount of the data received by the subscribers, which is represented by $s \times (n + 1)$ directed edges with the infinite-large transmission rate from the subscribers to the imaginary destination, denoted as $\boldsymbol{D}$ in Fig. 2, can be used to evaluate system performance [9]. When participating in D2D opportunistic communication, helpers can use cellular resources to selectively download content from BSs and store it in their buffer, and then share it. As a result of a limited buffer, helpers cannot store all content desired by subscribers. Furthermore, BSs and helpers can keep the stored content in the selected time periods, which depends on the obtained system optimization results.

To recap, all the BSs and UEs are involved in this weighted directed graph that models the temporal and spatial distributions of the network topology. Although the accessing relationships between UEs and BSs are dynamic and the communication contacts are time-varying, each row in the graph has the static topology for the duration of one time frame since the access states of all the participating network nodes remain unchanged for the duration of each frame. The objective of our optimization framework [9] is to maximize the total amount of data received by all subscribers, which is equal to the total amount of the flows to the destination $\boldsymbol{D}$ of Fig. 2.

System Constraints and Solutions

There exist three key system constraints in this buffer-limited D2D cellular network:

Flow Conservation*:* For any vertex in the graph, the amount of incoming flows must equal the amount of outgoing flows plus the amount of data stored if the vertex is a BS or helper.

Transmission Rate and Channel Access*:* Given the limited spectral resources for the D2D and cellular direct communications, the weight of each edge is directly associated with the allocated resource. Specifically, the total transmitted flow of each edge must meet the transmission bandwidth constraint. Moreover, the transmitted content flows must be strictly circumscribed within the connected UEs at each time frame, and they must also meet the interference requirements for channel access.

Finite Buffer*:* The buffer of a BS is constrained and the buffer of a helper is limited. Also, a BS typically has larger buffering capacity than a helper. To investigate the impact of a finite buffer on the theoretical performance bound of D2D communication underlaying cellular networks, we set the upper bound of the BS buffering data flows and that of the helper buffering data flows separately under the realistic assumption that every BS or helper has a limited buffer size.

After combining the above-introduced objective and constraints, we form a maximization

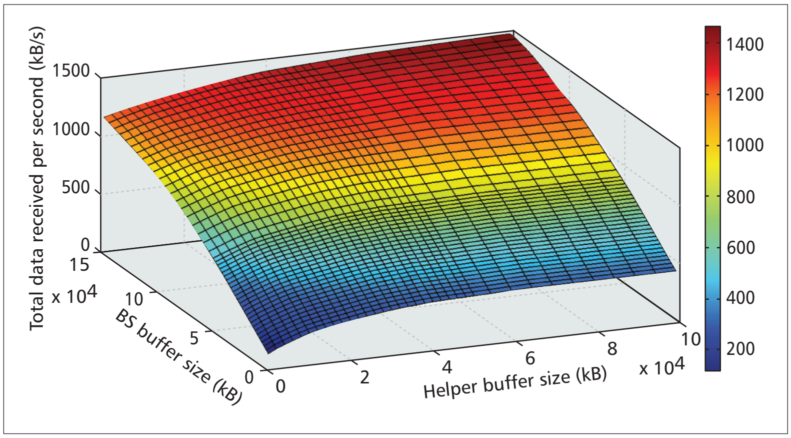


Figure 3. General trend of the total data received per second when both BS buffer size and helper buffer size are constrained and variable.

problem with the decision variables denoted by the set $\mathcal{C}$, which consists of all the data flows, i.e. the weights of all the directional edges in Fig. 2. A challenge is that not all the associated constraints are linear constraints, and thus the problem does not belong to the category of linear programming problems. Nevertheless, these nonlinear constraints can be transformed into linear expressions using the reformulation linearization technique (RLT) [10]. Consequently, this maximization problem can be solved using the existing optimization tool kits, such as CPLEX [11] and YALMIP [12].

Utilization of the Buffer

Intuitively, enlarging the buffer contributes to system performance and is accordingly a potential way of conserving bandwidth, but it will also result in an unavoidable increase in content delivering delay. Specifically, larger BS buffer sizes enable the BSs to receive more content from the source initially and to wait for appropriate opportunities to transmit them, while enlarging the buffers of helpers enables them to store sufficient amounts of data and to wait for appropriate D2D transmission opportunities to transmit more data to subscribers. By contrast, more limited buffer capacities will restrict the achievable system performance more severely.

To quantitatively exemplify the positive impacts of enlarging the buffer on the total amount of data received by subscribers, we implement simulations under a network scenario with 15 BSs and 100 UEs, among which 25 UEs are subscribers and the others are helpers. In order to yield general results, the network begins in zero-state, meaning that helpers have not retrieved content in the past and begin with an empty buffer. The number of UEs is sufficient for establishing D2D communications, and the human mobility model self-similar least action walk (SLAW) [13] is used to implement the traces of the simulated UEs. We use the typical settings in SLAW [13], where the Hurst parameter for self-similarity of waypoints is set to 0.75, the clustering range is set to 50 m, the Levy exponent for pause time is set to 1, the minimum pause time is set to 30 s, and the maximum pause time is set to 3600 s. The cell radius is 400 m and the D2D communication distance is 50 m. Since each LTE physical resource block (PRB) consists of 12 subcarriers with typically 15 kHz spacing, we allocate each UE with 800 kHz of bandwidth resources (approximately equal to four to five PRBs) to participate in cellular direct and D2D communications. Seventy percent of bandwidth resources is used for cellular direct transmissions, while the other 30 percent is allocated for D2D transmissions. In the simulation, we concentrate on investigating the influence of buffering, and we only consider the intra-cell interference, i.e. calculating the link transmission rate by only considering the interference caused by the nodes sharing the same spectrum resources [14]. We point out that there exist physical-layer techniques that can effectively manage inter-cell interference [6].

Figure 3 shows the general trend in the impacts of BS buffering and helper buffering on the capability of the system. It can be seen from Fig. 3 that there exists a significant positive correlation between the BS buffer size and the total data received per second (TDRPS) by all the subscribers, which indicates that enlarging the BS buffer contributes strongly to the enhanced performance of the entire system in the 1000-second simulation period. On the other hand, although enlarging the helper storage also has a positive impact on the system's achievable performance, it is much less effective compared to increasing the BS storage, especially when the helper buffer size is more than 50 MB. More specifically, given 100 MB of BS buffer, the TDRPS ascends only approximately 2.3 percent when the helper buffer increases from 51 MB to 100 MB, as can be clearly seen in Fig. 3. This is in contrast to more than 39 percent performance improvement due to increasing the BS buffer from 51 MB to 100 MB, with a fixed 100 MB helper buffer.

Since the significant performance improvement results from enlarging the BS buffer, a D2D content-downloading system can achieve the same required TDRPS performance with less bandwidth resources by increasing the BS buffer size. In Fig. 4a each line fitted to the selected simulation points has approximately a constant TDRPS. The results of Fig. 4a clearly demonstrate that the demand for bandwidth drops sharply with the increase in BS buffer size, given the same TDRPS requirement. This indicates that we can trade off the BS buffer size with the bandwidth. For example, with a 15 MB BS buffer, the total cellular bandwidth required is more than 700 kHz to achieve the 2 MB TDRPS, while with the 81 MB BS buffer, the system only needs 440 kHz bandwidth to achieve the same TDRPS performance. Thus, enlarging the BS buffer size can be utilized to enable the system to maintain the same level of TDRPS performance with less bandwidth. Of course, the BS buffer costs much less than cellular bandwidth.

Although enlarging the buffer offers an effective means of enhancing system performance, it will also increase data delivery delay. The average data delay in this 1000-second simulation period is calculated by computing the weighted arithmetic mean of the mid-times of every time frame, with the normalized weights set according to the total amount of data received

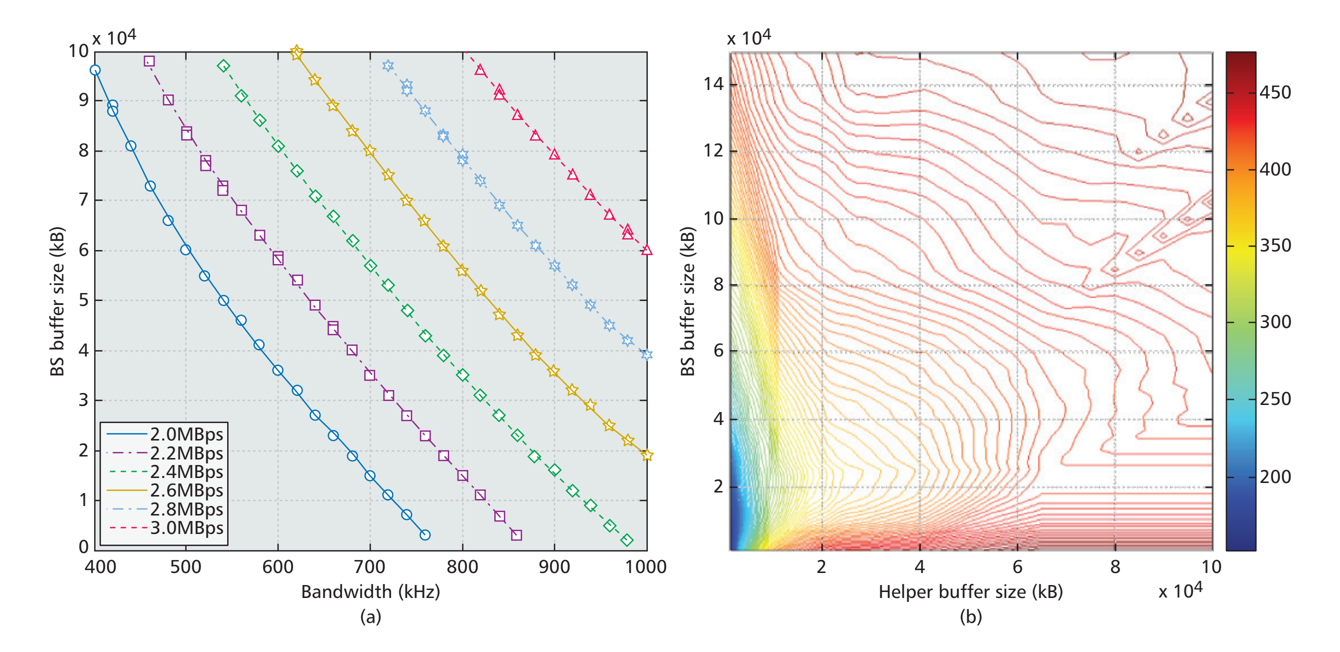


Figure 4. a) Relationship between the BS buffer size and the demand for bandwidth to meet the given TDRPS requirement; b) average data delivery delay as a function of the BS and helper buffer sizes, in the buffer-aided D2D content downloading.

in each frame, respectively. As shown in Fig. 4b, an increase in either the BS buffer or the helper buffer will result in a longer delay. The main reason for this unavoidable delay is that D2D opportunistic communication, which relies on the mobility of mobile devices, requires the helpers to store content temporally and to wait for opportunistic communication contacts.

Clearly, a long delay is always undesired as delay also impacts the user experience. For cellular services that are sensitive to both transmission delay and throughput, special protocols should be designed to circumscribe buffer size as well as the proportion of the data delivered by D2D transmission. In particular, for real-time applications, users should rely on cellular direct transmission instead of the D2D option in order to meet quality of service (QoS) requirements. However, certain delay is permissible in content downloading because this content is not real-time sensitive. Specifically, most users care more about the downloading rate but pay less attention to how long the data has stayed in the buffer of another device. In other words, it is the TDRPS instead of delay that mainly determines system performance and user experience in content-downloading services. Furthermore, with more users involved in D2D opportunistic communications, communication contacts occur more frequently, which can significantly accelerate the downloading speed of popular content. With a large proportion of content downloading services shifted to relying on D2D transmission, the network can in turn free more cellular direct transmission resources for real-time applications.

Further Discussions

In a buffer-limited D2D content-downloading underlaying cellular system, how the total system bandwidth is divided between cellular direct communication and D2D communication as well as the ratio of helpers to subscribers given the total number of UEs also influence the achievable performance. By carrying out further simulations to study the influence of these two parameters, our empirical results suggest that to achieve a reasonable optimal value of TDRPS, the proportion of the cellular direct-transmission bandwidth over the total system bandwidth should be in the range of 0.6 to 0.8, while the proportion of subscribers given the total number of UEs should be in the range of 0.5 to 0.65, respectively. Furthermore, other important issues, such as UE requirements, UE compensation, security, and energy consumption, are also discussed here.

Bandwidth Allocation

In contrast to the traditional D2D technologies that usually work on the crowded 2.4 GHz unlicensed band, in the D2D underlaying cellular network, D2D communication shares the bandwidth with cellular direct transmission. An appropriate allocation of the system bandwidth between these two communication modes is important for meeting the required system performance. After performing the simulation study under the same practical network setting (15 BSs, 25 subscribers, and 75 helpers), we acquire the results depicted in Fig. 5a and Fig. 5b for the variable BS buffer size and variable helper buffer size, respectively. The bandwidth resources allocated to each UE is also 800 kHz. Additionally, in Fig. 5a the helper buffer size is fixed (60 MB), while in Fig. 5b the BS buffer size is fixed (60 MB). Although the absolute measures may slightly fluctuate due to different mobility patterns, it is clear that a cellular direct-transmission bandwidth proportion in the range of 0.6 to 0.8 achieves the highest TDRPS. In this range, the impact of the helper buffer size is important when it is smaller than 30 MB. But a comparison between Fig. 5a

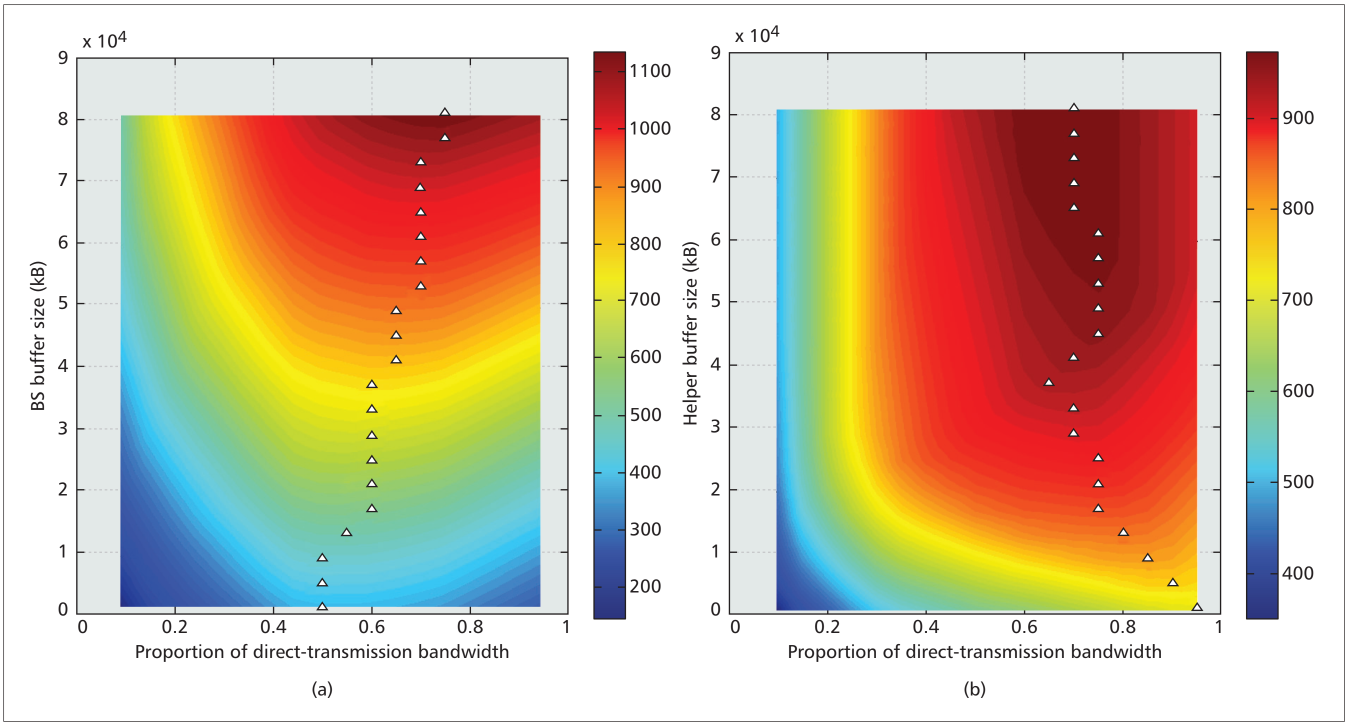


Figure 5. Total data received per second achieved for different allocated cellular direct-transmission bandwidth fractions, given fixed helper buffer size (60MB).

and Fig. 5b indicates that when the buffer size is large (larger than 30 MB in this case), or when the cellular direct transmission bandwidth proportion is small (smaller than 0.3 in this case), the impact of the helper buffer size on system performance is much less than that of the BS buffer size. For instance, given that the cellular direct-transmission bandwidth fraction is 0.2, the TDRPS remains a constant 686.3 kB when the helper buffer size increases from 33 MB to 81 MB, while the same growth in the BS buffer size leads to 47.6 percent TDRPS improvement. This observation is reasonable in that an ample D2D-transmission bandwidth fraction, which is equal to 1 minus the cellular direct-transmission bandwidth fraction, enables helpers to transmit their stored data at a rapid rate and to clear their buffers in a timely manner, and consequently the helper buffer size is less influential.

In Fig. 5 the optimal proportions of cellular direct-transmission bandwidth are marked by small black triangles. Obviously, the optimal proportion of cellular direct-transmission bandwidth has a positive correlation with the BS buffer size, which indicates that enlarging the BS buffer size contributes more to cellular direct transmission than to D2D communication. By contrast, the optimal proportion of cellular direct-transmission bandwidth tends to decrease with the increase in the helper buffer size when the helper buffer size is small, but the trend fluctuates when the helper buffer size becomes large. Considering that the allocation of resources does not vary among time frames in our graph model, we can only draw the conclusion that the optimal allocation of resources is influenced by both BS buffer size and helper buffer size. Although our scalable graph model is able to optimize the time-varying allocation of resources for different time frames, the linear programming problem will turn into a complex nonlinear programming problem and consequently reduce the efficiency of this model. Therefore, more flexible models are required to better investigate the optimal resource allocation and practical resource scheduling for D2D communication underlaying cellular networks, which calls for considerable future work, with buffer size taken into consideration.

Proportion of Subscribers

In a D2D content-downloading underlaying cellular system, the ratio of helpers to subscribers will naturally impact system performance in terms of achievable TDRPS. Figure 6 depicts the TDRPS as the function of the proportion of subscribers and the BS buffer size, given a fixed helper buffer size of 60 MB. It can be seen from Fig. 6 that when the subscriber fraction is less than 0.3, the TDRPS increases quickly as the subscriber fraction increases, and the TDRPS attains the highest values when the subscriber fraction is approximately in the range of 0.5 to 0.65. Further increasing the proportion of subscribers leads to a reduction in the TDRPS. Based on these results, we may conclude that in an optimal D2D content-downloading underlaying cellular system under the previously-mentioned practical assumptions, approximately 50 percent to 65 percent of all UEs should be subscribers, i.e. the ratio of helpers to subscribers should be approximately in the range of 0.54 to 1.

Other Issues

As revealed in our analysis framework and simulation, helpers are required to devote sufficient storage to D2D transmission in order to ensure

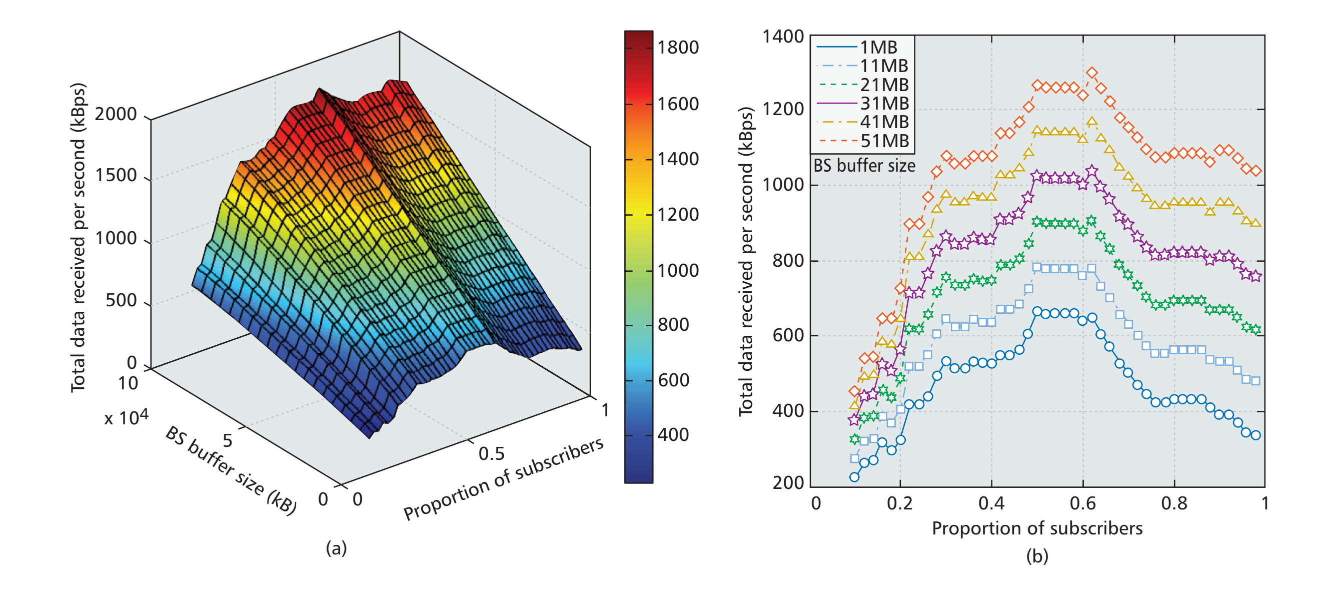


Figure 6. Total data received per second as: a) the function of the proportion of subscribers and the BS buffer size; b) the functions of the proportion of subscribers for various BS buffer sizes. The helper buffer size is fixed to 60 MB.

good system performance. In practice, subscribers may be asked to pay compensation to helpers and service providers for the related D2D data-transmission data flow and direct-transmission data flow, respectively. Furthermore, as in other forms of collaborative communication, buffer-aided D2D communication may raise security problems as well. The security issue has been discussed and potential solutions have been provided in [15]. In addition, energy consumption is also a challenging issue for D2D content-downloading underlaying cellular systems, but an energy-efficient device discovery radio with cellular network assistance has been proposed in [8]. However, it is still open to debate whether buffer size will influence energy consumption. If this influence is not negligible, related research on the trade off between buffer size and energy consumption will also be promising.

Conclusions

We have proposed an optimization framework for analyzing the performance of a buffer-limited D2D content-downloading underlaying cellular system. In particular, we have quantitatively evaluated the positive impact of enlarging the BS and helper buffer sizes on enhancing achievable content downloading performance. Moreover, we have demonstrated that enlarging the BS buffer size leads to a significant performance enhancement and, consequently, it can be utilized as an effective means of saving the required system bandwidth, while maintaining the same level of performance. We have also investigated the negative impact of enlarging the buffer size, which may increase content-downloading delay. Furthermore, based on the proposed optimization framework, we have investigated the optimal bandwidth allocation between the cellular direct communication and the D2D communication, as well as the optimal ratio of helpers to subscribers for the simulated buffer-limited D2D content-downloading underlaying cellular system under realistic assumptions. Similar to other forms of collaborative communications, mobile users are required to operate cooperatively and unselfishly to transmit the data for other users in this framework. However, if we consider the social-domain features, most users behave in a more or less selfish way. Thus, social altruism is another key factor that needs to be considered in future work. Thus, our study also opens a new research direction for bandwidth conserving, delay control, and altruistic preserving in cellular networks.

References

[1] L. Lei ***et al.***, "Operator Controlled Device-to-Device Communications in LTE-Advanced Networks," ***IEEE Wireless Commun.***, vol. 19, no. 3, June 2012, pp. 96–104.

[2] K. Doppler ***et al.***, "Device-to-Device Communication as an Underlay to LTE-Advanced Networks," ***IEEE Commun. Mag.***, vol. 47, no. 12, Dec. 2009, pp. 42–49.

[3] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Device-to-Device Collaboration Through Distributed Storage," ***Proc. Globecom 2012***, Anaheim, CA, Dec. 3–7, 2012, pp. 2397–402.

[4] A. Prasad ***et al.***, "Energy-Efficient D2D Discovery for Proximity Services in 3GPP LTE Advanced Networks: ProSe Discovery Mechanisms," ***IEEE Vehic. Tech. Mag.***, vol. 9, no. 4, Dec. 2014, pp. 40–50.

[5] M. Chen, ***et al.***, "On the Computation Offloading at Ad Hoc Cloudlet: Architecture and Service Modes," ***IEEE Commun. Mag.***, vol. 53, no. 6, June 2015, pp. 18-24.

[6] P. Pahlevan ***et al.***, "Novel Concepts for Device-to-Device Communication Using Network Coding," ***IEEE Commun. Mag.***, vol. 52, no. 4, Apr. 2014 , pp. 32–39.

[7] D. H. Lee ***et al.***, "Two-Stage Semi-Distributed Resource Management for Device-to-Device Communication in Cellular Networks," ***IEEE Trans. Wireless Commun.***, vol. 13, no. 4, Apr. 2014, pp. 1908–20.

[8] H. Nishiyama, M. Ito, and N. Kato, "Relay-by-Smartphone: Realizing Multihop Device-to-Device Communications," ***IEEE Commun. Mag.***, vol. 52, no. 4, Apr. 2014 , pp. 56–65.

[9] Y. Li ***et al.***, "Optimal Mobile Content Downloading in Device-to-Device Communication Underlaying Cellular Networks," ***IEEE Trans. Wireless Commun.***, vol. 13, no. 7, July 2014, pp. 3596–608.

[10] H. D. Sherali and C. H. Tuncbilek, "A Global Optimization Algorithm for Polynomial Programming Problems Using a Reformulation-Linearization Technique,"***J. Global Optimization***, vol. 2, no. 1, Mar. 1992, pp. 101–12.

[11] ***CPLEX: Linear Programming Solver***, available: http://www.ilog.com/.

[12] J. Löfberg, "YALMIP: A Toolbox for Modeling and Optimization in MATLAB," ***Proc. 2004 IEEE Int'l. Symp. Computer Aided Control Systems Design***, Taipei, China, Sept. 2–4, 2004, pp. 284–89.

[13] K. Lee *et al.*, "SLAW: A New Mobility Model for Human Walks," *Proc. INFOCOM 2009*, Rio de Janeiro, Brazil, Apr. 19–25, 2009, pp. 855–63.
[14] C. Xu *et al.*, "Efficiency Resource Allocation for Device-to-Device Underlay Communication Systems: A Reverse Iterative Combinatorial Auction Based Approach," *IEEE JSAC*, vol. 31, no. 9, Sept. 2013, pp. 348–58.
[15] M. Alam *et al.*, "Secure Device-to-Device Communication in LTE-A," *IEEE Commun. Mag.*, vol. 52, no. 4, Apr. 2014, pp. 66–73.

Biographies

Haoming Zhang received the B.S. degree from Tsinghua University, Beijing, China, in 2015. His research interests are in the areas of mobile computing and wireless communications. He is currently pursuing his master's degree at Carnegie Mellon University.

Yong Li [M'09] (liyong07@tsinghua.edu.cn) received the B.S. and Ph.D. degrees from Huazhong University of Science and Technology and Tsinghua University in 2007 and 2012, respectively. From 2012 to 2013 he was a visiting research associate with Telekom Innovation Laboratories and Hong Kong University of Science and Technology, respectively. From 2013 to 2014 he was a visiting scientist with the University of Miami. He is currently a faculty member in the Department of Electronic Engineering, Tsinghua University. His research interests are in the areas of mobile computing and social networks, urban computing and vehicular networks, and network science and future Internet. He has served as general chair, technical program committee (TPC) chair, and TPC member for several international workshops and conferences. He is currently an associate editor of the *Journal of Communications and Networking* and the *EURASIP Journal of Wireless Communications and Networking*.

Depeng Jin received the B.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 1995 and 1999, respectively, both in electronics engineering. He is an associate professor at Tsinghua University and vice chair of the Department of Electronic Engineering. He was awarded the National Scientific and Technological Innovation Prize (Second Class) in 2002. His research fields include telecommunications, high-speed networks, ASIC design, and future Internet architecture.

Mohammad Mehedi Hassan [M'12] is an assistant professor in the Information Systems Department at the College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He received his Ph.D. degree in computer engineering from Kyung Hee University, South Korea in February 2011. He has authored and co-authored more than 70 publications in refereed IEEE/ACM/Springer journals, conference papers, books, and book chapters. His research interests include cloud collaboration, media cloud, sensor-cloud, mobile cloud, IPTV, and wirless sensor networks.

Abdulhameed Alelaiwi [M'12] is an assistant professor in the Software Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He received his Ph.D. degree in software engineering from the College of Engineering, Florida Institute of Technology-Melbourne, USA in 2002. His research interests include software testing analysis and design, cloud computing, and multimedia.

Sheng Chen [M'90, SM'97, F'08] obtained his B.Eng. degree from the East China Petroleum Institute, Dongying, China, in January 1982, and his Ph.D. degree from City University, London, in September 1986, both in control engineering. In 2005 he was awarded the higher doctorate degree, doctor of sciences (DSc), from the University of Southampton, Southampton, UK. From 1986 to 1999 he held research and academic appointments at the Universities of Sheffield, Edinburgh, and Portsmouth, all in the UK. Since 1999 he has been with the Department of Electronics and Computer Science, University of Southampton, UK, where he currently holds the post of professor in intelligent systems and signal processing. He is a distinguished adjunct professor at King Abdulaziz University, Jeddah, Saudi Arabia. He is a chartered engineer (CEng) and a Fellow of IET (FIET). His recent research interests include adaptive signal processing, wireless communications, modelling and identification of nonlinear systems, neural network and machine learning, intelligent control system design, evolutionary computation methods and optimization. He has published more than 470 research papers. Dr. Chen is an ISI highly cited researcher in the engineering category (March 2004).

Benefits and Challenges of Virtualization in 5G Radio Access Networks

The authors focus on the benefits, challenges, and limitations that accompany virtualization in 5G radio access networks (RANs). Within the context of virtualized RAN, they consider its implementation requirements and analyze its cost. They also outline the impact on standardization, which will continue to involve 3GPP but will engage new players whose inclusion in the discussion encourages novel implementation concepts.

Peter Rost, Ignacio Berberana, Andreas Maeder, Henning Paul, Vinay Suryaprakash, Matthew Valenti, Dirk Wübben, Armin Dekorsy, and Gerhard Fettweis

Abstract

Future 5G deployments will embrace a multitude of novel technologies that will significantly change the air interface, system architecture, and service delivery platforms. However, compared to previous migrations to next-generation technologies, this time the implementation of mobile networks will receive particular attention. The virtualization of network functionality, the application of open, standardized, and inter-operable software, as well as the use of commodity hardware, will transform mobile-network technology. In this article we focus on the benefits, challenges, and limitations that accompany virtualization in 5G radio access networks (RANs). Within the context of virtualized RAN, we consider its implementation requirements and analyze its cost. We also outline the impact on standardization, which will continue to involve 3GPP but will engage new players whose inclusion in the discussion encourages novel implementation concepts.

Introduction

Cloud computing has dramatically transformed the information technology (IT) sector by introducing new ways to store and process data, create and offer services, and operate complex systems. Recognizing this power, mobile network operators are beginning to leverage cloud-computing technologies by migrating mobile network functionality to the cloud. At first, operator services and functions in the core network were the focus of the research and standards communities [1], e.g. in the European Telecommunication Standards Institute (ETSI) Network Functions Virtualization (NFV) Industry Specification Group (ISG) [2]. Recent attention has shifted to meeting the baseband-processing requirements of the radio access network (RAN) on high-volume IT hardware [3, 4].

Concurrently, an increasing number of mobile terminals and an increased demand for data motivate massive network densification through the use of small cells. In a macro-cell network, each cell serves a large number of users, which enables modeling aggregated traffic as being homogeneous even if the users have different traffic and mobility profiles. In contrast, each cell in a small-cell network serves fewer users, and hence the traffic profile observed is less homogeneous; i.e. there are areas with significant peak traffic (e.g. metro stations) and areas with no (or low) traffic (e.g. a business district during weekends). Additionally, finer spatial sampling of traffic by small cells implies stronger traffic variation per cell. For instance, [5] shows that macro-cell utilization is typically around 20 percent to 40 percent. However, since each base station (BS) must be equipped with sufficient computing resources to handle its peak load, resources are over-provisioned by a factor of 5 to 10, which is both expensive and wasteful. Centralized RAN and resource virtualization avoids over-provisioning by assigning resources intelligently and elastically based on the actual need.

Virtualization in the Context of RAN

Forms of RAN Virtualization

Virtualization can be applied to different aspects of the RAN, through spectrum virtualization, hardware sharing, virtualization of multiple radio access technologies (RATs), and virtualization of computing resources. Spectrum virtualization allows the available spectrum to be utilized more efficiently by permitting multiple network operators to share the same spectrum. Hardware and network sharing is of particular relevance for small cells in order to avoid massive over-provisioning. Virtualization of multiple RATs allows simplified management of different RATs, each dedicated to different services and offering a different quality of service (QoS). Virtualization of computing resources is a new option that builds upon the idea of co-locating the processing resources of multiple BSs at a central processing center. While early implementations provided each physical BS with its own dedicated computing resources, which resulted in an over-provisioning of computing resources, more advanced implementations permit a dynamic reassignment of processing resources to BSs. This article focuses on the potentials and challenges of moving the processing required for a mobile network to a centralized computing cloud that houses a virtualized computing infrastructure based on commodity hardware. In the following, we refer to this system as *cloud-RAN*.

Cloud-RAN as an Enabler of RAN Virtualization

A fully commoditized implementation permits complete programmability and flexibility, and it facilitates realizing the gains of a cloud-RAN implementation to the fullest extent. However, the question of whether (or not) the computational power of commoditized hardware is sufficient remains open. In order to fulfill real-time guarantees, computationally intensive parts may be executed on dedicated support hardware, e.g. co-processors similar to a graphical processing unit (GPU). To avoid a hardware lock-in, these

Peter Rost and Andreas Maeder are with Nokia Networks.

Ignacio Berberana is with Telefonica I+D.

Henning Paul, Dirk Wübben, and Armin Dekorsy are with the University of Bremen.

Vinay Suryaprakash and Gerhard Fettweis are with Technische Universität Dresden.

Matthew Valenti is with West Virginia University.

0163-6804/15/$25.00 © 2015 IEEE

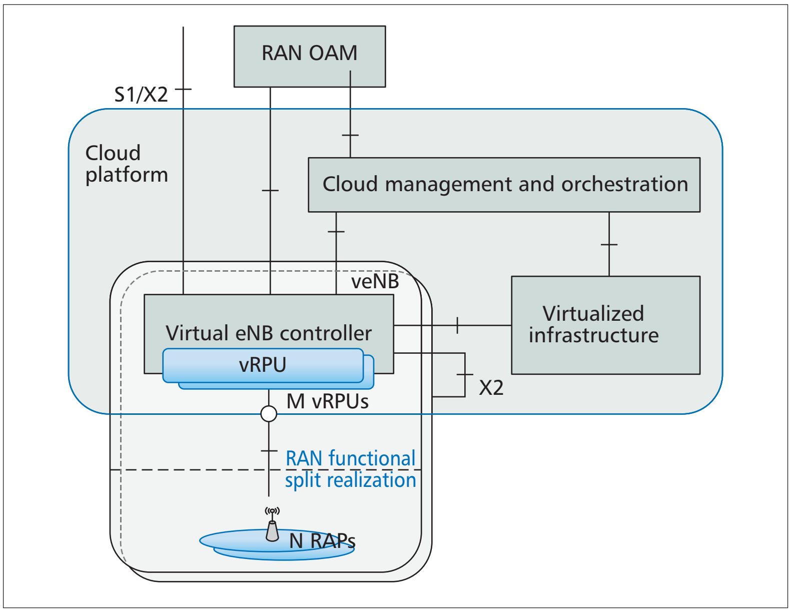

RAN OAM
S1/X2
Cloud platform
Cloud management and orchestration
veNB
Virtual eNB controller
vRPU
Virtualized infrastructure
X2
M vRPUs
RAN functional split realization
N RAPs

Figure 1. Cloud-RAN architecture implementation example.

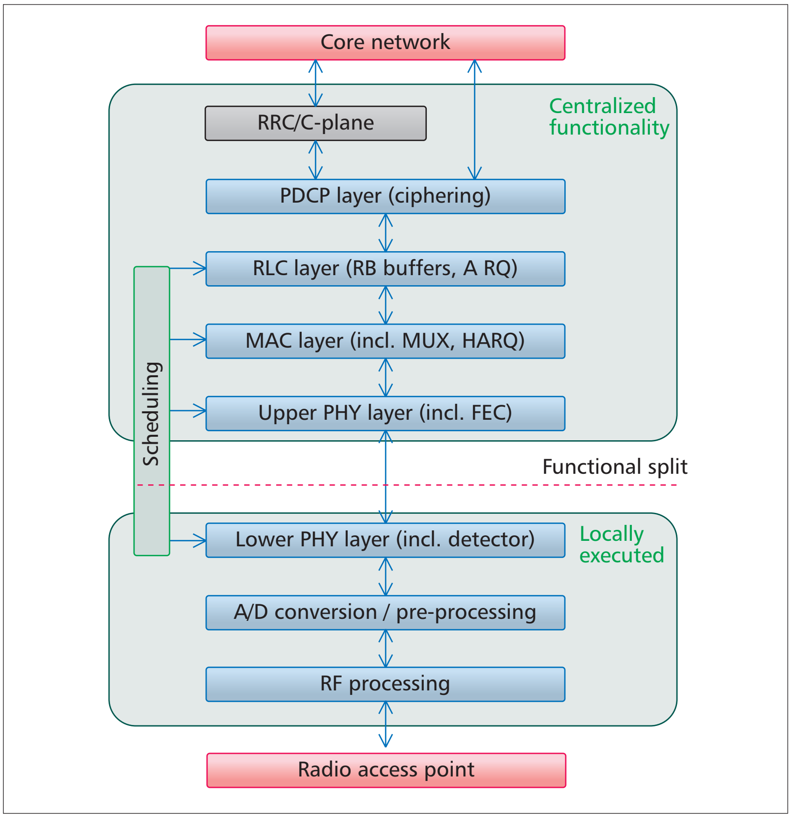

Core network
RRC/C-plane
Centralized functionality
PDCP layer (ciphering)
RLC layer (RB buffers, A RQ)
MAC layer (incl. MUX, HARQ)
Upper PHY layer (incl. FEC)
Scheduling
Functional split
Lower PHY layer (incl. detector)
Locally executed
A/D conversion / pre-processing
RF processing
Radio access point

Figure 2. RAN functional split.

co-processors could be made accessible through open interfaces (similar to OpenGL). Furthermore, the size and location of the computing centers are important design choices for a cloud-RAN system. Larger computing pools improve efficiency by reducing the likelihood of insufficient computing resources, and moving the data and its processing to remote centers may leverage potentially reduced operating costs (for instance, cheaper electricity, land, or labor). However, moving the data to a remote location will add to the communication latencies and costs. Please note that this article uses 3GPP LTE RAN functionality to explain RAN virtualization, although RAN virtualization itself will be an essential part of future 5G mobile networks.

The virtualization of computing resources and its application to RANs will require an interaction among standardization bodies concerned with RAN such as 3GPP (for which the implementation platform has so far been out of scope) and those focusing on virtualization aspects such as the ETSI NFV and Mobile Edge Computing (MEC) ISGs. Ideally, the standardization of RAN protocols takes into account characteristics stemming from virtualization and partial centralization, allowing “cloud-native” RAN implementations. Alternatively, the computing platform and its interaction with BSs may also be adapted to the RAN protocol constraints, as in the case of legacy system implementations.

Figure 1 shows an example cloud-RAN architecture that applies the ETSI NFV architectural principle to an E-UTRAN based system. The enhanced NodeB (eNodeB) as a logical network entity defined in 3GPP is implemented both in the cloud platform and at physical radio access points (RAP). The “cloudified” part of the eNodeB consists of a number of virtual RAN processing units (vRPUs), responsible for executing the RAN protocol stack, and a virtual eNodeB controller that terminates 3GPP interfaces toward the core network and other eNodeBs (virtual or non-virtual). In the context of the ETSI NFV framework, the virtualized eNodeB can be implemented as a virtual network function (VNF), which is instantiated on a virtualized infrastructure. Scaling, lifecycle management, and performance monitoring are handled by corresponding management and orchestration functions that take information from the RAN OAM (operations, administration and management/maintenance) system into account.

RAN Functional Decomposition

One of the central issues in a cloud-RAN environment is determining which functionality is executed centrally at the data center and which remains local to the RAP. Based on the results in [2] and [3], this article considers a cloud-RAN that applies the RAN functional split that is illustrated in Fig. 2. We assume that all RAN functionality starting with forward error correction (FEC) and including Layer 2 and Layer 3 processing is centralized and processed on high-volume commodity IT hardware. All functionality below FEC is executed locally at the RAP. Note that other degrees of centralization are possible but are not elaborated upon here (see [2, 3] for further details). The functional split highlighted by this article centralizes a majority of the computations, thereby exploiting most of the centralization gains, yet it relaxes the latency and throughput requirements on the backhaul as compared to a fully centralized solution that requires substantial “fronthaul” connectivity. Hence, the solution can be implemented using today’s backhaul and switching technologies.

Implementation Aspects of Cloud-RAN Systems

Cloud-RAN requires novel technologies that are provisioned along three dimensions:

- Radio access equipment: power-efficient and cost-efficient multi-RAT BSs.
- Backhaul: flexible and economical connectivity of BSs and centralized data-processing resources.
- Data processing: economical, elastic, and easily programmable centralized processing resources.

The virtualization and centralization of the RAN requires a platform that lies at the intersection of real-time architectures for processing communication signals and large-scale information processing systems. This creates dependencies between the data-processing capabilities of the computing infrastructure and the achievable communication rates of the RAN. Hence, the design, optimization, and analysis of such a system require a new conceptual framework that links the theories of data processing and communications. In this section we elaborate upon the implementation aspects of cloud-RAN and discuss challenges (and opportunities) that arise.

Real-Time Requirements of Cloud-RAN

Cloud-RAN implementations must take into account the stringent real-time requirements of the RAN. For instance, hybrid automatic repeat-request (HARQ) in LTE requires that a positive or negative acknowledgement be sent 3 ms after receiving a transport block. Failure to do so induces an unnecessary HARQ transmission, thereby lowering the throughput. In the downlink, link adaptation and radio frame generation are the main challenges. Link adaptation becomes suboptimal as the channel information becomes outdated, and radio frame generation must be synchronized between the cloud platform and the BS. Furthermore, the QoS profile enforces end-to-end latency guarantees, which require processing to be completed within a stipulated amount of time.

The actual real-time constraints that need to be fulfilled depend strongly on the functionality that is performed centrally, the dependencies between individual functional components of the RAN, and the ability to predict the processing requirements. Depending on the degree of centralization, some of the real-time constraints may be relaxed, e.g. performing decoding locally at the BS gives more time to meet the HARQ timing constraint. However, the dependencies between individual RAN functions play an important role, e.g. link adaptation determines the actual data rates on the air interface but is susceptible to changes in channel quality. If link adaptation is performed locally, then it is difficult to perform scheduling and packet segmentation at the central processor. Finally, the predictability of processing requirements is important because software jitter may violate real-time constraints. For instance, turbo-decoding requires about 80 percent of the uplink processing [11]. However, depending on the number of iterations and the actual number of information bits processed, the required complexity and decoding time may vary significantly.

Implementation Constraints of Cloud-RAN

Software implementation of RAN functionality requires a new way to design and operate the RAN. Until now, RAN functionality has been executed on dedicated hardware such as digital signal processors (DSPs) or application specific integrated circuits (ASICs). Dedicated hardware is precisely dimensioned and provides the required resources to cope with peak-traffic demands; it is highly reliable and has high performance, but does not permit sharing or virtualization of resources. In contrast, software implementation on commodity hardware may be more flexible and allow for resource sharing and virtualization. However, it is usually less reliable and has lower performance. Therefore, such implementations need to be “cloud-native” and must be designed for resilience. This cannot be achieved by merely porting existing implementations, but rather requires more advanced concepts.

Commodity hardware may be implemented by general purpose processors (GPPs) or a mix of GPPs for upper-layer processing and complementary network processors for lower-layer processing, similar to GPUs in current computer architectures. The network processors may be addressed through open interfaces (similar to OpenGL) to allow flexibility. Additionally, the processing may be performed in virtual machines or in more lightweight environments such as containers [6].

Cloud-RAN will pose new challenges to data-center architectures since it may require dedicated platforms rather than the existing platforms that have been optimized for Internet services. However, they will still be considered “commodity” due to the pervasiveness of mobile network technology. In particular, the distribution and execution of RAN processing jobs in data centers requires high-performance software defined networking (SDN) architectures that route RAN data and address processing elements within data centers efficiently. Similarly, the real-time requirements in a RAN may not allow simple migration of virtual machines (or containers) but require new mechanisms that facilitate fast transfer of processing states or RAN protocol states. The efficiency with which processing elements (containers or virtual machines) are assigned to data packets has a major impact on the elasticity of the system.

The requirements on the data center will also depend on the manner in which the processing is implemented. For instance, processing may be performed on a per-user-terminal basis, a per-BS basis, or a per-cluster basis. The first option provides higher scheduling granularity; the second option may simplify the process of merging data originating from different user terminals (e.g. for scheduling); the third option may simplify the joint processing of data across multiple BSs. Furthermore, the parallelization of processing could be done with a very low granularity (on a per packet basis) or with higher granularity (on a per-BS basis). The need

The actual real-time constraints that need to be fulfilled depend strongly on the functionality that is performed centrally, the dependencies between individual functional components of the RAN, and the ability to predict the processing requirements.

for synchronization objects (semaphores) also increases with an increase in granularity, and this limits the processing performance significantly. In contrast, processing each packet on a separate processor (or core) allows for decoupling processes, and therefore avoids the need for synchronization objects.

Joint RAN/Cloud Resource Management

In a cloud-RAN system, the radio and data-processing resources should be managed jointly, i.e. radio resource allocation must adapt not only to the prevailing channel conditions and required quality of service, but must take into account the demand for computational resources imposed by the radio allocation. This is a predictive task, i.e. the system has to estimate the required computational complexity, estimate the available computational resources, and then adapt the RAN resource allocation accordingly.

In a cloud-RAN system, the radio and data-processing resources should be managed jointly; i.e., radio resource allocation must adapt not only to the prevailing channel conditions and required quality of service, but must take into account the demand for computational resources imposed by the radio allocation.

One possibility for carrying out this joint optimization is to account for the data-processing load during link adaptation, i.e. the resource scheduler could incorporate a weighted metric that penalizes choices that lead to high computational demands. Furthermore, as the number of users served by a BS increases, the expected traffic and processing requirements will also increase. This may require a re-assignment of processing resources to virtual machines within the data-processing center and should be anticipated by the scheduler. Additionally, the scheduler must be able to operate at the computational capacity, i.e. the maximum system throughput using a given amount of computational resources. This requirement is particularly important during peak-traffic hours when many users connect to the mobile network and the computational load approaches the system limit.

The previous examples described operational tasks. However, there are also dimensioning and positioning challenges. For instance, [7] provides a framework for estimating the amount of computational resources required for an expected number of users. The concept of computational outage, which is the likelihood that the available computational resources are insufficient to meet the instantaneous computational load, is introduced. Using this framework, the required computational resources can be predicted, and computationally aware schedulers that maximize the system utilization and prevent computational outage can be designed.

Data-Processing Complexity of RAN Protocols

In a cloud-RAN system, the data-processing requirements depend on many different factors. For example, if the transmission rate increases, more information bits need to be processed, which in turn linearly increases the computational load. Additionally, if a communication link operates close to its Shannon capacity, even more receiver processing is required, which can be attributed to the need for additional turbo decoder iterations. As a result, the processing load increases super-linearly as the system operates increasingly close to capacity, and the load depends on both the instantaneous channel conditions and the scheduling policy, which determines how close to capacity the system operates.

Therefore, in a manner similar to exploiting channel diversity in mobile networks (e.g. multi-user diversity in scheduling or spatial diversity in multiple antenna systems), *computational* diversity can also be exploited. Computational diversity exploits the large fluctuations in the data-processing load imposed by multiple users with diverse channel conditions. Hence, if multiple users are served by a cloud-RAN instance, their diverse computational requirements may be used to improve the resource utilization since the computational assets need to be provisioned according to the expected cumulative load of the users rather than the peak load of any given user. Furthermore, by dynamically adapting the modulation and coding scheme through the use of appropriate *computationally aware* link-adaptation algorithms, the data-processing load can be controlled.

From a user's perspective, there is no difference between a channel outage and a computational outage: in either case, the communications fail and another attempt must be made to transmit the packets. The model introduced in [7] accurately predicts the data-processing resources required to perform uplink decoding in a multicell scenario for a given threshold on computational outage probability. Using the empirically determined computational load discussed in [11] and assuming a 10 MHz LTE channel and that each turbo-decoder iteration requires 1000 FLOPS per data bit, we can estimate the overall required data-processing capabilities for a reference setup involving server blades equipped with four Intel Xeon 4870 (10-core processor) and 128GB RAM.

Based on these assumptions and the framework introduced in [7], Fig. 3 shows the computational resources required for LTE in a data center. We compare two cases of centralized computing: in the cloud-RAN case with virtualization, processing resources may be flexibly re-assigned to BSs, while in the second case without virtualization, each BS is serviced by its own dedicated computational resources (as is the case in fully centralized RAN). For both cases, cells are assumed to be fully loaded and the computational outage probability is set to 10 percent. We quantify the computational requirement by the number of servers using the aforementioned architecture. When resources are shared, we see a reduction of approximately 50 percent in the data-processing resources required.

Cost Analysis

A major issue in cloud-RAN implementation is its impact on the cost of mobile networks (and the capital expenditure (CAPEX) in particular). RAN virtualization and centralization over a non-ideal backhaul may allow cost-efficient implementations. The backhaul deployment and the ability to use existing infrastructure as well as non-ideal backhaul technologies play a critical role in the economic analysis of cloud-RAN. This section presents a cost-analysis for the RAN functional split illustrated in Fig. 2.

Cost-Components in Cloud-RAN

An evaluation of the CAPEX for a cloud-RAN system must consider costs stemming from different network components. In [8] four different network layers[1] as illustrated in Fig. 4 are distinguished: users, BSs, backhaul nodes, and data centers. The lowest layer represents the users and assumes a particular average traffic demand per user. Both micro and macro BSs are overlaid on the same layer. Backhaul nodes consist of aggregation points that are then connected to data centers wherein centralized processing is performed. This model captures the most important cost components and facilitates analysis of the salient trends in a cloud-RAN deployment.

BSs in a cloud-RAN environment process may perform only part of the RAN protocol stack at the local site. Hence, their size and costs might depend upon the amount of processing performed locally. Furthermore, the difference in cost between macro and micro BSs can vary significantly because the latter use lower transmit power and fewer antennas, thereby reducing the cost per access point significantly. In contrast, increased centralization also requires adequate backhaul technologies that cater to the throughput and latency requirements. In fully centralized systems, high-performance optical fiber is required. Since it is very expensive, its cost may even outweigh the RAN cost reduction described above. The functional split considered in this article (Fig. 2) does not require specific backhaul technologies and can also be applied to non-ideal backhaul (with latencies above 1 msec). Therefore, the backhaul costs may be significantly lower compared to a fully centralized RAN. Finally, the deployment of additional data centers will be necessary. However, since each data center hosts only a few servers (as seen in the number of servers required in Fig. 3), the additional data-processing hardware required may be deployed at preexisting points in order to simplify site acquisition and reduce costs.

Example Cost Analysis

In [8] a generic framework and an expression for the CAPEX of the entire network have been derived, which can be applied to three different scenarios by substituting the appropriate component cost values. The first scenario is cloud-RAN with virtualization and using a mix of optical and wireless backhaul technologies (each 50 percent). The second scenario is a fully centralized RAN, which refers to a functional split above A/D conversion in Fig. 2, without virtualization, and requiring optical fiber connectivity (as prevalent today). The third scenario is distributed RAN (DRAN), which refers to a conventional implementation where all RAN processing is performed locally at the BSs. Table 1 shows an example budget for a cloud-RAN network that uses the functional split illustrated in Fig. 2. It compares the costs for DRAN and cloud-RAN. In our example, we assume 170 active users per km^2, an average traffic demand per user of 10 Mb/s, and a mix of 50 percent microwave and 50 percent optical fiber technology for the backhaul.

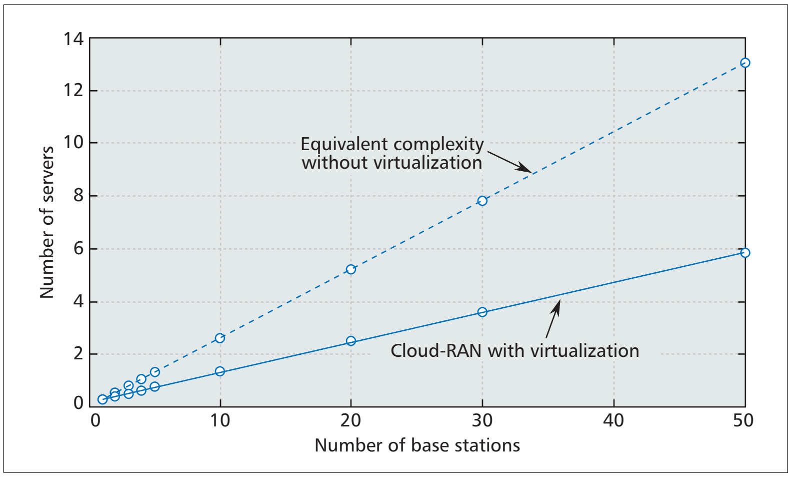


Figure 3. The number of IT server blades required depending upon the number of centralized (fully loaded) BSs.

Type of cost	DRAN	Cloud-RAN
Macro base station	$50k	$25k
Micro base station	$20k	$10k
Microwave BH	$50k per link plus $5k per kilometer	
Optical fiber BH	$5k per link plus $100k per kilometer	
Data center		$40k
Server blades		$20k each (Fig. 3)

Table 1. Exemplary budget for Cloud-RAN analysis (more details are given in [7]).

Figure 5 shows the resulting CAPEX for the three cases of DRAN, cloud-RAN, and fully centralized RAN. In all three cases, the expected area throughput is used as a basis for normalization and results are plotted over different data center densities. It is important to note that one data center may consist of only a few server racks at an existing point of presence within the mobile network. This reduces the operational expenditure (OPEX) because no additional site rental is necessary. Furthermore, small data centers promote greater failure resilience and they reduce the traffic within the metropolitan transport network. Therefore, considering Fig. 5, a density of one or two data centers per square kilometer appears realistic in a very dense urban small-cell deployment. If we further increase the density of small-cells, e.g. due to higher data rate demands and user density, then the cost effectiveness of cloud-RAN would increase even further as the exploited centralization gains in cloud-RAN also increase (similar to the over-provisioning of distributed RAN) and the cost-reduction per BS becomes more dominant.

The results show that cloud-RAN based on the applied RAN functional split can be more cost effective than a DRAN implementation. However, the actual benefit may depend on the scenario, parameterization, and actual traffic

[1] Note that the term "layer" here does not refer to the layers in the Open Systems Interconnection (OSI) model, which standardizes the internal functions of a communication system, but instead refers to each network component.

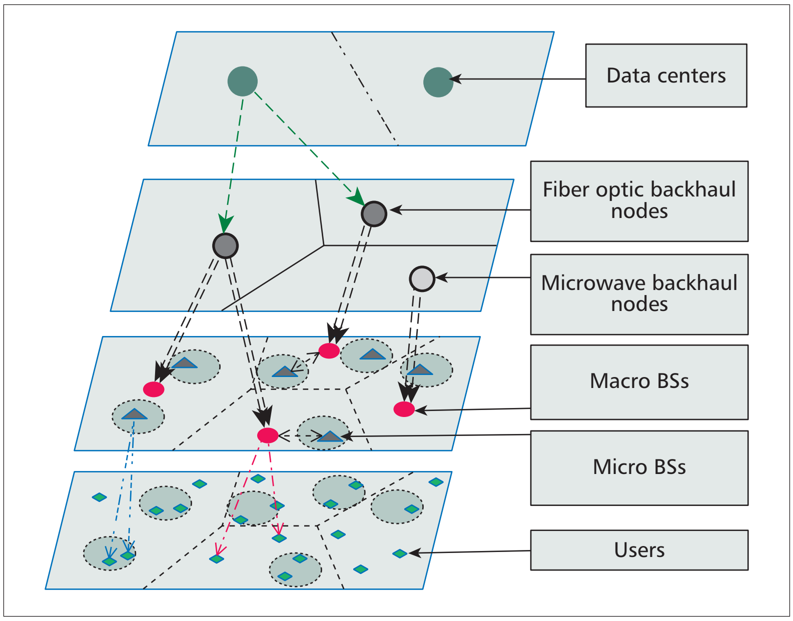


Figure 4. Model of a heterogeneous network with multiple network components.

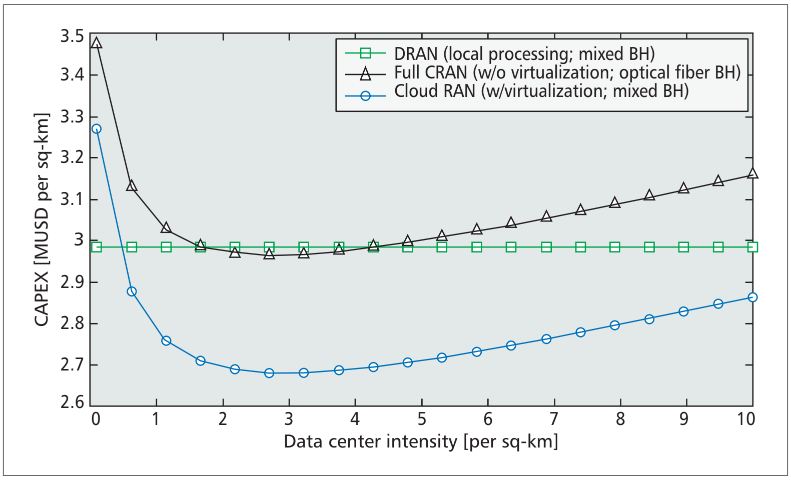


Figure 5. Fost efficiency analysis.

demand. Additionally, the architecture presented here holds the potential for reduced OPEX due to lower maintenance costs on site as well as easier management through standard IT management mechanisms.

Standardization Impact

Several standards development organizations (SDOs) have recognized virtualized cloud-RAN as one of the key technologies to meet the requirements of 5G networks. The mobile communication industry (including operators, vendors, and chipmakers) collaborates in various industry fora and projects on drafting 5G requirements. The Next Generation Mobile Networks (NGMN) Alliance created its 5G Initiative, focusing on an operator's view of 5G requirements. In Europe, the ETSI and several 5G-related projects funded by the European Commission, such as iJOIN[2] and METIS[3], work toward a common view on 5G. For the North American market, the 4G Americas industry forum published recommendations on 5G requirements and solutions [12]. Similar publications are available from Asian 5G-related activities, e.g. from the 5G Forum in Korea, ARIB in Japan, and the IMT-2020 Promotion Group as well as the Chinese Ministry of Science and Technology project 863-5G. Finally, the International Telecommunication Unit promotes its "IMT-2020 and beyond" program in the ITU-R working group 5D.

While all these activities are pre-standardization efforts, they frequently mention coordinated transmission and flexible networks/services as key enabling technologies, and cost and resource efficiency as key requirements for future 5G solutions, all of which strongly point toward virtualized cloud-RAN as a solution.

Standardization of virtualized cloud-RAN requires activities in both networking and computing frameworks, such as SDN and NFV, as well as enhancements to the mobile network functionalities and architecture in order to take full advantage of the virtualization approach.

Network and computing frameworks are addressed by several SDOs such as the Open Networking Foundation (ONF, [15]) for SDN based on the OpenFlow protocol, ETSI NFV[4] and ETSI MEC.[5] Virtualized cloud-RAN as a use case implies new and stringent requirements (e.g. on latency) for these frameworks. The ONF Wireless & Mobile project, which aims to collect use cases and determine architectural as well as protocol requirements for extending ONF-based technologies to wireless and mobile domains, is a first step toward the identification of such requirements in the transport-network area.

For computing frameworks, ETSI NFV aims to evolve quasi-standard IT virtualization technology to consolidate many network equipment types into industry-standard high-volume servers, switches, and storage. It enables the implementation of network functions in software that can run on a range of industry-standard server hardware and can be moved to, or loaded in, various locations in the network as required, without the need to install new equipment. RAN virtualization use cases are described, but not yet addressed, in the current ETSI NFV recommendations. Meanwhile, the newly created ETSI MEC aims to offer application developers and content providers cloud-computing capabilities and an IT service environment at the edge of the mobile network.

The mobile network aspect of 5G will be led by the 3rd Generation Partnership Project (3GPP). Partners in 3GPP are still awaiting a consensus on 5G requirements before concrete actions are taken. Some aspects of virtualized network functions have already been addressed, such as in the SA2 system architecture working group with the new work item on flexible mobile service steering (FMSS) in the operator's core network. However, substantial impact on specifications in 3GPP RAN working groups cannot be expected before future LTE Releases 14, 15, and beyond. Thus it can be expected that virtualization and software control may help simplify the network architecture and support the flexible allocation of radio processing functionalities.

[2] http://www.ict-ijoin.eu

[3] www.metis2020.com

[4] http://www.etsi.org/technologies-clusters/technologies/nfv

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

A first glimpse of RAN-related efforts is already visible in the network virtualization work stream of the Small Cell Forum [14], which analyzes requirements of RAN virtualization. In particular, different functional splits and their associated performance benefits and constraints are discussed.

Conclusions and Further Challenges

This article discussed the challenges, benefits, and opportunities of virtualizing RAN functions. We paid particular attention to the data-processing requirements, which are directly influenced by the operation and design of the RAN itself. Using results from this complexity analysis, we discussed the main contributors to the cost of a virtualized RAN system. We further discussed prominent implementation challenges such as joint resource optimization for RAN and the cloud computing platform.

Based on the discussion in this article, we can conclude that virtualized RAN provides greater flexibility to the mobile network operator and potentially reduces network costs. It is our opinion that RAN virtualization will be an integral part of 5G and that commodity IT platforms have the potential to host cloud RAN networks. However, there are many challenges beyond those tackled in this article, such as communication interfaces within data centers, parallelization of RAN functions, state maintenance, and the impact of the RAN protocol stack. Many of these aspects are detailed in [9].

There are many challenges beyond those tackled in this article, such as communication interfaces within data centers, parallelization of RAN functions, state maintenance, and the impact of the RAN protocol stack

Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Program FP7/2007-2013 under grant agreement no. 317941 – project iJOIN, http://www.ict-ijoin.eu. The European Union and its agencies are not liable or otherwise responsible for the contents of this document; its content reflects the view of its authors only.

References

[1] T. Taleb, "Toward Carrier Cloud: Potential, Challenges, and Solutions," *IEEE Commun. Mag.*, June 2014.

[2] P. Rost **et al.**, "Cloud Technologies for Flexible 5G Radio Access Networks," *IEEE Commun. Mag.*, May 2014.

[3] D. Wübben **et al.**, "Benefits and Impact of Cloud Computing on 5G Signal Processing," *IEEE Signal Proc. Mag.*, Nov. 2014.

[4] J. Kerttula **et al.**, "Implementing TD-LTE as Software Defined Radio in General Purpose Processor," *ACM SIGCOMM Software Radio Implementation Forum 2014*, Chicago (IL), USA, Aug. 2014.

[5] H. Guan, T. Kolding, and P. Merz, "Discovery of Cloud-RAN", *Cloud-RAN Wksp.*, Apr. 2010.

[6] G. Banga, P. Druschel, and J. C. Mogul, "Resource Containers: A New Facility for Resource Management in Server Systems," *Proc. Symp. Operating Systems Design and Implementation*, New Orleans, LA, USA, vol. 99, Feb. 1999, pp. 45–58.

[7] P. Rost, S. Talarico, and M.C. Valenti, "The Complexity-Rate Tradeoff of Centralized Radio Access Networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6164-6176, Nov. 2015.

[8] V. Suryaprakash, P. Rost, and G. Fettweis, "Are Heterogeneous Cloud-Based Radio Access Networks Cost-Effective?" *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, October 2015

[9] EU FP7 Project iJOIN, "iJOIN Deliverable D5.2: Final Definition of iJOIN Requirements and Scenarios," Nov. 2014.

[10] Open Networking Foundation, "Software-Defined Networking: The New Norm for Networks" (online document]), white paper, Apr. 2012.

[11] S. Bhaumik **et al.**, "CloudIQ: A Framework for Processing Base Stations in a Data Center," *IEEE MobiCom*, Istanbul, Turkey, Aug. 2012.

[12] 4G Americas, "Recommendations on 5G Requirements and Solutions," white paper, Oct. 2014.

[13] 3GPP, "TR 36.839 V11.1.0, Mobility Enhancements in Heterogeneous Networks," technical report, Sept. 2012.

[14] Small Cell Forum, "Virtualization for Small Cells: Overview," technical report, June 2015.

[15] Open Networking Foundation, "OpenFlow-Enabled Mobile and Wireless Networks," technical report, Sept. 2013.

Biographies

Peter Rost [SM] (peter.rost@ieee.org) received his Ph.D. degree from Technische Universität Dresden, Dresden, Germany, in 2009 (supervised by Prof. G. Fettweis), and his M.Sc. degree from the University of Stuttgart, Stuttgart, Germany, in 2005. Since May 2015 Peter has been a member of the Radio Systems research group at Nokia Network, focusing on 5G RAN architecture. From 2010 to 2015 he was a member of the Wireless and Backhaul Networks group at NEC Laboratories Europe. He has been active in 3GPP RAN2 and several EU projects (e.g. EU FP7 iJOIN as technical manager (www.ict-ijoin.eu)). He is a member of IEEE ComSoc GITC, IEEE Online GreenComm Steering Committee, VDE ITG Expert Committee "Information and System Theory." He is an executive editor of *IEEE Transactions on Wireless Communications*.

Ignacio Berberana (ignacio.berberana@telefonica.com) received the M.S. degree in mining engineering from Madrid Polytechnic University in 1987. In 1987 he enjoyed a National Research Grant for studying adaptive control systems. In 1988 he joined Telefonica I+D (Telefonica research labs), where he has worked in areas covering satellite and wireless communications, including several European projects (CODIT, MONET, Rainbow, Artist4G, iJOIN). Currently he is responsible for the Innovation unit in the Radio Access Networks direction of the Telefónica Global CTO office, which deals with long term evolution of mobile access, including 5G systems.

Andreas Maeder [M] (andreas.maeder@nokia.com) is a senior researcher at Nokia Networks, Munich, Germany. He has been with NEC Laboratories Europe in Heidelberg, Germany from 2008 to 2015. Andreas received his Ph.D. from the Unversity of Wuerzburg, Germany. Currently, his main area of research is the convergence of IT and telecommunication technologies. Andreas has contributed to the standardization of broadband wireless access technologies in IEEE and 3GPP since 2008. He was rappparteur of the work item on user plane congestion management in 3GPP SA2. He was chair of the IEEE BWA workshop 2013, and has authored numerous scientific articles, conference papers, and patents.

Henning Paul [M] (paul@ant.uni-bremen.de) received his Dr.-Ing. (Ph.D.) and Dipl.-Ing. (equivalent to M.Sc.) degrees from the University of Bremen, Germany in 2012 and 2007, respectively. He has been with the Department of Communications Engineering, University of Bremen, Germany since 2007, where he currently is a senior researcher and lecturer. His research is focused in the field of wireless sensor networks, in-network processing, and distributed signal processing algorithms for mobile communications. He is member of VDE ITG.

Vinay Suryaprakash [M] (vinay.suryaprakash@alcatel-lucent.com) received his doctorate from the Technische Universität Dresden, Germany under the supervision of Prof. Gerhard Fettweis in 2014. He received his master of science in electrical engineering from the University of Southern California, Los Angeles in 2007, after which, as an employee of Cisco Systems Inc., San Jose, CA from 2008 to 2010, he was involved in analysis and testing of load balancers that help regulate traffic in large networks. His current research focuses on using stochastic geometry for the system level analysis of wireless networks. In 2013 he was nominated as one of the six finalists of the Qualcomm Innovation Fellowship 2013 from contestants all across Europe.

Matthew C. Valenti [SM] (Matthew.Valenti@mail.wvu.edu) received his Ph.D. from Virginia Tech, USA, and has been on the faculty of West Virginia University (WVU) since 1999, where he is currently a professor and the Director of the Center for Identification Technology Research (CITeR). His research interests are in wireless communications, cloud computing, and biometric identification. He serves on the Executive Editorial Committee of *IEEE Transactions on Wireless Communications*, as an editor for *IEEE Transactions on Communications*, and as chair of the Communication Theory Technical Committee of the IEEE Communications Society (ComSoc). He is active in the organization of several ComSoc sponsored conferences, including MILCOM, ICC, and Globecom.

Dirk Wübben [SM] (wuebben@ant.uni-bremen.de) received the Dipl.-ing. (FH) degree in electrical engineering from the University of Applied Science Münster, Germany, in 1998, and the Dipl.-ing. (Uni) degree and the Ph.D. degree in electrical engineering from the University of Bremen, Germany in 2000 and 2005, respectively. In 2001 he joined the Department of Communications Engineering, University of Bremen, Germany, where he is currently a senior researcher and lecturer. His research interests include wireless communications, signal processing, cooperative communication systems, and channel coding.

Armin Dekorsy (Dekorsy@ant.uni-bremen.de) is the head of the Department of Communications Engineering, University of Bremen. He received his Dipl.-Ing. (FH) (B.Sc.) degree from Fachhochschule Konstanz, Germany; the Dipl.-Ing. (M.Sc.) degree from the University of Paderborn, Germany; and the

Ph.D. degree from the University of Bremen, Germany, all in communications engineering. From 2000 to 2007 he worked as a research engineer at Deutsche Telekom AG, and as a distinguished member of technical staff (DMTS) at Bell Labs Europe, Lucent Technologies. In 2007 he joined Qualcomm GmbH as European research coordinator conducting Qualcomms' internal and external European research activities. He has long-term expertise in the research of wireless communication systems, baseband algorithms, and signal processing. Prof. Dekorsy has published more than 160 journal and conference publications and holds more than 17 patents in the area of wireless communications. Prof. Dekorsy is a member of the IEEE Communications Society and IEEE Signal Processing Society, the VDE/ITG expert committee on "Information and System Theory", and represents the University at ETSI, NETWORLD2020, and at the OneM2M forum.

Gerhard P. Fettweis [F] (gerhard.fettweis@vodafone-chair.com) earned his Ph.D. under H. Meyr's supervision from RWTH Aachen in 1990. After one year at IBM Research in San Jose, CA, he moved to TCSI Inc., Berkeley, CA. Since 1994 he has been Vodafone Chair Professor at TU Dresden, Germany, with 20 companies from Asia/Europe/US sponsoring his research on wireless transmission and chip design. He coordinates two DFG centers at TU Dresden, namely cfaed and HAEC. He is an IEEE Fellow and a member of the German academy Acatech. His most recent award is the Stuart Meyer Memorial Award from IEEE VTS. In Dresden his team has spun-out 13 start-ups, and set up funded projects in volume of close to EUR 1/2 billion. He has helped organize IEEE conferences, most notably as TPC Chair of ICC 2009 and TTM 2012, and as General Chair of VTC Spring 2013 and DATE 2014.

XG-FAST: The 5th Generation Broadband

Traditionally, copper network operators complement a fiber-to-the-home strategy with a hybrid fiber-copper deployment in which fiber is gradually brought closer to the consumer, and digital subscriber line technology is used for the remaining copper network. The authors propose the system concepts of XG-FAST, the 5th generation broadband (5GBB) technology capable of delivering a 10 Gb/s data rate over short copper pairs.

Werner Coomans, Rodrigo B. Moraes, Koen Hooghe, Alex Duque, Joe Galaro, Michael Timmers, Adriaan J. van Wijngaarden, Mamoun Guenach, and Jochen Maes

The authors are with Bell Labs, Alcatel-Lucent.

COMMUNICATIONS STANDARDS

Abstract

Traditionally, copper network operators complement a fiber-to-the-home (FTTH) strategy with a hybrid fiber-copper deployment in which fiber is gradually brought closer to the consumer, and digital subscriber line (DSL) technology is used for the remaining copper network. In this article we propose the system concepts of XG-FAST, the 5th generation broadband (5GBB) technology capable of delivering a 10 Gb/s data rate over short copper pairs. With a hardware proof-of-concept platform, it is demonstrated that multi-gigabit rates are achievable over typical drop lengths of up to 130 m, with net data rates exceeding 10 Gb/s on the shortest loops. The XG-FAST technology will make fiber-to-the-frontage (FTTF) deployments feasible, which avoids many of the hurdles accompanying a traditional FTTH roll-out. Single subscriber XG-FAST devices would be an integral component of FTTH deployments, and as such help accelerate a worldwide roll-out of FTTH services. Moreover, an FTTF XG-FAST network is able to provide a remotely managed infrastructure and a cost-effective multi-gigabit backhaul for future 5G wireless networks.

Introduction

About two decades ago, copper networks carried just a few kilobits per second. Today, digital subscriber line (DSL) technology offers rates exceeding 100 Mb/s over those same copper networks. Communication, computing, and storage capacity continue to increase significantly every year, and fixed access network technologies must provide ever higher data rates to support a wide variety of high-speed applications, such as video-on-demand, fast downloads, support for cloud-based applications, and transport and backhaul capabilities for wireless networks.

For copper networks, the achievable capacity is typically dominated by attenuation. Optical fiber technology has a much longer reach due to its inherent low attenuation, and is therefore ideal to transport high data rates over long distances. Current passive optical network (PON) technologies provide data rates up to 10 Gb/s, and the next-generation PON systems are expected to carry up to 40 Gb/s. However, a direct transition to full fiber-to-the-home (FTTH) connectivity is hampered by enormous capital expenditures and the very long roll-out time needed to build these networks. At the same time, nearly every household or residential building in developed countries is connected to the copper-based telephone network. Reusing these copper cables as a physical transport medium for broadband access networks has proven to be of great economical value. For this reason, the fiber connection has gradually been brought closer to the end-user, shortening the remaining copper loop to within a few hundred meters from the customer. This evolution has been accompanied by new generations of DSL technologies that leverage the higher capacity of these shorter copper loops.

For short and medium length copper loops, instead of attenuation, crosstalk across adjacent twisted pair cables proved to be the limiting factor to improve data rates. Vectored VDSL2 [1], the most recently deployed DSL technology, effectively suppresses crosstalk across adjacent copper lines in a cable. Using this technology, single line performance can be attained, making the loop length the limiting factor once again. This has led to the development of the fourth generation broadband (4GBB) technology, called "fast access to subscriber terminals," in short, G.fast, which was standardized in Dec. 2014 [2–4]. The target deployment scenarios for G.fast include multi-home deployments such as fiber-to-the-distribution-point (FTTdp) and fiber-to-the-building. The G.fast standard is optimized for copper loops up to 250 m with a 106 MHz bandwidth and a maximum aggregate data rate that approaches 1 Gb/s. Since the impact of crosstalk interference increases with frequency, vectoring has become mandatory in G.fast.

This article presents concepts for a fifth generation access technology, referred to as XG-FAST [5]. It aims to provide data rates up to 10 Gb/s to the end user over very short existing copper-based lines. Target deployment scenarios for XG-FAST include fiber-to-the-frontage (FTTF), where an optical network terminal (ONT) is installed near the boundary between public and private property to shorten the copper loop length while avoiding construction work on customer premises. Other deployment scenarios include home networks. Multi-gigabit data rates can also provide connectivity to 5G wireless networks, which, compared to 4G networks, should achieve 1,000 times the system capacity, 10 times the spectral efficiency, energy efficiency, and data rate (i.e. a peak data rate of 10 Gb/s for low mobility and a peak data rate of 1 Gb/s for high mobility) [6]. As such, XG-FAST then becomes synergistic with 5G wireless, where the XG-FAST drop into the home is extended with multi-gigabit 5G wireless technologies (the latest Wi-Fi 802.11ad standard already aims for maximum data rates of almost 7 Gb/s [7]). Although XG-FAST is presented here as a technology for twisted pair telephone cables, it can essentially be used with different cable types, including Ethernet cables, coaxial cables, and power lines.

0163-6804/15/$25.00 © 2015 IEEE

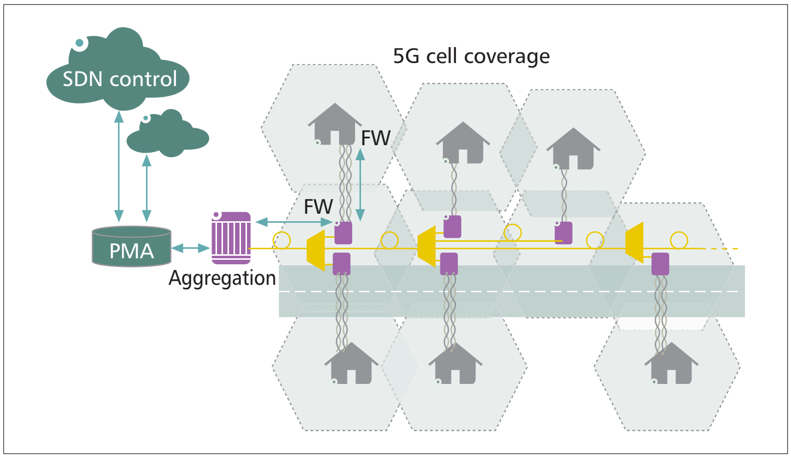


Figure 1. A homes-passed fiber network provides the basis for XG-FAST drops and dense 5G wireless access points. The persistent management agent (PMA) of G.fast is extended to XG-FAST and wireless. It provides the foundation of joint fixed-wireless SDN control as well as containing firmware (FW) for upgrade of network functions inside programmable universal remotes and customer premises equipment.

This article first discusses the role of XG-FAST in future mobile network architectures, which will require a centralized control of traffic referred to as software defined networking (SDN). The XG-FAST system concept is introduced next, and techniques are presented to improve performance. The viability of the XG-FAST technology is demonstrated with measurements performed on several short copper cables, and next steps are proposed to turn the XG-FAST technology concept into practice.

SDN-Controlled 5G Networks

Future 5G networks aim to support a wide variety of use cases, ranging from human-centric communication scenarios that demand high peak rates and low latency, and good connectivity when moving or in a crowd, to machine-oriented communication scenarios with many connected devices and an emphasis on low power consumption [8]. It is a common belief that a single radio access technology will not be able to satisfy all these different requirements simultaneously, and that the future 5G network will consist of a flexible combination of both existing and new radio concepts [8]. This means that 5G will have to be designed as a system rather than as a stand-alone technology.

One of the major 5G trends are small cell overlay networks [9]. This makes it possible to tackle the capacity challenge, and when combined with centralized control, this can be done in a power efficient way. The backhaul of these high capacity small cells must be realized by a high capacity backhaul technology. Often cited examples are mm waves, optical fiber and copper, depending on the targeted use case. We envisage XG-FAST to be a copper backhaul technology with multi-gigabit capacity and sub-ms latency. Low latency is essential for efficient mobile backhauling, and was also targeted for G.fast [3]. XG-FAST relies on a dense deployment of optically backhauled fixed access nodes, and hence is synergistic with the small cell deployment use case of dense urban areas. XG-FAST will consist of a homes-passed fiber network, where each home is connected over copper through XG-FAST, as illustrated in Fig. 1. Such a deployment provides an excellent basis for a 5G network with a low infrastructure deployment cost. Using XG-FAST for backhauling, indoor small cells can be provisioned at one or more locations within the home. In addition, outdoor small cells can be provisioned at the XG-FAST drop units in the street, directly multiplexing into the optical backhaul of XG-FAST. This requires adequate traffic management for the combined wireless and fixed traffic at the XG-FAST node.

Centralized Control

The centralized control plane that is desired for small cells [9] comes naturally with XG-FAST. A dense FTTF deployment requires an "install-and-forget" deployment model, where, after initial installation by a field technician, the access point is managed remotely. In reverse-powered G.fast, a persistent management agent (PMA) is maintained deeper in the network, such that the G.fast distribution point unit is virtually visible and manageable in the absence of reverse power. A PMA serves to provide virtualization of the network management function and can be extended to software defined control and management [10, 11]. A PMA thus becomes a cloud function that can be integrated with SDN control of the wireless access points for consistent management of the end-to-end link. The separation of the management and data plane is not a goal in itself, but follows from the reverse powering targeted for XG-FAST nodes. It enhances the benefits of SDN such as ease of service creation, firmware upgradeability, and facilitating control and diagnostics of a single network with multiple service providers through virtual or bitstream unbundling.

For communication between the PMA and the fixed access node, the industry selected the NETCONF protocol, which provides a means for exchanging configuration information using the data modeling language YANG. The high data rate and low latency of XG-FAST itself facilitates cloudification of 5G control and data plane functions. End users cannot, or will not, deal with in-home Wi-Fi or 5G access point management. Centralized configuration and management is necessary to avoid misconfigurations and conflicting settings of neighboring cells. A further step entails virtualizing layer 2 and layer 3 of fixed and wireless technologies, such as authorization and forwarding rules.

By taking the 5G system design into consideration for XG-FAST, one could imagine a node containing universal hardware components, which are programmable to act as either a fixed or a wireless resource. Centralized control can then be exploited even more efficiently to decrease power consumption. For example, during the day (more mobile users) more resources can be assigned to 5G outdoor wireless access points, while in the evening (more users inside homes) more resources can be assigned to the fixed processing, benefiting the in-home wireless access point capacity.

Given the massive number of required devices for an FTTF deployment, they will have to be as

Technology	Cable type	Duplexing	Modulation	RF bandwidth	Aggregate data rate (US+DS)
G.fast	TP	TDD	DMT (up to 4096-QAM)	104 MHz	1 Gb/s (1 pair)
G.hn	TP, PL, coax	TDD	DMT (up to 4096-QAM)	100 MHz	1 Gb/s (1 pair)
1000/10G BASE-T	TP CAT5/CAT6	Full duplex	PAM-5/PAM-16	≈ 80 MHz/400 MHz	2 Gb/s/20 Gb/s (four pairs)
XG-FAST	TP, coax	Full duplex/TDD	DMT (up to 32768-QAM)	500 MHz	10 Gb/s (one/two pairs)

Table 1. Overview of some relevant gigabit copper access technologies for twisted pair. US: upstream; DS: downstream; TDD: time-division duplexing; DMT: discrete multi-tone; PAM: pulse-amplitude modulation.

Given the massive number of required devices for an FTTF deployment, they will have to be as low-cost and low-power as possible to be viable. Besides centralized control, XG-FAST devices can therefore benefit from centralizing part of their physical layer data functionality.

low-cost and low-power as possible to be viable. Besides centralized control, XG-FAST devices can therefore benefit from centralizing part of their physical layer data functionality. Moreover, in that central aggregation node the centralized functionality can be virtualized to facilitate management and increase flexibility. For downstream traffic, one option could be to centralize some functionality up to the frequency domain, consisting of framing, forward error correction (FEC), and constellation mapping. This keeps the fiber link digital (upholding the compatibility with FTTH), while not blowing up the required fiber capacity.

Copper Broadband

In access networks, existing copper infrastructure mainly consists of twisted pair (TP) and coaxial cabling. Power lines (PL) can also be exploited, especially for in-home distribution. Table 1 highlights several technologies that have been developed for these cable types and which can achieve data rates of at least 1 Gb/s.

The Ethernet technologies 1000 BASE-T and 10G BASE-T are P2P technologies that provide a symmetric data rate of 1 Gb/s and 10 Gb/s, respectively. The physical layer design of these technologies requires a specific cable type (minimum CAT5 or CAT6) containing four twisted pairs, which essentially makes them unsuitable for deployment in copper access networks. The copper access network has mostly been built prior to the advent of the Internet and Ethernet, using telephony-grade cabling with a variable number of pairs available per home. The data rate for XG-FAST will be matched to the cable quality. This flexibility is a crucial advantage in brown field deployments that reuse an existing copper infrastructure.

The other technologies in Table 1 do allow operation on legacy copper access infrastructure. The home network standard G.hn is a P2MP technology specified to achieve data rates up to 1 Gb/s, and can operate over twisted pair, coaxial cable, and power lines [12]. It uses DMT modulation over 100 MHz of baseband or passband spectrum, and low density parity check (LDPC) coding for forward error correction. Similarly, G.fast uses DMT modulation over a 104 MHz block of spectrum, ranging from 2.2 MHz to 106 MHz, to avoid spectral overlap with legacy ADSL technologies. The current version of the standard is optimized for loops shorter than 250 m and uses time division duplexing (TDD) to duplex upstream and downstream traffic. The FEC is Reed Solomon (RS) combined with trellis coded modulation (TCM). Unlike G.hn, G.fast has a carrier-grade management protocol. Despite the similarity of their physical layer, G.fast outperforms G.hn in access networks thanks to crosstalk cancellation and increased framing efficiency.

XG-FAST further expands the signal bandwidth to 500 MHz, making it possible to increase the achievable data rate over the shortest loops. In the next Section, we propose synergistic system concepts for XG-FAST that make it possible to increase its spectral efficiency compared to G.fast, such as adaptive modulation, MIMO vectoring in combination with bonding and phantom mode transmission, and full duplex transmission.

XG-FAST System Concepts

Powering

Because of the sheer number and distributed nature of XG-FAST devices in an FTTF deployment scenario, the devices will preferably be powered by the customer premises equipment (CPE), a scenario known as "reverse powering." As the XG-FAST device is tailored for a single user, reverse powering becomes easier compared to multi-user distribution point units. The shorter loop lengths also reduce the loss of the reverse power provided over that loop. This should simplify the reverse powering requirements compared to G.fast.

Bandwidth

The biggest contributor to the data rate increase will be the expansion of the signal bandwidth, enabled by shortening the copper loops through increasing fiber penetration. The 70 m operator cable in the proof-of-concept later uses less than 400 MHz, while the 30 m cable is able to exploit a bandwidth larger than 500 MHz. Based on these cable measurements and modeling, a signal bandwidth of the order of 500 MHz is considered to be a good choice for XG-FAST [5].

Modulation

Current xDSL technologies use discrete multitone (DMT) modulation, which divides the frequency spectrum into equally spaced parallel channels or "tones" that are independent due to orthogonality. The bit-loading per tone can be separately adjusted for each tone. DMT modulation for XG-FAST is a logical choice, in order to be able to exploit the full potential of every copper loop (whatever its quality) with a finely-tuned bit-loading on every tone.

XG-FAST will allow for simultaneous upstream and downstream on the same frequency (full duplex), doubling the spectral efficiency compared to currently used time or frequency division duplexing schemes.

Adaptive Modulation and Coding

Transients in wideband interference originating from external sources have an impact over a large part of the signal bandwidth. To accommodate such noise transients, the gap to capacity is typically artificially increased by adjusting the signal-to-noise-ratio margin (SNRM). This is essentially a proactive capacity reduction mechanism that ensures quality of service in case of an unexpected increase in wideband noise. A typical SNRM value used in practice is 6 dB, corresponding to a spectral efficiency loss of 2 bit/s/Hz. A modulation technique called transmitter controlled adaptive modulation (TxCAM) makes it possible to reduce this loss in spectral efficiency [13]. TxCAM is a hierarchically-layered modulation scheme in which the transmitter can autonomously adapt the data rate by turning off layers in case of a sudden noise increase. This autonomy mitigates the need for a lengthy command-and-response procedure to negotiate a change between transmitter and receiver to do a data rate adaptation. With TxCAM, the line can be operated without a SNRM and hence increase the spectral efficiency by up to 2 bit/s/Hz. Regarding forward error correction, TxCAM is unsuited for use with coded modulation such as the TCM currently used in G.fast. So although a shift from TCM to low density parity check (LDPC) coding on the merits of coding gain alone may not be worthwhile [14], the use of TxCAM would be synergetic with LDPC coding as a forward error correction scheme.

Bonding

In many places, customer sites have been equipped with two copper pairs, one for dedicated voice service and another for fax or early dial-up data access. This can be exploited by XG-FAST, which has the benefit of being a single subscriber device and allows each customer to be served with an optimized technology.

These pairs can be bonded together, meaning that the two physical layer data streams are multiplexed to provide a single pipe to the user. When vectoring is used, the achievable data rate gain is proportional to the number of channels.

Phantom Mode Transmission

Two available twisted pairs can also be used as two "virtual" wires to create a third "virtual" pair, called the phantom mode. This concept is well known and was already considered at the end of the 19th century. The differential signal of the phantom mode is the difference between the common mode signals on both twisted pairs. The channel characteristics of the phantom mode are typically worse than those of the twisted pairs because its two "wires" are essentially two different twisted pairs, resulting in unbalances. Due to the large crosstalk interference generated by phantom channels, it is essential to use vectoring to yield a data rate gain that is worth the hardware effort.

The three data streams associated with the three modes can again be bonded together, potentially providing a threefold increase in data rate compared to a single twisted pair. Further on, we will experimentally demonstrate the feasibility of using phantom mode transmission at the high frequencies targeted for XG-FAST.

Two-Sided Signal Coordination

Crosstalk cancellation techniques remain essential in XG-FAST when multiple pairs are being used by a single user. In VDSL2 and G.fast, crosstalk cancellation is achieved by signal coordination solely at the access node, which is a multiple-input-single-output (MISO) scheme. Since XG-FAST is primarily a single-user technology, it is now possible to achieve signal coordination in both the transmitter and the receiver, creating opportunities for new precoder and equalization schemes. It allows the use of better performing multiple-input-multiple-output (MIMO) schemes, which was not possible in previous DSL generations.

In G.fast, a linear gain-scaled precoder has been introduced to address the fact that crosstalk can be as large or larger than the direct channel, and nonlinear precoders are being studied [2]. With two-sided coordination, the power penalty due to linear or non-linear precoding can be removed by exploiting the eigenmodes of the channel with singular value decomposition (SVD). In this scheme, both precoder and postcoder matrices are unitary matrices. Precoding at the transmitter rotates the signal vector without increasing the transmit power, and postcoding at the receiver rotates the received signal vector without noise enhancement. These operations diagonalize the channel, decomposing it into separate virtual sub-channels (the eigenmodes of the channel). The SVD transmission scheme can be interpreted as the coherent combination of the signals so as to achieve improved SNR. Variants for the precoder and equalizer matrices can be obtained based on singular value decomposition that take a minimum mean-square error criterion, rather than zero-forcing, or that take noise covariance into account.

Although XG-FAST is primarily a single-user technology, it does not preclude coordination at the central side among multiple bonded vectoring groups. In that case, zero-forcing may be applied across groups, while maintaining the above precoder and postcoder structure [15].

Full Duplex Transmission

In multi-user DSL deployments, simultaneous upstream and downstream communication on identical frequencies is made impossible by near-end crosstalk at the customer side of the network. When used as a single-subscriber technology, XG-FAST will have no NEXT between CPEs, hence allowing for simultaneous upstream and downstream on the same frequency, which we will refer to as full duplex, doubling the spectral efficiency compared to currently used time or frequency division duplexing schemes.

It requires an analog hybrid at the transceivers that attenuates the transmit signal to the level of or below that of the received signal, not to substantially increase the required dynamic range of the analog-to-digital converter. The residual echo signal is further removed digitally using vectoring techniques. Note that compared to TDD, full duplex transmission also allows a reduced latency and framing overhead.

Proof of Concept

In this section, we demonstrate the potential of XG-FAST with a hardware proof-of-concept platform. It incorporates the system concepts mentioned above: increased bandwidth, two-sided coordinated vectoring, bonding and phantom mode, and full duplex transmission. We also operate the system at an SNRM of 0 dB, which emulates the benefit of using TxCAM. The purpose of this demonstration is to show the raw potential of XG-FAST. It is not meant to be a fully optimized platform.

A 500 MHz baseband spectrum is used for the DMT signal, and forward error correction is provided using a (255, 239) Reed Solomon code. We show results for two duplexing schemes: TDD as in G.fast, and full duplex. The vectoring matrices are calculated based on channel data measured by transmitting a pilot sequence of 16 DMT symbols. The short length of the copper loops allows the use of a short cyclic extension (CE), which was never longer than 1.2 μs. The framing overhead is 6.25 percent for TDD, as in G.fast, and 3.47 percent for full duplex. The transmit power was never larger than 5 dBm per direction. The performance was measured on CAT5e cables (AWG-24 or 0.51 mm diameter) of length 30 m to 120 m, and on telephone cables provided by a European operator (0.6 mm copper diameter) of length 30 m to 130 m.

Figure 2 shows the aggregate (sum of upstream and downstream) net data rate (NDR) that was achieved on the different cables. The squares correspond to the CAT5e cables, and the stars correspond to the operator cables. The solid (blue) lines show the NDR over two pairs using bonding, phantom mode, and TDD; the dashed (orange) lines show the rate over a single pair using full duplex; the dotted (green) lines show the rate achieved over a single pair using TDD.

Using a single pair, an aggregate NDR of 5 Gb/s and 4.5 Gb/s is achieved on the shortest 30 m CAT5e and operator cable. Up to 70 m operator cable and 120 m CAT5e cable, it is possible to achieve an aggregate NDR exceeding 2 Gb/s, allowing for 1 Gb/s symmetric, which is an important milestone for data communication over a single telephone line.

Still using a single pair, but operating in full duplex instead of TDD, the aggregate NDR can be almost doubled to 9.7 Gb/s (CAT5e) and 8.8 Gb/s (operator cable) on the shortest loops. The reach of symmetric gigabit service is also greatly increased to 130 m, achieving a 2.5 Gb/s aggregate rate on operator cabling. The ratio of the full duplex rate over twice the TDD rate ranges from 83 percent to 98 percent for the different cables. Further improvements are expected by improving the analog front end of the transceivers.

When using TDD on two twisted pairs while exploiting bonding and phantom mode transmission, the aggregate NDR exceeds 10 Gb/s on a 30 m operator cable and a 50 m CAT5e cable. Note that the two pair measurements currently had to be performed with an inferior analog front end compared to the single pair measurements, yielding a downward bias in data rates. Compared to a single line rate with the same analog front end (not shown), bonding and phantom mode combined provide a gain between 2.6 and 2.8 for all cables, except for the 0.6 mm operator cable of 70 m, which yields a combined gain as high as 3.2 due to the MIMO benefits of vectoring with two-sided coordination.

The operator cables of 70 m and 130 m are only able to use 400 MHz and 250 MHz of signal bandwidth, respectively, due to the increasing channel loss with increasing length. Expanding the spectrum beyond 500 MHz yields modest improvements on the shortest cables. For example, expanding the bandwidth to 650 MHz and 800 MHz on the 30 m operator cable, respectively, allows a modest 400 Mb/s and 500 Mb/s rate increase per direction on top of the 4.5 Gb/s that is achieved over 500 MHz.

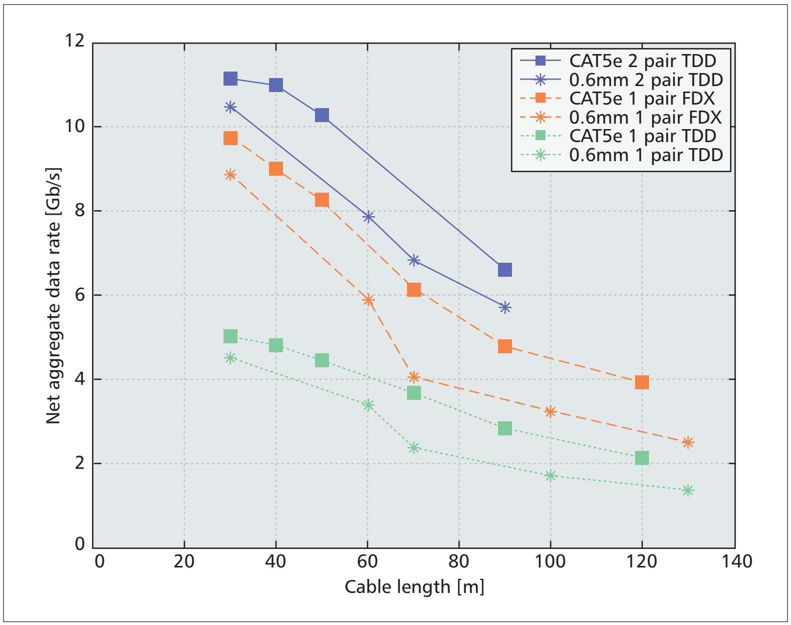


Figure 2. Rate-reach curve for CAT5e cables and 0.6 mm cables provided by a European operator on one and two twisted pairs with the proof-of-concept hardware platform.

We also performed measurements on a coaxial cable of RG6 type, which is a typical drop cable used in hybrid fiber-coaxial networks. The superior channel characteristics compared to twisted pair cabling make it possible to extend the reach of XG-FAST. An NDR of 5.38 Gb/s was achieved on a 20 m long RG6 cable using TDD and a transmit power of 3 dBm. Increasing the length of the RG6 cable to 100 m only reduced the NDR to 5.25 Gb/s using a transmit power of 7 dBm, indicating a nearly flat rate-reach curve for loop lengths up to at least 100 m.

Conclusion

In this article, we introduced the basic concepts of XG-FAST, a novel multi-gigabit transmission technology capable of transmitting more than 10 Gb/s over short copper pairs. One of the possible deployment scenarios for XG-FAST is FTTF (ONT-outside-the-home), in which XG-FAST is used as the next-generation DSL technology succeeding G.fast. The performance will depend on the cable quality and will gracefully degrade with increasing copper loop length, which is desired in brown field deployments that reuse an existing copper infrastructure.

We experimentally demonstrated an aggregate net data rate exceeding 8.8 Gb/s and 2.5 Gb/s over a single pair of, respectively, a 30 m and 130 m 0.6 mm twisted pair telephone cable from a European operator, using full duplex transmission and a signal bandwidth of 500 MHz. Using two pairs and time division duplexing instead of full duplex, we

The hybrid fiber-copper network with XG-FAST provides an ideal basis for dense deployment of 5G small cells. With uniform SDN control across fixed and wireless and the virtualization of functions within the access points, a single universal and programmable access point comes within reach.

demonstrated a 10 Gb/s rate using bonding and phantom mode transmission for loops up to 30 m.

As potential future work, the combination of full duplex transmission and multi-pair bonding and vectoring will be able to achieve aggregate rates exceeding 20 Gb/s, which is 10 Gb/s symmetric, on only two pairs. As a reference, 10GBASE-T delivers an aggregate data rate of 20 Gb/s over four pairs of CAT6 cable up to 55 m long. Other future work includes the exploration of vectoring schemes in a multi-user setting with multiple pairs per user (e.g. apartment buildings), and channel characterization of the installed copper plant for the high frequency range of 500 MHz considered for XG-FAST.

The hybrid fiber-copper network with XG-FAST provides an ideal basis for dense deployment of 5G small cells. With uniform SDN control across fixed and wireless and the virtualization of functions within the access points, a single universal and programmable access point comes within reach. The XG-FAST technology would complement fiber and next-generation 5G wireless technologies and be a further evolution of the G.fast technology. FTTF XG-FAST deployments will naturally allow the existing copper plant to serve both future wireline access and backhaul for 5G mobile technologies that are specified to operate at multi-gigabit data rates.

The fourth generation DSL technology, G.fast (G.9701), was recently standardized at ITU-T in December 2014. Ongoing efforts are focused on amendments for low-power modes and an updated band plan up to 212 MHz. Standardization of XG-FAST is beneficial, even though it is a single-user technology. The standardization of the physical layer helps grow the market for operators, leading to more cost-effective chip sets due to economies of scale. The physical layer of XG-FAST could leverage the consented ITU-T G.fast standard, requiring an updated band plan covering frequencies up to 500 MHz, or give rise to a new standard allowing for novel concepts to be included (e.g. adaptive modulation, multiple pairs, phantom mode, two-sided vectoring, full duplex). Special attention to transmit PSD limitations by the ITU Terrestrial and Radio groups will be required given the larger bandwidths used for signaling.

References

[1] Self-FEXT Cancellation (Vectoring) for Use with VDSL2 Transceivers, ITU-T Recommendation G.993.5, Apr. 2010.

[2] M. Timmers ***et al.***, "G.fast: Evolving the Copper Access Network," *IEEE Commun. Mag.*, vol. 51, no. 8, Aug. 2013, pp. 74–79.

[3] *Fast Access to Subscriber Terminals (FAST) — Physical Layer Specification*, ITU-T Recommendation G.9701, Dec. 2014.

[4] P. Ödling ***et al.***, "The Fourth Generation Broadband Concept," ***IEEE Commun. Mag.***, vol. 47, no. 1, Jan. 2009, pp. 63–69.

[5] W. Coomans ***et al.***, "XG-FAST: Towards 10 Gb/s Copper Access," *Proc. IEEE Global Commun. Conf. 3rd IEEE Wksp. Telecommunication Standards: From Research to Standards*, Austin, TX, Dec. 2014, pp. 630–35.

[6] C.-X. Wang ***et al.,*** "Cellular Architecture and Key Technologies for 5G Wireless Communication Networks," ***IEEE Commun. Mag.***, vol. 52, no. 2, Feb. 2014, pp. 122–30.

[7] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications — Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band*, IEEE Std. 802.11ad, Dec. 2012.

[8] A. Osseiran ***et al.***, "Scenarios for 5G Mobile and Wireless Communications: The Vision of the METIS Project," ***IEEE Commun. Mag.***, vol. 52, no. 5, May 2014, pp. 26–35.

[9] V. Jungnickel ***et al.***, "The Role of Small Cells, Coordinated Multipoint, and Massive MIMO in 5G," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 44–51.

[10] B. Kozicki ***et al.***, "Software-Defined Networks and Network Functions Virtualization in Wireline Access Networks," *Proc. IEEE Global Commun. Conf. 3rd IEEE Wksp. Telecommunication Standards: From Research to Standards*, Austin, TX, Dec. 2014, pp. 595–600.

[11] K. J. Kerpez ***et al.***, "Software-Defined Access Networks," ***IEEE Commun. Mag.***, vol. 52, no. 9, Sept. 2014, pp. 152–59.

[12] Unified High-Speed Wireline-Based Home Networking Transceivers — System Architecture and Physical Layer Specification, ITU-T Recommendation G.9960, Dec. 2011.

[13] M. Timmers ***et al.***, "Transmitter-Controlled Adaptive Modulation: Concepts and Results," *Bell Labs Tech. J.*, vol. 18, no. 1, 2013, pp. 153–69.

[14] J. Neckebroek ***et al.***, "Comparison of Error-Control Schemes for High-Rate Communication over Short DSL Loops Affected by Impulsive Noise," *Proc. IEEE Int. Commun. Conf.*, Budapest, Hungary, June 2013, pp. 4014–19.

[15] R. B. Moraes ***et al.***, "DMT MIMO IC Rate Maximization in DSL with Combined Signal and Spectrum Coordination," *IEEE Trans. Signal Proc.*, vol. 61, no. 7, July 2013, pp. 1756–69.

Biographies

Werner Coomans (werner.coomans@alcatel-lucent.com) has been an access technology researcher at Bell Labs since 2013. His research focuses on next-generation fixed access technologies over copper. Prior to joining Bell Labs he obtained master and Ph.D. degrees in engineering sciences, with a focus on semiconductor lasers, from the Vrije Universiteit Brussel (VUB).

Rodrigo B. Moraes [S'08, M'14] obtained the bachelor degree at the UFPA, Belém, Brazil, in 2005, the M.Sc. degree at the Pontifical Catholic University, Rio de Janeiro, Brazil, in 2009, and the Ph.D. degree in 2014 at the KU Leuven, Belgium, all in electrical engineering. Since 2014 he has been a research engineer at Bell Labs, Alcatel-Lucent, in Antwerp, Belgium. He has received best paper awards at IEEE ICC (2013) and IEEE Globecom (2014).

Koen Hooghe received an electrotechnical engineering degree from the University of Ghent. He is a hardware expert with Bell Labs Access Node Technology and has participated in the development of access systems and core routers as a hardware designer, project lead, and systems architect. He has been involved in research for next-generation access nodes at the physical layer, packet processing, systems architecture, and energy efficiency improvements. He contributed to the ETSI Environmental Engineering standards group.

Alex Duque received a BSEE degree from Lehigh University and an MSEE degree from Worcester Polytechnic Institute. He began his career designing communication satellite hardware for Lockheed-Martin Corporation. He then designed cell phone modem ASICs for Lucent Technologies. In 1998 he shifted from the Consumer Products Division into various research groups within Bell Laboratories, where he continues to develop ASIC and FPGA designs both for research projects and for Alcatel-Lucent products.

Joe Galaro joined Bell Labs in 1990, after spending 10 years with Lockheed Electronics working on DSP, ASIC, and hardware design for aircraft, satellite, and shipboard systems. In his tenure at the labs he developed various technologies providing higher bandwidth data transmission. These technologies have included dial-up modems, ISDN, cable modem CMTS, DSL, HD IBOC radio, XGPON, and XG-FAST. He provides significant hardware design experience (circuit design, FPGA, and ASIC) as well as system architecture insights.

Michael Timmers is a technology strategist in the Fixed Networks Division of Alcatel-Lucent turning ideas into reality by closely engaging with research, development, and customers. Previously, he researched and prototyped the latest DSL advancements in Bell Labs Access Research Domain in Antwerp, Belgium. He received an M.S. in engineering from Katholieke Universiteit Leuven in 2005, and a Ph.D. with a dissertation on cognitive radio in 2009 at the Interuniversitary MicroElectronics Center (IMEC) in Belgium.

Adriaan J. van Wijngaarden [S'87, M'98, SM'03] is a research scientist at Bell Laboratories, Murray Hill. He has been deeply engaged in theoretical and application-driven research in communications, information theory, coding, and algorithmic optimization, and provided key contributions to optical transport and access networks. He has more than 75 publications and 65 patents. He is a co-recipient of the Bell Labs President's Award (2011) and a best paper award (Globecom, 2014).

Mamoun Guenach (guenach@gmail.com) has been a research scientist at Bell Labs, Alcatel-Lucent, since 2006. He received the degree of engineer in electronics and communications from the Ecole Mohamadia d'Ingénieurs in Morocco. Following that, he moved to the faculty of applied sciences at the Université Catholique de Louvain (UCL) Belgium, where he received an M.Sc. degree in electricity and a Ph.D. degree in applied sciences. Since 2015 he has been a part-time visiting professor at Ghent University.

Jochen Maes (jochen.maes@alcatel-lucent.com) [SM'11] joined Bell Labs in 2006, where he continuously shifts the limits of copper. He holds an M.Sc. in physics and a Ph.D. in science. He heads the Copper Access team within the Fixed Networks program, focused on leveraging legacy infrastructure through innovations in systems and transceiver design. He contributes to several ITU standards. His work has received Broadband Infovision Awards (2010, 2014) and the Bell Labs President's Award (2011).